Technische Universität München

TUM School of Computation, Information and Technology

Phase Retrieval from Short-Time Fourier Measurements and Applications to Ptychography

Oleh Melnyk

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Johannes Zimmer

Prüfer*innen der Dissertation:

1. Prof. Dr. Felix Krahmer
2. Prof. Mark Iwen, Ph. D.
3. Prof. Dr. Stefan Kunis

Die Dissertation wurde am 17.06.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 15.02.2023 angenommen.

Abstract

We consider recovery from ptychographic measurements also known as Short-Time Fourier Transform phase retrieval. That is, the unknown object of interest has to be reconstructed from a set of diffraction patterns resulting from a series of localized illuminations. In this thesis, we work with three iterative methods for reconstruction, namely, Amplitude Flow, Error Reduction and Ptychographic Iterative Engine. We show that each can be seen as a gradient method for an amplitude-based squared loss, which allows to establish sublinear convergence guarantees for these methods. In addition, we study the Block Phase Retrieval algorithm, a non-iterative method designed specifically for ptychographic measurements and propose modifications for an improved performance.
Furthermore, we consider blind ptychographic reconstruction, a scenario where the illumination function is unknown and has to be estimated along with the object. We propose a version of the Amplitude Flow algorithm based on the alternating minimization technique with guaranteed sublinear convergence to a fixed point.
Finally, some of the previously discussed recovery methods are extended to the more general case of polychromatic illumination.

Zusammenfassung

Wir betrachten das Problem der Rekonstruktion eines Objektes aus ptychographischen Messungen, welches auch als Kurzzeit-Fourier-Transformation Phasenrekonstruktion bezeichnet wird. Das Ziel eines ptychographischen Experiments ist die Rekonstruktion eines unbekannten Objektes aus Beugungsbildern, welche durch eine Serie lokalisierter Beleuchtungen erhalten werden. Im Rahmen dieser Arbeit betrachten wir drei iterative Rekonstruktionsmethoden, welche als Amplitude Flow, Error Reduction sowie Ptychographic Iterative Engine bezeichnet werden. Wir zeigen, dass jede dieser Methoden als eine Variante eines Gradienten-Abstiegs-Verfahrens angewandet auf eine amplituden-basierte quadratische Verlustfunktion, interpretiert werden kann. Mit Hilfe dieses Zusammenhangs beweisen wir, dass sublineare Konvergenz für alle aufgeführten Methoden garantiert werden kann. Darüber hinaus untersuchen wir der sogenanten Block-Phase-Retrieval-Algorithmus. Diese Methode wurde speziell für ptychographische Messungen entwickelt. Wir stellen Anpassungen dieser Methode vor, welche eine verbesserte Leistungsfähigkeit versprechen.
Des Weiteren betrachten wir das Problem der sogenannten blinden Ptychographie. Dies bezeichnet ein Rekonstruktionsverfahren, in welchem neben dem Objekt auch die Beleuchtungsfunktion unbekannt ist und gleichzeitig mit dem Objekt aus den Daten rekonstruiert werden muss. Für diese Problemstellung entwickeln wir eine Version des Amplitude Flow Algorithmus, die auf einem alternierenden Minimierungsverfahren beruht. Wir beweisen, dass dieses Verfahren sublinear zu einem Fixpunkt konvergiert.
Zuletzt erweitern wir einige der zuvor diskutierten Rekonstruktionsmethoden für den verallgemeinerten Fall einer polychromatischem Beleuchtung.

Acknowledgment

First of all, I would like to thank my supervisor Felix Krahmer. Starting with master thesis and continued in doctoral studies, he gave me creative freedom to pursue research directions and ideas of choice and helped with useful advise when necessary.

Secondly, I am grateful to my advisor at Helmholtz Center Munich, Frank Filbir for his involvement, dedication, skillfulness and sense of humor. Without him, these years would be completely different.

I also thankful to Benedikt Diederichs, Patricia Römer and Hanna Veselovska for filling the working time with interesting and fun chattering and sharing their ideas and opinions.

Moreover, I express my gratitude to Tim Fuchs, Peter Jung, Christian Kümmerle, Marco Mondelli, Nada Sissouno, Dominik Stöger and Claudio Verdun for insightful scientific discussions. Furthermore, I thank Florian Boßmann, Mark Iwen, Stefan Kunis, Dominik Nagel, Ryan Saab and Pierre Weiss for an occasional exchange of thoughts and references.

Special thanks go to the members of the Phase Retrieval Journal Club and, in particular, Lukas Liehr and Arseniy Tsipenyuk. Our discussions broadened my knowledge about phase retrieval and helped to address some of the open problems.

I thank to Massimo Fornasier, Markus Muhr, Isabella Wiegand and Barbara Wohlmuth for their initiative to help Ukrainian scientists who escaped the full-scale russian invasion in Ukraine. This program helped me to feel useful in spring 2022 and allowed to meet Dmytro Sytnyk, Oksana Chernova and, by extension, Oleksandr Zadorozhnyi.

As a part of my teaching duties at Technical University of Munich, I advised master theses of Anton Forstner, Sarah Dörr and Anastasia Kireeva. This collaboration a very pleasant experience.

I thankful to my office mate Carlos Améndola Cerón for being the biggest Star Wars fanboi. I thank colleagues and guests at Technical University of Munich who I met over the years: Stefano Almi, Ricardo Acevedo Cabra, Stefan Bamberger, Sandro Belz, Vjosa Blakaj, Mauro Bonafini, Cristina Cipriani, Olga Graf, Martina Gschwendtner, Hui Huang, Richard Huber, Sara Krause-Solberg, Johannes Maly, Peter Massopust, Olga Minevich, Felipe Pagginelli Patricio, Michael Rauchensteiner, Konstantin Riedl, Bernhard Schmitzer, Giacomo Sodini, Philippe Sünnen, Felix Voigtländer, Laure Vuaille and Jan Vybiral. Our discussions were fun and made the office life more enjoyable. I also appreciated company of Wolfgang zu Castell, Werner Dubitzky, Marion Engel, Carlos

Garcia Perez, Keiichi Ito, Martina Mangold-Troup, Sandra Mayer, Josef Obermaier, Murali Sukumaran, Mahyar Valizadeh, Hannah Zoller and other colleagues at Helmholtz Center Munich.

This research was funded by Helmholtz Association, project Ptychography 4.0 (ZT-I-0025). I would like to thank all collaborators within this project and, in particular, to Arya Bangun, Alexander Clausen, Wilhelm Eschen, Nico Hoffmann, Christian Kübel, Benjamin März, Xaoke Mu, Knut Müller-Caspary, Christian Schroer, Erik Thiessenhusen, Maximilian Töllner and Dieter Weber.

Last but not least, I am grateful to my girlfriend Svitlana, my parents Leonid and Olena, and my sister Antonina and brother Sergii for their continuous support during these years and beyond.

Contents

Introduction

In microscopic and optical experiments, an object of interest is illuminated by a ray of light. It further travels through a lens and as a result an inverted picture of the object can be seen on a screen. However, for nanoscale microscopy this setup is not feasible due to the absence of good quality lenses. Without optics it is only possible to capture a diffraction pattern of the object and the recovery of the object from the diffraction pattern has to be performed numerically. This problem is also called Fourier phase retrieval. It is well-known [1] that, without additional assumptions on the object function, the unique recovery is not possible, which is a major drawback in many applications. The desire to mitigate this drawback led to alternative imaging methods, one of which is the so-called ptychography [2, 3, 4].

Ptychography (Figure 1) is a lensless imaging technique, which aims to recover the object of interest from a set of diffraction patterns. Each of these images is obtained by illuminating a small region of the specimen at a time by penetrating light such as an electron beam or an X-ray. As a result, the light encodes the information about the object and propagates further to the detector, where the resulting diffraction pattern is recorded. Then, the object is shifted and the next region is illuminated. The regions overlap and, thus, the obtained diffraction patterns contain a surplus of information, which allows a unique reconstruction of the object. Ptychographic recovery is also referred to as Short-Time Fourier Transform (STFT) phase retrieval as the measurements are given by the magnitudes of STFT of the specimen.

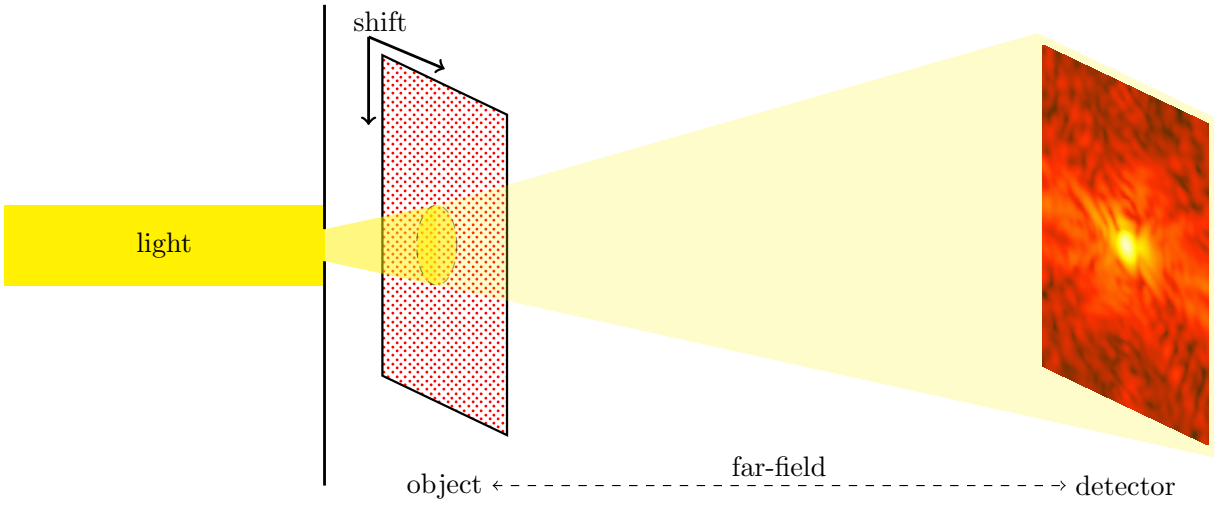


Figure 1: An illustration of ptychographic experiment.

The successful applications of ptychography [5, 6, 7, 8, 9, 10, 11, 12] also led to an extension of this technique to other imaging scenarios, which include tomographic ptychography [13, 14, 15, 16, 17, 18, 19], multislice ptychography [20, 21, 22, 23] and polychromatic ptychography [24, 25]. These works contain applications in biology [13, 8, 18, 11], material sciences [15, 14, 9, 16, 10, 17], crystallography [12, 23] and many other fields [7, 19, 21].
As the popularity of ptychography rose among the practitioners, it caused a growth of data volumes and increased the demand for efficient reconstruction methods. In response, many algorithms and techniques [26, 27, 28, 29, 30, 31, 32, 33, 34, 35] were developed and used for reconstruction. Commonly, a typical algorithm requires an initial guess and then generates a sequence of iterates in some way. At some point the algorithm reaches a predefined stopping criteria and the final iterate is returned as the reconstructed object. The mathematical justification for a good and fast performance of such methods is a complicated task and often the available analysis is restricted to specific scenarios. Therefore, these methods often depend on a good starting point, which is frequently found by trial-and-error.
Despite of many forms of ptychographic imaging and related algorithms, many aspects of ptychography lack an analytical understanding, which created a large number of contributions by the mathematical community. We hope, that this thesis will provide a contribution to a better understanding of ptychographic reconstruction. It contains the following chapters and sections:

- In Chapter 1 we review the physics behind ptychography. Three core phenomena will be considered. These are propagation of light, intensity of light and the interaction of light with an object. The combination of these three steps provides the mathematical description of the diffraction patterns captured by the detector.

- Chapter 2 provides the reader with notation and mathematical concepts used throughout the thesis. In particular, we review Wirtinger derivatives and results about standard and stochastic gradient descent for real-valued functions of complex variables.

- Chapter 3 is dedicated to the recovery from ptychographic measurements. Its first part familiarizes the reader with the background on ptychographic recovery, which includes the continuous ptychographic problem (Section 3.1) and its discretization (Section 3.2). The main results regarding ambiguities, uniqueness and stability of reconstruction for discrete ptychographic problem will be discussed in Section 3.3. An overview of reconstruction algorithms can be found in Section 3.4.

  The second part of Chapter 3 focuses on three iterative methods widely discussed in the literature. Section 3.5.1 is dedicated to the gradient descent technique for amplitude-based squared loss known as Amplitude Flow [36]. Section 3.5.2 is about the Error Reduction algorithm [37, 38], which is an alternating projections approach. Finally, the Ptychographic Iterative Engine (PIE) [39, 33], a computationally fast method utilizing a single diffraction pattern at the time, is discussed in Section 3.5.3. We show that the latter two algorithms can also be viewed as gradient methods for the same amplitude-based squared loss function. More precisely, we show that Error Reduction is a scaled gradient descent and PIE is nothing else but a stochastic gradient descent. Furthermore, based on the convergence theory for gradient meth-

ods, we establish the guaranteed convergence of both algorithms and show that the convergence speed is sublinear.

The last part of Chapter 3 studies the Block Phase Retrieval method for ptychography [35, 40]. It is a non-iterative approach that possesses theoretical error bounds for the reconstructed object, which makes this approach unique. As a drawback, it is strongly restricted to the specific experimental setup. Section 3.6 covers the main results on Block Phase Retrieval, including our modifications for a better performance and relaxations of the setup restrictions.

- Blind ptychography is the main topic of Chapter 4. While in ptychography the distribution of the light inside the illuminated region is assumed to be known and is actively used during the recovery process, in blind ptychography it is considered unknown and has to be reconstructed along with the object. Consequently, the joint recovery is more complicated and requires special care. We propose a version of Amplitude Flow for blind ptychography based on the alternating minimization technique and derive its sublinear convergence speed.

- In Chapter 5, polychromatic ptychography is considered. Unlike ptychography where the light consist of a single coherent wave, in polychromatic ptychography the light is a mixture of several waves. We discuss how the mathematical description of the resulting measurement process changes. Furthermore, we extend the results on Amplitude Flow established for non-blind (Section 3.5.1) and blind ptychography (Chapter 4) to the case of polychromatic light.

- Finally, in Chapter 6 we perform numerical trials to evaluate the performance of all proposed algorithms. Based on specific performance measures, we select an optimal modification of the Block Phase Retrieval algorithm. Then, it is used as an initialization for Amplitude Flow, Error Reduction and PIE to improve the reconstruction quality. Furthermore, we study reconstruction in case of blind, polychromatic and blind polychromatic ptychography.

Chapter 1

Light, diffraction and ptychography

In this section, we review the physical background on the light diffraction and derive the diffraction formulas for the ptychographic experiment. The presented results are a compilation of the material found in [41, 42]. Our main goal is to familiarize the reader with the physical model which is the starting point for the analysis provided in the rest of the thesis. Hence, for the sake of readability and in order to avoid technicalities, we allow ourselves to be not strictly mathematically rigorous. Namely we will assume that all functions, measures, etc. belong to the appropriate spaces, so that all desired properties apply.
In order to establish the mathematical description of the ptychographic measurements, we decompose a single illumination within the ptychographic experiment into separate steps (Figure 1.1):

1. The light from the source reaches the aperture plane, where it is mostly absorbed and only propagates further through a small slit.

2. The localized light travels from the aperture plane to the object plane.

3. The light penetrates the object and encodes information about it.

4. The resulting light travels from the object plane to the detector plane.

5. The detector captures the incoming light.

Each step is separately explained on the basis of the *electromagnetic* (EM) wave theory of the light and, in particular, the Maxwell's equations [43]. That is, steps 2 and 4

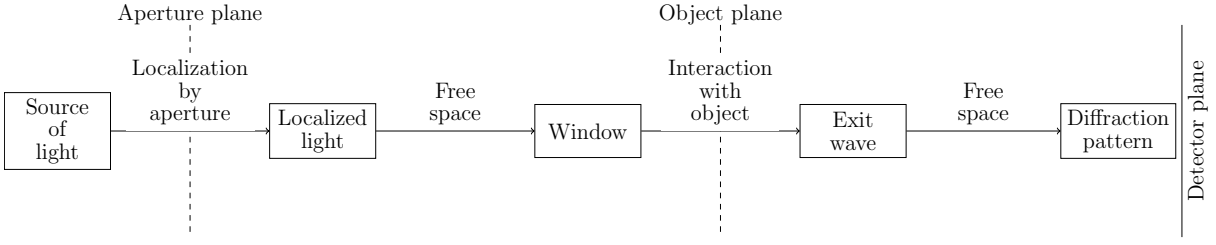


Figure 1.1: Schematics of the ptychographic experiment.

correspond to the propagation of an EM wave in free space, i.e., without interaction with the environment and obstacles, which is mathematically explained in Section 1.1. Steps 1 and 3 are essentially an interaction of light wave with matter, which is covered in Section 1.2. The last step is related to the notion of the optical intensity of EM wave, which is the topic of Section 1.3.
The first two steps combined describe the distribution of the light reaching the object, also known as the *probe* or the *window*. Given the window, steps 3-5 completely describe the resulting image on the detector called the *diffraction pattern* and its mathematical description can be found in Section 1.4. Furthermore, the formulas for the window corresponding to the circular aperture, a common choice in practice, and its approximation are derived in Section 1.5.

1.1 Electro-magnetic wave in free space

In this section we study the propagation of an EM wave in free space. In particular, in the setting of the ptychographic experiment, we are interested in the propagation of the wave in a certain direction from the starting plane (either aperture or object) to the target plane (object or detector, respectively). In such scenarios, the EM wave entering the free space is assumed be known and its values at the target plane are to be determined.
A starting point in understanding the behavior of the light in free space are the Maxwell's equations [42]. They establish the relationship between an electric field $\mathcal{E}(s,t) : \mathbb{R}^3 \times \mathbb{R} \mapsto \mathbb{R}$ and a magnetic field $\mathcal{H}(s,t) : \mathbb{R}^3 \times \mathbb{R} \mapsto \mathbb{R}$ as a system of differential equations. In the following $\nabla = (\partial_x, \partial_y, \partial_z)^T$ denotes a vector of partial space derivatives and ∂_t is the time derivative, $\cdot$ is the inner or dot product and $\times$ is the vector product. The operations $\nabla\times$ and $\nabla\cdot$ are also known as curl and divergence, respectively.
With these notations, the Maxwell's equations in free space read as

$$\begin{aligned} \nabla \times \mathcal{E} = -\mu_0 \partial_t \mathcal{H}, \quad \nabla \times \mathcal{H} = \varepsilon_0 \partial_t \mathcal{E}, \\ \nabla \cdot \mathcal{E} = 0, \quad \nabla \cdot \mathcal{H} = 0, \end{aligned}$$

where constants μ_0 and ε_0 are the magnetic permeability and the electric permittivity, respectively. An application of the curl operation $(\nabla\times)$ to the first two equations and usage of the identity

$$\nabla \times (\nabla \times \mathcal{E}) = \nabla(\nabla \cdot \mathcal{E}) - \nabla^2 \mathcal{E} \tag{1.1}$$

combined with the third and fourth equations results in

$$\nabla^2 \mathcal{E} - \mu_0 \varepsilon_0 \partial_t^2 \mathcal{E} = 0 \text{ and } \nabla^2 \mathcal{H} - \mu_0 \varepsilon_0 \partial_t^2 \mathcal{H} = 0.$$

Hence, all components of the EM field satisfy the scalar wave equation

$$\Delta u - c_0^{-2} \partial_t^2 u = 0, \tag{1.2}$$

where we replaced ∇^2 by the notation Δ and used $c_0 = (\mu_0 \varepsilon_0)^{-1/2}$. Note that c_0 is the speed of light. Therefore, from now on we will consider the scalar wave u as either of components of electric $\mathcal{E}$ or magnetic $\mathcal{H}$ fields.

A common way to approach the wave equation (1.2) is through the Fourier transform of the wave u given by

$$u(s,t) = \int_{\mathbb{R}} e^{2\pi i\nu t} d\sigma_s(\nu). \tag{1.3}$$

The measure σ_s is called the spectral measure and it describes the frequency distribution in each point $s \in \mathbb{R}^3$ of the space. Since u is a physical quantity, it is real-valued and the spectral measure σ_r satisfies the symmetry condition

$$d\sigma_s(-\nu) = \overline{d\sigma_s(\nu)}. \tag{1.4}$$

In an idealized ptychographic experiment, σ_s is the Dirac delta measure corresponding to a single frequency $\nu \geq 0$. Such a wave is called *monochromatic* and in all other cases, the wave is referred to as *polychromatic*. A large part of this thesis is concerned with monochromatic ptychography. Polychromatic ptychography will be discussed in Chapter 5.

For the monochromatic illumination, the spectral representation of u is given by

$$u(s,t) = \frac{1}{2}\boldsymbol{u}(s)e^{2\pi i\nu t} + \frac{1}{2}\overline{\boldsymbol{u}(s)}e^{-2\pi i\nu t} = \mathrm{Re}(\boldsymbol{u}(s)e^{2\pi i\nu t}), \tag{1.5}$$

with the spectral density function $\boldsymbol{u} : \mathbb{R}^3 \to \mathbb{C}$. Due to one-to-one correspondence between the density $\boldsymbol{u}$ and the wave function u, it suffices to investigate the changes of $\boldsymbol{u}$ between starting and target planes.

Returning to the wave equation (1.2), the substitution of the representation (1.5) implies that the spectral density $\boldsymbol{u}$ is a solution of the Helmholtz equation

$$\Delta\boldsymbol{u} + k^2\boldsymbol{u} = 0, \tag{1.6}$$

where $k = 2\pi\nu/c_0$ denotes the wavenumber.

Without loss of generality, let the direction of propagation coincide with the z-axis and let the starting and target planes be perpendicular to z-axis. We assume that the starting plane is located at $z = 0$ and the target plane at $z = d$ for a fixed distance parameter $d > 0$. We will denote the projection of the density $\boldsymbol{u}$ on the corresponding planes as $\boldsymbol{u}_0 = \boldsymbol{u}\big|_{z=0}$ and $\boldsymbol{u}_d = \boldsymbol{u}\big|_{z=d}$. Assume that the behavior of the density $\boldsymbol{u}$ at the starting plane is a priory known and the values of $\boldsymbol{u}_d$ in the target plane are to be determined. Since the Helmholtz equation (1.6) with boundary condition $\boldsymbol{u}_0 = \boldsymbol{u}\big|_{z=0}$ is linear and shift invariant, the mapping $\boldsymbol{u}_0 \mapsto \boldsymbol{u}_d$ can be viewed as an input-output system. Thus, the transformation between the starting and target planes is given by

$$\boldsymbol{u}_d = \boldsymbol{h} * \boldsymbol{u}_0, \tag{1.7}$$

where $*$ denotes the convolution operation. The function $\boldsymbol{h}$ is called the impulse response function since for the pointwise impulse input

$$\boldsymbol{u}_0(\tilde{x},\tilde{y}) = \mathcal{I}_{(x,y)=(\tilde{x},\tilde{y})} := \begin{cases} 1, & \text{if } x = \tilde{x} \text{ and } y = \tilde{y}, \\ 0, & \text{otherwise,} \end{cases}$$

the system output will be $\boldsymbol{u}_d = \boldsymbol{h}(x,y)$ for all $x, y \in \mathbb{R}$. It is more convenient to study the input-output system in the Fourier domain related to the spatial coordinates. An

application of the Fourier transform (denoted by $\widehat{\ }$ or $\mathcal{F}$) from both sides combined with the convolution theorem [44, p. 8] leads to the equation

$$\hat{\boldsymbol{u}}_d(\nu_x, \nu_y) = \hat{\boldsymbol{h}}(\nu_x, \nu_y)\hat{\boldsymbol{u}}_0(\nu_x, \nu_y), \tag{1.8}$$

for all spatial frequencies $\nu_x, \nu_y \in \mathbb{R}$. The Fourier transform $\hat{\boldsymbol{h}}$ of the impulse response function also known as the transfer function, can be determined by considering a plane wave

$$\boldsymbol{u}(x, y, z) = e^{i(k_x x + k_y y + k_z z)}$$

with the spatial wave numbers satisfying $k_x^2 + k_y^2 + k_z^2 = k^2$. This condition implies that $\boldsymbol{u}$ is indeed a solution of the Helmholtz equation (1.6). Plugging $\boldsymbol{u}$ into (1.8) and performing some computations leads to

$$\hat{\boldsymbol{h}}(\nu_x, \nu_y) = e^{ik_z d} = e^{id\sqrt{k^2 - k_x^2 - k_y^2}} = e^{2\pi i d\sqrt{\nu^2 - \nu_x^2 - \nu_y^2}}.$$

The exact formula for $\boldsymbol{h}$, the inverse Fourier transform of the obtained function, has no known closed form. The common approach in physics is to use the Fresnel or small angle approximation

$$\sqrt{\nu^2 - \nu_x^2 - \nu_y^2} = \nu\left[1 - \frac{\nu_x^2 + \nu_y^2}{2\nu^2} + \frac{1}{8}\left(\frac{\nu_x^2 + \nu_y^2}{\nu^2}\right)^2 - \dots\right] \approx \nu - \frac{\nu_x^2 + \nu_y^2}{2\nu},$$

which is most accurate when $(\nu_x^2 + \nu_y^2)$ is significantly smaller than ν^2, so that the wave mainly propagates along the z-axis. The usage of the Fresnel approximation leads to the following expression for the transfer function

$$\hat{\boldsymbol{h}}(\nu_x, \nu_y) \approx e^{2\pi i d\nu} e^{-\frac{\pi i d}{\nu}(\nu_x^2 + \nu_y^2)}.$$

The application of the inverse Fourier transform gives us the approximation of the impulse transfer function by

$$\boldsymbol{h}(x, y) \approx \frac{-i\nu}{d} e^{2\pi i d\nu} e^{\frac{2\pi i\nu}{2d}(x^2 + y^2)}.$$

Finally, we are able to return back to the input-output relation (1.7), which gives us

$$\boldsymbol{u}_d(x, y) \approx \frac{-i\nu}{d} e^{2\pi i d\nu} \int_{\mathbb{R}^2} \boldsymbol{u}_0(\tilde{x}, \tilde{y}) e^{\frac{\pi i\nu}{d}((x-\tilde{x})^2 + (y-\tilde{y})^2)} d\tilde{x} d\tilde{y}.$$

Note that the squares in the last exponent can further be split into three parts,

$$(x - \tilde{x})^2 + (y - \tilde{y})^2 = (x^2 + y^2) - 2(x\tilde{x} + y\tilde{y}) + (\tilde{x}^2 + \tilde{y}^2).$$

When the target plane is sufficiently far away from the starting plane, in the so-called *far-field*, so that the fraction $\pi\nu(\tilde{x}^2 + \tilde{y}^2)/d$ is much smaller than 1, the third term can be neglected. Then, the formula for $\boldsymbol{u}_d$ is given by

$$\boldsymbol{u}_d(x, y) \approx \frac{-i\nu}{d} e^{2\pi i d\nu} e^{\frac{\pi i\nu}{d}(x^2 + y^2)} \int_{\mathbb{R}^2} \boldsymbol{u}_0(\tilde{x}, \tilde{y}) e^{\frac{-2\pi i\nu}{d}(x\tilde{x} + y\tilde{y})} d\tilde{x} d\tilde{y}.$$

After a change of variables in the integral, we observe that the input-output system of the Helmholtz equation with Fresnel assumption and the target plane in the far-field acts as a (scaled) Fourier transform in the spatial domain,

$$\boldsymbol{u}_d(s) \approx \frac{-i\nu}{d} e^{2\pi i d\nu} e^{\frac{\pi i\nu}{d}(x^2+y^2)} \mathcal{F}\boldsymbol{u}_0\left(\frac{\nu s}{d}\right), \tag{1.9}$$

for all $s = (x, y) \in \mathbb{R}^2$. We will use this approximation for the propagation of the light wave in free space.

1.2 Light interaction with object

If an EM wave passes through a medium, the wave will be modified according to the specific properties of the medium. In the process, it encodes the information about the object and this phenomenon is the foundation for imaging applications. Thus, it is important to properly describe the change of the EM wave. While it is possible to study the interaction process starting from Maxwell's equations similarly to free space, it requires involved investigation of the differential equations, which goes beyond the scope of this thesis. Instead, a more phenomological model is used to describe this process. It describes the exit wave $\boldsymbol{u}_e$ as a multiplication of the incoming wave $\boldsymbol{u}_i$ by an *object transfer function* $\boldsymbol{x}$, that means

$$\boldsymbol{u}_e(s) = \boldsymbol{x}(s)\boldsymbol{u}_i(s), \quad s \in \mathbb{R}^2. \tag{1.10}$$

This model is known as multiplication assumption and is commonly used in imaging applications [45, 46, 4].
The object transfer function $\boldsymbol{x}$ models two phenomena, the *absorption* and the *refraction* of light, which are represented by the amplitude $|\boldsymbol{x}|$ and the phases of $\boldsymbol{x}/|\boldsymbol{x}|$. Both the absorption and the refraction characterize the object locally and provide its spatial representation, which is the main interest in the ptychographic experiment.
As illumination reaches the specimen, the light may be absorbed by the object, which causes the reduction of energy in the exit wave. It is more likely to happen in the denser areas of the specimen, which links absorption to the atomic density. In particular, if the light is absorbed completely, in (1.10) it corresponds to $|\boldsymbol{x}(s)| = 0$, and if nothing is absorbed, then $|\boldsymbol{u}_e(s)| = |\boldsymbol{u}_i(s)|$, so that $|\boldsymbol{x}(s)| = 1$.
Refraction, on the other hand, represents the change in the direction of propagation of light caused by the scattering effects. In (1.10), this is represented by the multiplication of the incoming wave $\boldsymbol{u}_i(s)$ with $\boldsymbol{x}(s)/|\boldsymbol{x}(s)|$.

1.3 Intensity of light

The optical intensity $\boldsymbol{I}$ of the wave u is proportional to the time average of the squared wave function,

$$\boldsymbol{I}(s) \propto \langle |u(s,t)|^2 \rangle = \frac{1}{2T} \int_{-T}^{T} |u(s,t)|^2 dt,$$

where $T > 0$ is the exposure time. For the monochromatic wave u with spectral representation (1.5), we obtain

$$\begin{aligned}\frac{1}{2T}\int_{-T}^{T}|u(s,t)|^2dt &= \frac{1}{2T}\int_{-T}^{T}\frac{1}{4}|\boldsymbol{u}(s)e^{2\pi i\nu t}+\overline{\boldsymbol{u}(s)}e^{-2\pi i\nu t})|^2dt\\ &= \frac{1}{2T}\int_{-T}^{T}\frac{1}{2}|\boldsymbol{u}(s)|^2+\frac{1}{4}\boldsymbol{u}^2(s)e^{4\pi i\nu t}+\frac{1}{4}\overline{\boldsymbol{u}}^2(s)e^{-4\pi i\nu t}dt\\ &= \frac{1}{2}|\boldsymbol{u}(s)|^2+\frac{1}{4}(\boldsymbol{u}^2(s)+\overline{\boldsymbol{u}}^2(s))\,\mathrm{sinc}(4\pi\nu T),\end{aligned}$$

where $\mathrm{sinc}(t) := \sin(t)/t$. If T is much larger than $1/\nu$, the sinc function and the second term vanish. Consequently, the optical intensity of the monochromatic wave u is proportional to

$$\boldsymbol{I}(s) \propto |\boldsymbol{u}(s)|^2. \tag{1.11}$$

1.4 Formula of diffraction pattern

Now, we are equipped for the mathematical description of the ptychographic experiment and resulting diffraction patterns. Let us start by restating the selected coordinate system in which z axis is aligned with the propagation of the light and axes x and y are perpendicular to it. The object plane coincides with the x-y plane with $z = 0$. The detector plane is parallel to the object plane with $z = p$, with the distance $p > 0$ sufficiently large to satisfy the far-field assumption.
In this section, we assume that the window is already formed by the aperture and start directly from its interaction with the object. Let w be the monochromatic light wave representing the window with the spectral density in the object plane $\boldsymbol{w} : \mathbb{R}^2 \to \mathbb{C}$. The object transfer function is denoted by $\boldsymbol{x} : \mathbb{R}^2 \to \mathbb{C}$. As the object is being shifted after each illumination, its shift is represented mathematically via the family of the translation operators $\mathcal{T}_r$ as

$$\mathcal{T}_r\boldsymbol{u}(s) = \boldsymbol{u}(s-r), \text{ with the dislocation parameter } r = (r_x, r_y) \in \mathbb{R}^2,$$

for all $s \in \mathbb{R}^2$. Hence, by (1.10), the exit wave corresponding to the shift position $r \in \mathbb{R}^2$ of the object $\boldsymbol{x}$ illuminated by the window w is given by

$$\boldsymbol{u}_e(s) = (\mathcal{T}_{-r}\boldsymbol{x})(s)\boldsymbol{w}(s).$$

The next step is the propagation of the exit wave from the object to the detector plane. By (1.9), the spectral density of the exit wave in the detector plane is approximated by

$$\boldsymbol{u}_p(s) \approx \frac{-i\nu}{p}e^{2\pi ip\nu}e^{\frac{\pi i\nu}{p}(x^2+y^2)}\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}]\left(\frac{\nu s}{p}\right), \quad s = (x,y) \in \mathbb{R}^2.$$

The detector can only capture the energy of the incoming wave, which is proportional to the squared magnitude $|\boldsymbol{u}_p(x,y)|^2$. Furthermore, the detector doesn't record the energy instantly and only captures the average energy over the period of exposure time, which

is precisely the optical intensity of the wave u_p. We recall that by (1.11), the intensity of u_p is proportional to

$$\boldsymbol{I}(r,s) \propto |\boldsymbol{u}_p(s)|^2 = \left| \frac{\nu}{p} \mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}] \left(\frac{\nu s}{p} \right) \right|^2, \quad s \in \mathbb{R}^2. \tag{1.12}$$

The obtained formula is the general representation of the intensity observed at the detector plane. Finally, the proportional factor does not play a role in the reconstruction process and will be, therefore, ignored.

1.5 Circular aperture

In this section, we derive the mathematical description of the window if a circular aperture is used for the localization of the light. In terms of our mathematical setup, let the aperture plane be placed in parallel to the object plane at $z = -p_a < 0$. We will assume that the distance p_a is sufficiently large so that the far-field assumption applies. Furthermore, the center of the circular aperture is at $(x, y) = (0, 0)$ and its radius is denoted by $R > 0$.
In a perfect experimental setup, the source generates a plane monochromatic wave aligned with the z-axis, which is given by

$$u(x,y,z,t) = \frac{1}{2} e^{2\pi i(\nu t + k(z+p_a))} + \frac{1}{2} e^{-2\pi i(\nu t + k(z+p_a))},$$

for some frequency ν and corresponding wavenumber k. The spectral density of the wave u at $z = -p_a$ is given by

$$\boldsymbol{u}(x,y) = 1.$$

The interaction of the wave u with the circular aperture leads to the spectral density of the form

$$\boldsymbol{u}_l(x,y) = \boldsymbol{u}(x,y)\mathcal{I}_{x^2+y^2 \le R^2} = \mathcal{I}_{x^2+y^2 \le R^2}. \tag{1.13}$$

Next, the wave propagates from the aperture to the object, and the spectral density $\boldsymbol{w}$ of the window is approximated by (1.9),

$$\begin{aligned} \boldsymbol{w}(x,y) &\approx \frac{-i\nu}{p_a} e^{2\pi i p_a \nu} e^{\frac{\pi i \nu}{p_a}(x^2+y^2)} \mathcal{F}[\boldsymbol{u}_l] \left(\frac{\nu x}{p_a}, \frac{\nu y}{p_a} \right) \\ &\propto \frac{-i\nu}{p_a} e^{2\pi i p_a \nu} e^{\frac{\pi i \nu}{p_a}(x^2+y^2)} \frac{J_1(\pi \nu R \sqrt{x^2+y^2}/p_a)}{\pi \nu R \sqrt{x^2+y^2}/p_a}, \end{aligned} \tag{1.14}$$

where J_1 is the Bessel function of the first kind. Note that $\boldsymbol{w}$ is a radial function. The behavior of the function $J_1(r)/r$ is shown in Figure 1.2. The intensity of the window $\boldsymbol{w}$ corresponding to the circular aperture is referred to as the Airy disc.

Remark 1.5.1. *Note that in the example above, the truncation of the wave by the aperture is a crude approximation. As the wave propagates in all directions, the wave $\boldsymbol{u}_a$ is non-zero along the aperture plane so that the equation (1.13) is not valid as well as the equation (1.14) for $\boldsymbol{w}$. A correct derivation of the window $\boldsymbol{w}$ requires solving the*

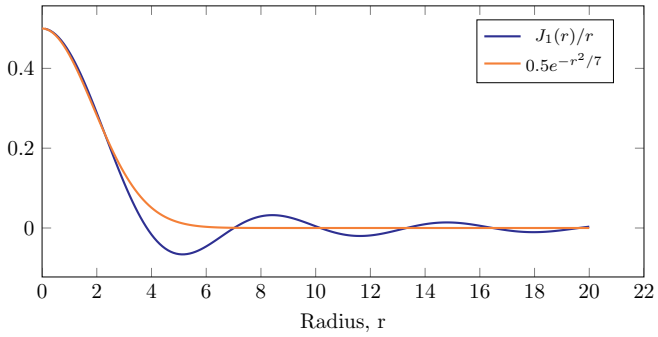


Figure 1.2: Fourier transform of the circular aperture $J_1(r)/r$ and Gaussian bell $0.5e^{-r^2/7}$ as a functions of the radius r.

Maxwell's equations in free space with boundary conditions corresponding to the shape of the aperture known as Rayleigh-Sommerfeld diffraction formula. Nevertheless, both the heuristic approach above and the correct derivation of $\boldsymbol{w}$ with circular aperture coincide, which is not true in general.

Often, only the bright spot in the middle of the Airy disc is of interest and the varying tails can be neglected. In such cases, the main peak of $J_1(r)/r$ is approximated by the Gaussian bell $0.5e^{-r^2/7}$ (see Figure 1.2). Thus, the window is approximated by

$$\boldsymbol{w}_g(x,y) \propto \frac{-i\nu}{p_a} e^{2\pi i p_a \nu} e^{\frac{\pi i \nu}{p_a}(x^2+y^2)} e^{-\frac{\pi^2\nu^2R^2(x^2+y^2)}{7p_a^2}}. \tag{1.15}$$

Chapter 2

Preliminaries and notation

In this chapter, we provide the notation and basic results, which will be used throughout the thesis. Some section-specific definitions will appear later when required.

2.1 Sets, vectors and matrices

We start by recalling the standard definitions involving vectors, matrices and sets. For more detailed material we refer the reader to [47, 48].
The *complex unit* will always be denoted by i. For a complex number $z = \alpha + i\beta$, $\alpha, \beta \in \mathbb{R}$, its *conjugate* is given by $\overline{z} = \alpha - i\beta$. The number $\alpha = \mathrm{Re}(z)$ is called the real part of z and $\beta = \mathrm{Im}(z)$ is the imaginary part. The *absolute value* (magnitude, amplitude) of a complex number is $|z| = \sqrt{\alpha^2 + \beta^2}$ and the *phase* of a non-zero complex number is given by $z/|z|$. For $z = 0$ the phase is set to 1. The set of all complex numbers with magnitude one will be denoted by $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$.
In this thesis, we will mainly work with a-dimensional vectors in either real or complex vector spaces $\mathbb{R}^a$ and $\mathbb{C}^a$, respectively. For our convenience, the entries of the vectors are indexed from 0 to $a - 1$ and we will use the notation $[a] := \{0, \ldots, a - 1\}$ for index sets. The vector in $\mathbb{C}^a$ containing only zero entries is denoted by $\mathbb{0}_a$ and the vector with all entries equal to one is denoted by $\mathbb{1}_a$. A vector $v \in \mathbb{C}^a$ is called *non-vanishing* if all its entries are non-zero. The *support* of v is the set of indices corresponding to non-zero entries of the vector v, i.e.,

$$\mathrm{supp}(v) = \{j \in [a] : v_j \neq 0\}.$$

The *span* of vectors $v_0, \ldots, v_{K-1}$ in $\mathbb{C}^a$ is the subspace of $\mathbb{C}^a$ containing all possible linear combinations of these vectors,

$$\mathrm{span}\{v_k,\ k \in [K]\} = \left\{ \sum_{k \in [K]} \alpha_k v_k :\ \alpha \in \mathbb{C}^K \right\}.$$

For double indexed objects, it is convenient to use the set product notation, that is for index sets $\mathcal{J}_1$ and $\mathcal{J}_2$, their product is given by

$$\mathcal{J}_1 \times \mathcal{J}_2 := \{(j_1, j_2) : j_1 \in \mathcal{J}_1, j_2 \in \mathcal{J}_2\}.$$

In case $\mathcal{J}_1 = \mathcal{J}_2 = \mathcal{J}$, we may also write $\mathcal{J}^2$. The *cardinality* of the set $\mathcal{J}$ is denoted by $|\mathcal{J}|$.
For $p \geq 1$, the ℓ_p*-norm* of a vector $v \in \mathbb{C}^a$ is defined as

$$\|v\|_p = \begin{cases} \left(\sum_{k\in[a]} |v_k|^p\right)^{1/p}, & 1 \leq p < \infty, \\ \max_{k\in[a]} |v_k|, & p = \infty. \end{cases}$$

The *inner product* of two vectors $u, v \in \mathbb{C}^a$ is given by

$$\langle u, v\rangle = v^* u = \sum_{k\in[a]} u_k \overline{v}_k.$$

If $\langle u, v\rangle = 0$, the vectors u and v are *orthogonal*, which is denoted by $u \perp v$. The *standard basis* $\{e_k\}_{k\in[a]}$ of $\mathbb{C}^a$ are the vectors with entries $(e_k)_j = 1$, if $k = j$ and $(e_k)_j = 0$ otherwise.
The *projection of* $u \in \mathbb{C}^a$ *onto a set* $\mathcal{S} \subseteq \mathbb{C}^a$ is an element $\tilde{u} \in \mathcal{S}$ fulfilling

$$\|u - \tilde{u}\|_2 \leq \|u - v\|_2$$

for all $v \in \mathcal{S}$. The operator, which maps u to $\tilde{u}$ is called the *projection operator onto* $\mathcal{S}$. In general, projection $\tilde{u}$ is not-unique, however, if $\mathcal{S}$ is a non-empty closed convex set, $\tilde{u}$ can be uniquely identified [49].
The space of complex $a \times b$ matrices is denoted by $\mathbb{C}^{a\times b}$. For a matrix $B \in \mathbb{C}^{a\times b}$, we will refer to its entries as $B_{j,k}$, $(j,k) \in [a] \times [b]$. The vector denoting the j-th *row* of B is denoted by $B_{(j)}$ and the vector denoting the k-th *column* is denoted by $B^{(k)}$.
We will also denote the *kernel* and the *image* of B as

$$\ker B := \{v \in \mathbb{C}^b : Bv = 0\},$$
$$\mathrm{im}(B) := \{Bv \in \mathbb{C}^a : v \in \mathbb{C}^b\},$$

respectively.
For $\ell \in [a]$ the ℓ-th *diagonal* of the square matrix $B \in \mathbb{C}^{a\times a}$ is given by

$$d^\ell(B)_j = B_{j,j-\ell \bmod a}, \quad j \in [a].$$

Larger values of ℓ correspond to the diagonals further below the main diagonal $d^0(B)$ and we will also use negative values of ℓ for the reverse order from the main diagonal $d^0(B)$.
The *rank* $\mathrm{rank}(B)$ of the matrix B is the dimension of the vector space generated by the columns of B. The *trace* of a square matrix $B \in \mathbb{C}^{a\times a}$ is the sum of its diagonal entries

$$\mathrm{tr}(B) = \sum_{k\in[a]} B_{k,k}.$$

The *transpose* and *complex conjugate transpose* of a vector v or a matrix B are denoted by v^T, v^* and B^T, B^*, respectively. The matrix in $\mathbb{C}^{a\times b}$ containing only zero entries is denoted by $O_{a\times b}$ and with all entries equal to one - by $\mathbb{1}_{a\times b}$.
The *Frobenius (Hilbert-Schmidt) inner product* of two matrices $U, V \in \mathbb{C}^{a\times b}$ is given by

$$\langle U, V\rangle_F := \mathrm{tr}(V^* U) = \sum_{(k,j)\in[a]\times[b]} U_{k,j} \overline{V}_{k,j}.$$

This inner product induces the *Frobenius (Hilbert-Schmidt) norm*

$$\|U\|_F = \sqrt{\langle U, U\rangle_F} = \sqrt{\sum_{(k,j)\in[a]\times[b]} |U_{k,j}|^2}.$$

For a vector $v \in \mathbb{C}^a$, the *diagonal matrix* $\mathrm{diag}(v) \in \mathbb{C}^{a\times a}$ is formed by placing the entries of the vector v onto the main diagonal, so that for $k, j \in [a]$ it holds that

$$\mathrm{diag}(v)_{k,j} := \begin{cases} v_k & k = j, \\ 0 & k \neq j. \end{cases}$$

In many cases, we will apply *entrywise operations* to vectors or matrices. For a function $f : \mathbb{C} \to \mathbb{C}$, we will understand $f(u)$ as a function f applied to each entry of u, $f(u)_j = f(u_j)$. The common examples of the entrywise operations are $|u|$, u^2 and $\operatorname{sgn} u$, where $\operatorname{sgn}(z) = z/|z|$ if $|z| \neq 0$ and 1 otherwise. Sometimes we would use an alternative version of the sign function, $\operatorname{sgn}_0(z)$, which is equal to $\operatorname{sgn}(z)$ for $z \neq 0$ and $\operatorname{sgn}_0(0) = 0$.
In the text, we will sometimes denote by $\cdot$ the argument of the function, with respect to which some transformation is performed, e.g. we will use the notation $f(-\cdot)$ for function $f(-s)$.
An *indicator function* $\mathcal{I}_{predicate}$ is a binary function, which is equal to 1 if the predicate is true and 0 otherwise. The rounding up or down operations will be respectively denoted by $\lceil a \rceil$ and $\lfloor a \rfloor$ a real number a.
The entrywise operations also apply to the arithmetic actions with vectors. In particular, the *entrywise (Hadamard) product* of u and v in $\mathbb{C}^a$ is given by

$$(u \circ v)_j = u_j v_j, \quad j \in [a].$$

The Hadamard product of two vectors u and v can be also presented as multiplication with a diagonal matrix, that is $u \circ v = \mathrm{diag}(u)v = \mathrm{diag}(v)u$.
The *entrywise division* of $u \in \mathbb{C}^a$ and non-vanishing $v \in \mathbb{C}^a$ is given by

$$(u/v)_j = u_j/v_j, \quad j \in [a].$$

Using the entrywise operations, we obtain the following properties of diagonal matrices,

$$\mathrm{diag}(u)\,\mathrm{diag}(v) = \mathrm{diag}(u \circ v), \quad \mathrm{diag}(u) + \mathrm{diag}(v) = \mathrm{diag}(u + v), \quad (\mathrm{diag}(u))^* = \mathrm{diag}(\overline{u}),$$
$$\mathrm{diag}(u)^{-1} = \mathrm{diag}(1/u), \text{ when } u_j \neq 0 \text{ for all } j \in [a].$$

We say that the mapping $f : \mathbb{C}^a \to \mathbb{C}^b$ is *injective* if for all pairs of vectors $u, v \in \mathbb{C}^a$ with $u \neq v$ it holds that $f(u) \neq f(v)$. The matrix $B \in \mathbb{C}^{b\times a}$ is *injective* if the corresponding mapping $v \mapsto Bv$ is injective. The injectivity of B is equivalent to the condition $\mathrm{rank}(B) = a$.
For a square matrix $B \in \mathbb{C}^{a\times a}$ with $\mathrm{rank}(B) = a$, its *inverse* $B^{-1} \in \mathbb{C}^{a\times a}$ satisfies $B^{-1}B = BB^{-1} = I_a$, where $I_a = \mathrm{diag}(\mathbb{1}_a)$ is the *identity matrix* in $\mathbb{C}^{a\times a}$. The matrix $B \in \mathbb{C}^{b\times a}, b \geq a$, is called *orthogonal* if it satisfies $B^*B = I_a$. The square orthogonal matrix $B \in \mathbb{C}^{a\times a}$ is a *unitary* matrix and its inverse is $B^{-1} = B^*$.

The square matrix $B \in \mathbb{C}^{a\times a}$ is called *Hermitian* if $B = B^*$ and the space of all Hermitian $a \times a$ matrices is denoted by $\mathbb{H}^a$. A Hermitian $a \times a$ matrix admits an *eigenvalue decomposition*

$$B = U\Lambda U^*,$$

where U is an $a \times a$ unitary matrix and Λ is an $a \times a$ real diagonal matrix. The entries λ_j, $j = 1, \ldots, a$, of the main diagonal of Λ are called *eigenvalues* and the corresponding columns $U^{(j)}$ are referred to as the *eigenvectors*. The eigenvalues are sorted in decreasing order and the number of non-zero eigenvalues coincides with the rank of the matrix B. The eigenvector corresponding to the largest eigenvalue is referred to as *top eigenvector*. In this thesis, we will require an error estimate for the recovery of the eigenvector of a rank-one matrix corrupted by noise.

Lemma 2.1.1 ([40, Lemma A.2])**.** *Let $u \in \mathbb{C}^a$ and $U = uu^*$. Consider $V \in \mathbb{H}^a$ and its largest magnitude eigenvalue λ and the corresponding eigenvector v, $\|b\|_2 = 1$. Then,*

$$\min_{|\alpha|=1} \left\| u - \alpha\sqrt{|\lambda|}v \right\|_2 \leq \frac{(1+2\sqrt{2})\,\|U-V\|_F}{\|u\|_2}.$$

A Hermitian matrix $B \in \mathbb{C}^{a\times a}$ is called *positive semidefinite* if for all $v \in \mathbb{C}^a$ it holds that $v^*Bv \geq 0$, which is equivalent to $\lambda_j(B) \geq 0$ for all $j = 1, \ldots, a$. Each positive semidefinite matrix admits decomposition $B = CC^*$ and for each $C \in \mathbb{C}^{a\times b}$ matrices CC^* and C^*C are positive semidefinite.
The analogue of the eigenvalue decomposition for non-square matrices is the *singular value decomposition* (SVD). For a matrix $B \in \mathbb{C}^{b\times a}$ its singular value decomposition is given by

$$B = U\Sigma V^*,$$

where $U \in \mathbb{C}^{b\times b}, V \in \mathbb{C}^{a\times a}$ are unitary matrices and $\Sigma \in \mathbb{R}^{b\times a}$ is a matrix with diagonal entries $\sigma_j(B) \geq 0$, $j = 1, \ldots, \min\{a, b\}$, sorted in decreasing order. The values $\sigma_j(B)$ are referred to as the *singular values* of the matrix B and the corresponding columns of U and V as *left and right singular vectors*. The singular values σ_j and corresponding columns $U^{(j)}$ correspond to the square root of eigenvalues of BB^* and their eigenvectors. Analogously, σ_j and columns $V^{(j)}$ correspond to the roots of eigenvalues of B^*B and their eigenvectors. Moreover, if B is Hermitian, then the singular values are the magnitudes of eigenvalues, sorted in decreasing order. If a matrix B has rank r, it holds that $\sigma_r(B) > 0$ and $\sigma_{r+1}(B) = 0$. Thus, it is sometimes more convenient to work with the SVD of B in the form $B = U\Sigma V^*$, where $U \in \mathbb{C}^{b\times r}, V \in \mathbb{C}^{a\times r}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{r\times r}$ is an invertible diagonal matrix with diagonal entries $\sigma_j(B) > 0$, $j = 1, \ldots, r$, sorted in decreasing order.
For $p \geq 1$, *the Schatten-p norm* of a matrix B is given by the ℓ_p-norm of the vector $(\sigma_1, \ldots, \sigma_r)$, that is

$$\|B\|_p = \|(\sigma_1, \ldots, \sigma_r)\|_p .$$

We will use the notation $\|\cdot\|_p$ to denote the Schatten-p norm of a matrix and the same notation for ℓ_p-norm of a vector. Since vectors are denoted by small letters and matrices by capital letters, it should be clear which norm is applied.

By definition, the Schatten-∞ norm is the largest singular value $\sigma_1(B)$ and equals to the spectral norm $\|B\|$ of a matrix B defined as

$$\|B\|_\infty := \max_{v\in\mathbb{C}^a, \|v\|=1} \|Bv\|_2 .$$

Furthermore, the Schatten-2 norm coincides with the Frobenius norm of a matrix B, that is $\|B\|_2 = \|B\|_F$, and we will only use the notation $\|B\|_F$.
Note that the singular values of a diagonal matrix $\mathrm{diag}(u)$ are given by $|u|$ and, therefore, the spectral norm is equal to $\|u\|_\infty$ and the Frobenius norm – to $\|u\|_2$.
Using the singular value decomposition, the *Moore-Penrose pseudoinverse* of the matrix B is defined as

$$B^\dagger := V\Sigma^{-1}U^*.$$

If B is a square invertible matrix, its pseudoinverse $B^\dagger$ coincides with B^{-1}. For an injective matrix $B \in \mathbb{C}^{b\times a}, b \geq a$, its pseudoinverse $B^\dagger$ can be expressed as

$$B^\dagger = (B^*B)^{-1}B^*. \tag{2.1}$$

It satisfies

$$B^\dagger B = I \text{ and } BB^\dagger \text{ is a projection operator onto the set } \mathrm{im}(B). \tag{2.2}$$

For two matrices $A \in \mathbb{C}^{a\times b}$ and $B \in \mathbb{C}^{c\times d}$, the *tensor product* $A \otimes B \in \mathbb{C}^{ab\times cd}$ is defined as

$$A \otimes B = \begin{bmatrix} A_{0,0}B & A_{0,1}B & \dots & A_{0,b-1}B \\ A_{1,0}B & A_{1,1}B & \dots & A_{1,b-1}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{a-1,0}B & A_{a-1,1}B & \dots & A_{a-1,b-1}B \end{bmatrix}. \tag{2.3}$$

Proposition 2.1.2 (Properties of tensor product)**.** *Let $A, B \in \mathbb{C}^{a\times b}$, $C, D \in \mathbb{C}^{c\times d}$, $E \in \mathbb{C}^{b\times e}$, $F \in \mathbb{C}^{d\times f}$ and $\alpha, \beta \in \mathbb{C}$.*

1. *The tensor product is bilinear,*

$$(\alpha A+\beta B)\otimes C = \alpha(A\otimes C)+\beta(B\otimes C) \quad \text{and} \quad A\otimes(\alpha C+\beta D) = \alpha(A\otimes C)+\beta(A\otimes D).$$

2. *The tensor product distributes over matrix multiplication,*

$$(A \otimes C)(E \otimes F) = AE \otimes CF.$$

3. *If the SVD of A is given by $U_1\Sigma_1V_1^*$ and of B by $U_2\Sigma_2V_2^*$, then the SVD of $A \otimes B$ is given by*

$$A \otimes B = (U_1 \otimes U_2)(\Sigma_1 \otimes \Sigma_2)(V_1 \otimes V_2)^*.$$

Consequently, the spectral and Frobenius norms of $A \otimes B$ admit

$$\|A \otimes B\|_\infty = \|A\|_\infty \|B\|_\infty \quad \text{and} \quad \|A \otimes B\|_F = \|A\|_F \|B\|_F .$$

In order to count the number of operations, we will use *big O notation* $\mathcal{O}(a)$, by which we mean that at most ca operations are required for some constant $c > 0$.

2.2 Fourier transform and related operators

The *discrete Fourier transform* of a vector $v \in \mathbb{C}^a$ is defined as

$$(F_a v)_k := \sum_{j\in[a]} v_j e^{-\frac{2\pi ijk}{a}}, \quad k \in [a]. \tag{2.4}$$

It corresponds to the matrix-vector multiplication $F_a v$ with the entries of the matrix $F_a \in \mathbb{C}^{a\times a}$ given by

$$(F_a)_{k,j} = e^{-\frac{2\pi ijk}{a}}, \quad j,k \in [a].$$

We will refer to the vector $F_a v$ as the *frequencies* of the vector v or the representation of v in the *frequency domain*.
The *inverse discrete Fourier transform* of a vector $v \in \mathbb{C}^a$ is defined as

$$(F_a^{-1} v)_k := \frac{1}{a}\sum_{j\in[a]} v_j e^{\frac{2\pi ijk}{a}}, \quad k \in [a]. \tag{2.5}$$

The corresponding matrix F_a^{-1} with entries $(F_a^{-1})_{k,j} = \frac{1}{a}e^{-\frac{2\pi ijk}{a}}$, $j,k \in [a]$, is indeed an inverse of F_a. Basic properties of the discrete Fourier transform are summarized in the next proposition.

Proposition 2.2.1. *Let $\alpha, \beta \in \mathbb{C}$, $r \in [a]$, $u, v \in \mathbb{C}^a$.*

1. *The Fourier transform is a linear operation,*

$$F_a(\alpha u + \beta v) = \alpha F_a u + \beta F_a v.$$

2. *The inverse Fourier transform matrix satisfies $F_a^{-1} = \frac{1}{a}F_a^*$ and we have*

$$F_a^* F_a = F_a F_a^* = aI_a.$$

3. *The matrices $\frac{1}{\sqrt{a}}F_a$ and $\sqrt{a}F_a^{-1}$ are unitary matrices and the Plancherel identity holds,*

$$\|F_a u\|_2^2 = a\|u\|_2^2.$$

The family of the *circular shift operators* $S_r : \mathbb{C}^a \to \mathbb{C}^a, r \in \mathbb{Z}$, is defined as

$$(S_r u)_j := u_{j-r \bmod a} \text{ for all } j \in [a],\ u \in \mathbb{C}^a, \tag{2.6}$$

and the family of the *modulation operators* $M_r : \mathbb{C}^a \to \mathbb{C}^a, r \in [a]$, is given by

$$(M_r u)_j := e^{\frac{2\pi ijr}{a}} u_j \text{ for all } j \in [a], u \in \mathbb{C}^a. \tag{2.7}$$

The corresponding matrix representations of S_r and M_r are

$$(S_r)_{k,j} = \mathcal{I}_{k=j-r \bmod a}, \quad \text{and} \quad (M_r)_{k,j} = e^{\frac{2\pi ijr}{a}}\mathcal{I}_{k=j}.$$

These operators are closely related by the Fourier transform and we state their main properties in the next proposition

Proposition 2.2.2. *Let $r \in \mathbb{Z}$ be arbitrary.*

1. *Both shift and modulo matrices are unitary and their adjoint/inverse satisfy*
$$S_r^* = S_r^{-1} = S_{-r} \text{ and } M_r^* = M_r^{-1} = M_{-r}.$$
2. *The modulation is the shift in the frequency domain, that is*
$$F_a S_r = M_{-r} F_a.$$

The *time reversal* operator R_a transforms the vector $v \in \mathbb{C}^a$ as

$$(R_a v)_j = v_{-j \bmod a}, \quad j \in [a], \tag{2.8}$$

with matrix representation

$$(R_a)_{k,j} = \begin{cases} 1, & k = -j \bmod a, \\ 0, & \text{otherwise.} \end{cases}$$

The *circular convolution* of two vectors $u, v \in \mathbb{C}^a$ is given by

$$(u *_a v)_k = \sum_{j \in [a]} u_j v_{k-j \bmod a}, \quad k \in [a]. \tag{2.9}$$

The circular convolution acts as an entrywise product of the frequencies of two vectors.

Theorem 2.2.3 (Circular convolution theorem)**.** *For all $u, v \in \mathbb{C}^a$ we have*

$$F_a(u *_a v) = (F_a u) \circ (F_a v) \quad \text{and} \quad a F_a(u \circ v) = (F_a u) *_a (F_a v).$$

The circular convolution can be represented by a matrix-vector product $C_u v$ with

$$C_u = \begin{bmatrix} u_0 & u_{a-1} & u_{a-2} & \dots & u_2 & u_1 \\ u_1 & u_0 & u_{a-1} & \dots & u_3 & u_2 \\ u_2 & u_1 & u_0 & \dots & u_4 & u_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{a-2} & u_{a-3} & u_{a-4} & \dots & u_0 & u_{a-1} \\ u_{a-1} & u_{a-2} & u_{a-3} & \dots & u_1 & u_0 \end{bmatrix}.$$

A matrix of this form is called a *convolution or circulant matrix* and it satisfies

$$(C_u)_{j,k} = u_{j-k \bmod a}.$$

The consequence of Theorem 2.2.3 is that all convolution matrices admit the following eigendecomposition.

Theorem 2.2.4 (Decomposition of circulant matrices)**.** *Let $C_u \in \mathbb{C}^{a \times a}$ be a circulant matrix. Then,*

$$C_u = \left(\frac{1}{\sqrt{a}} F_a\right)^* \operatorname{diag}(a F_a^{-1}[R_a u]) \left(\frac{1}{\sqrt{a}} F_a\right) = F_a^* \operatorname{diag}(F_a^{-1}[(C_u)_{(0)}]) F_a,$$

where F_A is the discrete Fourier transform (2.5) and $(C_u)_{(0)}$ denotes the 0-th row of the matrix C_u.

Returning to the time reversal operator, it possesses the following properties.

Proposition 2.2.5. *1. The time reversal matrix is Hermitian and unitary, that is*

$$R_a^* = R^{-1} = R_a.$$

2. The time reversal and the discrete Fourier transform commutate,

$$F_a R_a = R_a F_a,$$

and, furthermore, the following identities hold true,

$$F_a F_a = aR_a, \quad F_a\overline{u} = R_a\overline{F_a u}, \quad \text{and} \quad |F_a u|^2 = F_a(u *_a R_a\overline{u}).$$

3. Lastly, it satisfies

$$R_a S_r = S_{-r} R_a.$$

For dimensions $b \leq a$, the *projection operator* $P_b : \mathbb{C}^a \to \mathbb{C}^b$ is defined as

$$(P_b u)_j = u_j, \tag{2.10}$$

for all $j \in [b], u \in \mathbb{C}^a$, and it is represented by a $b \times a$ matrix

$$(P_b)_{j,k} = \mathcal{I}_{j=k}, \quad j \in [b], k \in [a].$$

The adjoint of P_b is the *embedding operator*, which appends a vector in $\mathbb{C}^b$ with $a - b$ zeros, so that

$$(P_b^* v)_k = \begin{cases} v_k & k \in [b], \\ 0, & k \notin [b], \end{cases} \text{ for all } k \in [a], v \in \mathbb{C}^b.$$

The matrix P_b^* is orthogonal, i.e., it satisfies

$$(P_b^*)^* P_b^* = P_b P_b^* = I_b \text{ and } P_b^* P_b = \begin{bmatrix} I_b & O_{b\times(a-b)} \\ O_{(a-b)\times b} & O_{(a-b)\times(a-b)} \end{bmatrix}, \tag{2.11}$$

so that for all v with $\operatorname{supp}(v) \subset [b]$ we have

$$P_b^* P_b v = v. \tag{2.12}$$

2.3 Wirtinger derivatives

2.3.1 Definitions

In many sections of this thesis, we will perform first order optimization of real-valued functions of complex variables, which will be based on Wirtinger derivatives [50]. In this section we recall some basic facts about the Wirtinger derivatives based on [51, 52]. A function $f : \mathbb{C} \mapsto \mathbb{C}$ can be viewed as a function of two real variables, the real and imaginary parts of the argument $z = \alpha + i\beta$. The function f is said to be differentiable (in real sense) if the derivatives with respect to α and β exist.

Then, the Wirtinger derivatives are defined as

$$\frac{\partial f}{\partial z} := \frac{1}{2}\left(\frac{\partial f}{\partial \alpha} - i\frac{\partial f}{\partial \beta}\right), \quad \frac{\partial f}{\partial \bar{z}} := \frac{1}{2}\left(\frac{\partial f}{\partial \alpha} + i\frac{\partial f}{\partial \beta}\right),$$

which is nothing else but a change of the coordinate system to the conjugate coordinates. In this sense, we treat the function f as a function of z and $\bar{z}$ instead of α and β.

Example 2.3.1. *Consider $f(z) = z = \alpha + i\beta$. Its Wirtinger derivatives are*

$$\frac{\partial z}{\partial z} = \frac{1}{2}\frac{\partial(\alpha + i\beta)}{\partial \alpha} - \frac{i}{2}\frac{\partial(\alpha + i\beta)}{\partial \beta} = \frac{1}{2} - \frac{i^2}{2} = 1 \text{ and } \frac{\partial z}{\partial \bar{z}} = 0.$$

The obtained inequalities imply that $\bar{z}$ can be treated as a constant when the derivative with respect to z is computed and vice versa.

Similarly to the real analysis of multivariate functions, the Wirtinger derivatives are extended for $f : \mathbb{C}^d \mapsto \mathbb{C}$, that is for $z \in \mathbb{C}^d$ they are given by

$$\frac{\partial f}{\partial z} = \left(\frac{\partial f}{\partial z_1}, \ldots, \frac{\partial f}{\partial z_d}\right) \quad \text{and} \quad \frac{\partial f}{\partial \bar{z}} = \left(\frac{\partial f}{\partial \bar{z}_1}, \ldots, \frac{\partial f}{\partial \bar{z}_d}\right).$$

The computation of the Wirtinger derivatives is analogous to the standard real analysis as the arithmetic operations and the chain rule extends to the complex case.

Theorem 2.3.2 (Properties of Wirtinger derivatives)**.** *Let $f, g : \mathbb{C}^d \mapsto \mathbb{C}$, $h : \mathbb{C} \mapsto \mathbb{C}$ be differentiable functions and let $\alpha, \beta \in \mathbb{C}$. Then, the following derivation rules apply:*

1. *Arithmetic actions*

$$\frac{\partial}{\partial z}(\alpha f + \beta g) = \alpha\frac{\partial f}{\partial z} + \beta\frac{\partial g}{\partial z}, \quad \frac{\partial}{\partial \bar{z}}(\alpha f + \beta g) = \alpha\frac{\partial f}{\partial \bar{z}} + \beta\frac{\partial g}{\partial \bar{z}},$$
$$\frac{\partial}{\partial z}(fg) = \frac{\partial f}{\partial z}g + \frac{\partial g}{\partial z}f, \quad \frac{\partial}{\partial \bar{z}}(fg) = \frac{\partial f}{\partial \bar{z}}g + \frac{\partial g}{\partial \bar{z}}f,$$

2. *Chain rule*

$$\frac{\partial}{\partial z}h(f(z)) = \frac{\partial h}{\partial f}(f(z))\frac{\partial f}{\partial z}(z) + \frac{\partial h}{\partial \bar{f}}(f(z))\frac{\partial \bar{f}}{\partial z}(z),$$
$$\frac{\partial}{\partial \bar{z}}h(f(z)) = \frac{\partial h}{\partial f}(f(z))\frac{\partial f}{\partial \bar{z}}(z) + \frac{\partial h}{\partial \bar{f}}(f(z))\frac{\partial \bar{f}}{\partial \bar{z}}(z).$$

3. *Conjugation rule*

$$\overline{\frac{\partial f}{\partial z}} = \frac{\partial \bar{f}}{\partial \bar{z}}, \quad \overline{\frac{\partial f}{\partial \bar{z}}} = \frac{\partial \bar{f}}{\partial z}. \tag{2.13}$$

The Wirtinger derivatives are particularly useful for optimization of real-valued functions of complex variables. Let $f : \mathbb{C}^d \mapsto \mathbb{R}$ be a differentiable real-valued function. Its differential can be rewritten in the form of the Witringer derivatives as

$$df = \frac{\partial f}{\partial \alpha}d\alpha + \frac{\partial f}{\partial \beta}d\beta = \frac{\partial f}{\partial z}dz + \frac{\partial f}{\partial \bar{z}}d\bar{z}.$$

Since f is real-valued, by (2.13) we have

$$\overline{\frac{\partial f}{\partial z}} = \frac{\partial f}{\partial \bar{z}},$$

and the differential simplifies to

$$df = 2\,\mathrm{Re}\left(\frac{\partial f}{\partial z} dz\right).$$

It is maximal when dz is a scaled version of $\overline{\frac{\partial f}{\partial z}} = \frac{\partial f}{\partial \bar{z}}$, and, thus, $\frac{\partial f}{\partial \bar{z}}$ gives the direction of the steepest ascent. Moreover, the critical points of f are those, where derivative with respect to $\bar{z}$ vanishes. For this reason, the gradient of f for variable z is defined as

$$\nabla_z f := \left(\frac{\partial f}{\partial \bar{z}}\right)^T = \left(\frac{\partial f}{\partial z}\right)^*$$

and the full gradient is

$$\nabla f = \begin{bmatrix} \nabla_z f \\ \nabla_{\bar{z}} f \end{bmatrix} = \begin{bmatrix} \nabla_z f \\ \overline{\nabla_z f} \end{bmatrix}.$$

The Hessian matrix of a function f is given by

$$\nabla^2 f = \begin{bmatrix} \nabla^2_{z,z} f & \nabla^2_{\bar{z},z} f \\ \nabla^2_{z,\bar{z}} f & \nabla^2_{\bar{z},\bar{z}} f \end{bmatrix},$$

with second order derivatives

$$\nabla^2_{z,z} f = \frac{\partial}{\partial z} \nabla_z f = \frac{\partial}{\partial z} \left(\frac{\partial f}{\partial z}\right)^*, \quad \nabla^2_{\bar{z},z} f = \frac{\partial}{\partial \bar{z}} \nabla_z f = \frac{\partial}{\partial \bar{z}} \left(\frac{\partial f}{\partial z}\right)^*,$$
$$\nabla^2_{z,\bar{z}} f = \frac{\partial}{\partial z} \left(\frac{\partial f}{\partial \bar{z}}\right)^*, \quad \nabla^2_{\bar{z},\bar{z}} f = \frac{\partial}{\partial \bar{z}} \left(\frac{\partial f}{\partial \bar{z}}\right)^*.$$

For a real-valued function f, by (2.13) we obtain the equalities

$$\nabla^2_{\bar{z},\bar{z}} f = \overline{\nabla^2_{z,z} f} \quad \text{and} \quad \nabla^2_{z,\bar{z}} f = \overline{\nabla^2_{\bar{z},z} f}. \tag{2.14}$$

The Wirtinger analogue of the second order Taylor's approximation theorem with residual in integral form states the following.

Theorem 2.3.3. *For all twice continuously differentiable functions $f : \mathbb{C}^d \mapsto \mathbb{R}$ and all $z, v \in \mathbb{C}^d$ we have*

$$f(z+v) = f(z) + \begin{bmatrix} \nabla_z f \\ \overline{\nabla_z f} \end{bmatrix}^* \begin{bmatrix} v \\ \bar{v} \end{bmatrix} + \begin{bmatrix} v \\ \bar{v} \end{bmatrix}^* \int_0^1 (1-s) \nabla^2 f(z+sv) ds \begin{bmatrix} v \\ \bar{v} \end{bmatrix}. \tag{2.15}$$

2.3.2 Gradient descent

The Wirtinger derivatives play the key role in first order minimization of real-valued functions of complex variables. Since ∇f_z provides the direction steepest ascent, we consider a gradient descent scheme with starting point $z^0 \in \mathbb{C}^d$ and a sequence $\{z^t\}_{t\geq 0}$ of iterates generated by

$$z^t = z^{t-1} - \mu_t \nabla_z f(z^{t-1}), \tag{2.16}$$

where μ_t denotes the learning rate for the t-th iteration. Under mild assumptions on the function f, we can guarantee that gradient descent with constant learning rate $\mu_t = \mu_c$, $t \geq 1$ will decrease the value of the function f and eventually stop.

Theorem 2.3.4. *Let $f : \mathbb{C}^d \to [0, \infty)$ be a twice differentiable function such that the Wirtinger Hessian satisfies*

$$\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 f(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \leq L \left\| \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \right\|_2^2 \tag{2.17}$$

for all $z, u \in \mathbb{C}^d$, where $L > 0$ is a constant independent of z. Let $\{z^t\}_{t\geq 0}$ be a sequence generated by (2.16) with arbitrary starting point $z^0 \in \mathbb{C}^d$ and learning rate $\mu_t = \mu_c$ such that $0 < \mu_c \leq 1/L$ holds. Then, we have

$$f(z^t) - f(z^{t-1}) \leq -\mu_t \left\| \nabla_z f(z^t) \right\|_2^2 \tag{2.18}$$

for all $t \geq 1$. In particular,

$$\lim_{t\to\infty} \left\| \nabla_z f(z^t) \right\|_2^2 = 0 \quad \text{and} \quad \min_{t\in[T]} \left\| \nabla_z f(z^t) \right\|_2^2 \leq \frac{f(z^0)}{\mu_c T}, \tag{2.19}$$

for all $T \geq 1$

Proof. The proof of (2.18) is based on the Taylor expansion of f using Wirtinger derivatives, which gives

$$f(z+u) = f(z) + \begin{bmatrix} \nabla_z f(z) \\ \overline{\nabla_z f(z)} \end{bmatrix}^* \begin{bmatrix} u \\ \bar{u} \end{bmatrix} + \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \int_0^1 (1-s) \nabla^2 f(z + s\,u)\, ds \begin{bmatrix} u \\ \bar{u} \end{bmatrix},$$

with $z = z^{t-1}, u = -\mu_c \nabla_z f(z^{t-1})$. Using inequality (2.17) and the assumption on μ_c we obtain that

$$\begin{aligned} f(z^t) - f(z^{t-1}) &\leq -\mu_c \left\| \begin{bmatrix} \nabla_z f(z^{t-1}) \\ \overline{\nabla_z f(z^{t-1})} \end{bmatrix} \right\|_2^2 + \frac{\mu_c^2 L}{2} \left\| \begin{bmatrix} \nabla_z f(z^{t-1}) \\ \overline{\nabla_z f(z^{t-1})} \end{bmatrix} \right\|_2^2 \\ &\leq -2\mu_c(1 - \mu_c L/2) \left\| \nabla_z f(z^{t-1}) \right\|_2^2 \\ &\leq -\mu_c \left\| \nabla_z f(z^{t-1}) \right\|_2^2, \end{aligned}$$

for every $t \geq 1$. To show (2.19) we note that for $T > 0$ we have

$$\mu_c \sum_{t=1}^{T} \left\| \nabla_z f(z^{t-1}) \right\|_2^2 \leq f(z^0) - f(z^T) \leq f(z^0),$$

which shows, that $\sum_{t=1}^{\infty} \|\nabla_z f(z^{t-1})\|_2^2$ is convergent. Consequently, $\|\nabla_z f(z^{t-1})\|_2^2 \to 0$ as $t \to \infty$. Finally, for $T > 0$, boundedness of the series yields

$$\min_{t\in[T]\}} \left\|\nabla_z f(z^t)\right\|_2^2 \leq \frac{1}{T}\sum_{t=0}^{T-1} \left\|\nabla_z f(z^t)\right\|_2^2 \leq \frac{f(z^0)}{\mu_c T}.$$

□

Remark 2.3.5. *The constant L in the condition (2.17) is* not *the Lipschitz constant of the gradient $\nabla_z f$ or smoothness constant of f. We recall that a function is said to be $\hat{L}$-smooth if for all $z_1, z_2 \in \mathbb{C}^d$ the inequality*

$$\|\nabla_z f(z_1) - \nabla_z f(z_2)\|_2 \leq \hat{L}\,\|z_1 - z_2\|_2\,,$$

holds with constant $\hat{L} \geq 0$. If f is twice continuously differentiable, $\hat{L}$-smoothness is equivalent to

$$\left|\begin{bmatrix} u \\ \bar{u}\end{bmatrix}^* \nabla^2 f(z) \begin{bmatrix} u \\ \bar{u}\end{bmatrix}\right| \leq \hat{L}\left\|\begin{bmatrix} u \\ \bar{u}\end{bmatrix}\right\|_2^2.$$

For convex functions, $\nabla^2 f(z)$ is a positive semidefinite matrix and constants L and $\hat{L}$ are the same. However, for non-convex functions f the inequality (2.17) is a weaker requirement than $\hat{L}$–smoothness.

The choice of the constant learning rate in Theorem 2.3.4 is based on the worst case scenario for all $z \in \mathbb{C}^n$ and may be suboptimal when

$$\begin{bmatrix} u \\ \bar{u}\end{bmatrix}^* \nabla^2 f(z) \begin{bmatrix} u \\ \bar{u}\end{bmatrix} \text{ is much smaller than } L\left\|\begin{bmatrix} u \\ \bar{u}\end{bmatrix}\right\|_2^2.$$

In this case, the so-called Armijo-Goldstein condition can be used to find a learning rate allowing for a larger decrease of the objective. This condition reads as

$$f(z - \mu\nabla_z f(z)) - f(z) \leq -\mu\,\|\nabla_z f(z)\|_2^2\,. \tag{2.20}$$

A suitable learning rate is now determined iteratively by the following backtracking line search algorithm, which we will call henceforth Armijo-Goldstein algorithm (AG, for short).

Algorithm 1: Backtracking search or Armijo-Goldstein condition (AG)

Input : Differentiable function $f : \mathbb{C}^d \to [0, +\infty)$, current position $z \in \mathbb{C}^d$, initial value $\mu_0 > 0$ and decrease factor $0 < \tau < 1$.
Output: Selected learning rate μ.
for $j = 0, 1, \ldots$ **do**
  if $f(z - \mu_j\nabla_z f(z)) - f(z) \leq -\mu_j\,\|\nabla_z f(z)\|_2^2$ **then**
    return $\mu = \mu_j$
  $\mu_{j+1} = \tau\mu_j$

In addition, we make use of the fact that by Theorem 2.3.4 the constant learning rate satisfies (2.20). Hence, by setting $\mu_0 = \mu_c\tau^{-N}$ for $N \in \mathbb{N} \cup \{0\}$, AG will always stop

at $\mu_N = \mu_c$ after N iterations. Note that the so determined parameter μ depends on f, z, τ, μ_c and the number of iterations N to meet the condition, i.e., $\mu = \mu(f, z, \tau, \mu_c, N)$. Moreover,

$$\mu_c \tau^{-N} \geq \mu \geq \mu_c \tag{2.21}$$

by construction. Also, in case $N = 0$, the learning rate selected by AG coincides with the constant learning rate μ_c.
The results of Theorem 2.3.4 extend to gradient descent with learning rates μ_t determined by AG.

Theorem 2.3.6. *Under the conditions of Theorem 2.3.4, a sequence $\{z^t\}_{t\geq 0}$ generated by (2.16) with arbitrary starting point $z^0 \in \mathbb{C}^d$ and learning rates $\mu_t = \mu_t(f, z^{t-1}, \tau, \mu_c, N)$ determined by Algorithm 1 satisfies (2.18) and (2.19).*

Proof. Inequality (2.18) holds by construction. The rest of the proof is analogous to the proof of Theorem 2.3.4. □

At last, we note that the quadratic form in the condition (2.17) can be rewritten using the equalities (2.14) as

$$\begin{aligned}
\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 f(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} &= \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \begin{bmatrix} \nabla^2_{z,z} f & \nabla^2_{\bar{z},z} f \\ \nabla^2_{z,\bar{z}} f & \nabla^2_{\bar{z},\bar{z}} f \end{bmatrix} \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \\
&= u^* \nabla^2_{z,z} f u + u^* \nabla^2_{\bar{z},z} f \bar{u} + u^T \nabla^2_{z,\bar{z}} f u + u^T \nabla^2_{\bar{z},\bar{z}} f \bar{u} \\
&= u^* \nabla^2_{z,z} f u + u^* \nabla^2_{\bar{z},z} f \bar{u} + \overline{u^* \nabla^2_{\bar{z},z} f \bar{u}} + \overline{u^* \nabla^2_{z,z} f u} \\
&= 2 \operatorname{Re} \left(u^* \nabla^2_{z,z} f u + u^* \nabla^2_{\bar{z},z} f \bar{u} \right),
\end{aligned} \tag{2.22}$$

which is more convenient when establishing the inequality (2.17).

2.3.3 Stochastic gradient descent

Stochastic gradient descent plays a major role in modern applications of optimization theory. One of the reasons behind the popularity of this method is the reduced computational complexity caused by only a partial evaluation of the gradient. For scenarios, where large datasets have to be processed, this allows for an efficient minimization of the objective function. We will consider a stochastic gradient descent scheme in the further sections and, thus, we provide the reader with the necessary results regarding its weak convergence based on [53]. The reader may also find more recent results on almost sure convergence of the stochastic gradient descent in [54, 55]. Since the results in [53] are derived for functions of real variables under the assumption that f is the L-smooth, we include the proofs for the complex case and functions satisfying condition (2.17) for completeness.
Let $f : \mathbb{C}^d \to \mathbb{R}$ be a continuous differentiable function, which admits a decomposition

$$f(z) = \sum_{r \in [R]} f_r(z),$$

for all $z \in \mathbb{C}^d$, where functions $f_r : \mathbb{C}^d \to \mathbb{R}$, $r \in [R]$ for some $R \in \mathbb{N}$ are also continuous differentiable functions. Then, the stochastic gradient of f in z direction is defined as

$$g_f(z) := \sum_{r \in [R]} v_r \nabla_z f_r(z), \tag{2.23}$$

for all $z \in \mathbb{C}^d$, with v_r being the sampling random variables. The random distribution of the vector $v \in \mathbb{C}^R$ is commonly chosen such that $g_f(z)$ is an unbiased estimate of $\nabla_z f(z)$, that is $\mathbb{E} g_f(z) = \nabla_z f(z)$.
Now, let us consider the stochastic gradient descent scheme with a starting point $z^0 \in \mathbb{C}^d$ and a sequence of iterates generated by

$$z^t = z^{t-1} - \mu_t g_f(z^{t-1}), \tag{2.24}$$

where μ_t denotes the learning rate for the t-th iteration. Similarly to the gradient descent discussed in the previous section, our goal is to show that for a suitable choice of a constant learning rate the algorithm converges.

Theorem 2.3.7 (Version of [53, Theorem 2]). *Let functions $f, f_r, r \in [R]$ be continuously differentiable and assume that f is bounded from below by 0, is twice continuously differentiable and satisfies the inequality (2.17) with constant $L > 0$. Assume that the stochastic gradient $g_f(z)$ given by (2.23) satisfies $\mathbb{E} g_f(z) = \nabla_z f(z)$ and*

$$\mathbb{E} \left\| g_f(z) \right\|_2^2 \leq A f(z) + B \left\| \nabla_z f(z) \right\|_2^2 + C, \tag{2.25}$$

for all $z \in \mathbb{C}^d$ and some constants $A, B, C \geq 0$. Let $\{z^t\}_{t \geq 0}$ be a sequence determined by (2.24) with an arbitrary starting point $z^0 \in \mathbb{C}^d$ and the constant learning rate $\mu_t = \mu_c$ satisfying $0 < \mu_c \leq \frac{1}{BL}$, where in the case $B = 0$ the right inequality is understood as $< +\infty$. Then, we have

$$\min_{t \in [T]} \mathbb{E} \left\| \nabla_z f(z^t) \right\|_2^2 \leq \begin{cases} AL\mu_c f(z^0) \left[1 + \frac{1}{TAL\mu_c^2} \right] + CL\mu_c, & A > 0, \\ \frac{f(z^0)}{T\mu_c} + CL\mu_c, & A = 0. \end{cases}$$

Proof. We start by applying the Taylor expansion (2.15) of f, which gives

$$f(z+u) = f(z) + \begin{bmatrix} \nabla_z f(z) \\ \overline{\nabla_z f(z)} \end{bmatrix}^* \begin{bmatrix} u \\ \bar{u} \end{bmatrix} + \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \int_0^1 (1-s) \nabla^2 f(z + s\,u)\, ds \begin{bmatrix} u \\ \bar{u} \end{bmatrix},$$

with $z = z^t, u = -\mu_c g_f(z^t)$. Using the inequality (2.17) for f, for every $t \geq 0$ we obtain

$$\begin{aligned} f(z^{t+1}) &\leq f(z^t) + \left\langle \begin{bmatrix} -\mu_c g_f(z^t) \\ -\mu_c \overline{g_f(z^t)} \end{bmatrix}, \begin{bmatrix} \nabla_z f(z^t) \\ \overline{\nabla_z f(z^t)} \end{bmatrix} \right\rangle + \frac{L}{2} \left\| \begin{bmatrix} -\mu_c g_f(z^t) \\ -\mu_c \overline{g_f(z^t)} \end{bmatrix} \right\|_2^2 \\ &= f(z^t) - 2\mu_c \operatorname{Re} \left\langle g_f(z^t), \nabla_z f(z^t) \right\rangle + L\mu_c^2 \left\| g_f(z^t) \right\|_2^2. \end{aligned} \tag{2.26}$$

If we condition on z^t, that is, we fix values of the random variables from previous iterations and only consider the randomness resulting from sampling in the stochastic gradient $g_f(z^t)$, by assumptions on g_f, the conditional expectation is bounded by

$$\begin{aligned} \mathbb{E}[f(z^{t+1}) \mid z^t] &\leq f(z^t) - 2\mu_c \left\langle \mathbb{E}[g_f(z^t) \mid z^t], \nabla_z f(z^t) \right\rangle + L\mu_c^2 \mathbb{E} \left[\left\| g_f(z^t) \right\|_2^2 \mid z^t \right] \\ &\leq f(z^t) - 2\mu_c \left\| \nabla_z f(z^t) \right\|_2^2 + L\mu_c^2 \left[A f(z) + B \left\| \nabla_z f(z^t) \right\|_2^2 + C \right] \\ &\leq -2\mu_c \left[1 - \frac{BL\mu_c}{2} \right] \left\| \nabla_z f(z^t) \right\|_2^2 + \left[1 + AL\mu_c^2 \right] f(z^t) + CL\mu_c^2 \\ &\leq -\mu_c \left\| \nabla_z f(z^t) \right\|_2^2 + \left[1 + AL\mu_c^2 \right] f(z^t) + CL\mu_c^2, \end{aligned}$$

where in the last line we used the condition on the learning rate for $B > 1$. If $B = 0$, the inequality is trivial. By defining $\beta := 1 + AL\mu_c^2$, taking the expectation and multiplying both sides with $\beta^{-(t+1)}\mu_c^{-1}$ we obtain

$$\beta^{-(t+1)}\mu_c^{-1}\mathbb{E}f(z^{t+1}) \leq -\beta^{-(t+1)}\mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2 + \beta^{-t}\mu_c^{-1}\mathbb{E}f(z^t) + CL\mu_c\beta^{-(t+1)}.$$

Next, we rearrange the terms and sum up the obtained inequalities for $t \in [T]$, which gives

$$\begin{aligned}\sum_{t\in[T]} \beta^{-(t+1)}\mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2 &\leq \mu_c^{-1}\sum_{t\in[T]}[\beta^{-t}\mathbb{E}f(z^t) - \beta^{-(t+1)}\mathbb{E}f(z^{t+1})] + CL\mu_c\sum_{t\in[T]}\beta^{-(t+1)} \\ &\leq \mu_c^{-1}[f(z^0) - \beta^T\mathbb{E}f(z^T)] + CL\mu_c\sum_{t\in[T]}\beta^{-(t+1)} \\ &\leq \mu_c^{-1}f(z^0) + CL\mu_c\sum_{t\in[T]}\beta^{-(t+1)}.\end{aligned}$$

Therefore, $\min_{t\in[T]}\mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2$ is bounded from above by

$$\begin{aligned}\min_{t\in[T]}\mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2 &\leq \frac{\sum_{t\in[T]}\beta^{-(t+1)}\mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2}{\sum_{t\in[T]}\beta^{-(t+1)}} \\ &\leq \frac{f(z^0)}{\mu_c\sum_{t\in[T]}\beta^{-(t+1)}} + CL\mu_c.\end{aligned}$$

If $A = 0$, then $\beta = 1$ and

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2 \leq \frac{f(z^0)}{\mu_c T} + CL\mu_c.$$

Otherwise, the sum in the denominator is the geometric sum, so that

$$\sum_{t\in[T]}\beta^{-(t+1)} = \beta^{-1}\frac{\beta^{-T}-1}{\beta^{-1}-1} = \frac{1-\beta^T}{\beta^T(1-\beta)} = \frac{\beta^T-1}{\beta^T(\beta-1)},$$

and, thus, we have

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2 \leq \frac{(\beta-1)f(z^0)}{\mu_c}\frac{\beta^T}{\beta^T-1} + CL\mu_c = \frac{(\beta-1)f(z^0)}{\mu_c}\left[1 + \frac{1}{\beta^T-1}\right] + CL\mu_c.$$

Note that the function $x \mapsto 1 + \frac{1}{x}$ is decreasing on $(0, +\infty)$ and

$$\beta^T - 1 = \left[1 + AL\mu_c^2\right]^T - 1 \geq TAL\mu_c^2.$$

Therefore, we obtain

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2 \leq AL\mu_c f(z^0)\left[1 + \frac{1}{TAL\mu_c^2}\right] + CL\mu_c.$$

Combining the two cases together concludes the proof. □

Theorem 2.3.7 further provides the following stopping criteria.

Corollary 2.3.8 ([53, Corollary 1])**.** *Let the assumptions of Theorem 2.3.7 hold and fix $\gamma > 0$. If the number of iterations T satisfies*

$$T \geq \max\left\{\frac{16LAf^2(z^0)}{\gamma^4}, \frac{4BLf(z^0)}{\gamma^2}, \frac{8CLf(z^0)}{\gamma^4}\right\},$$

and the constant learning rate fulfills

$$\mu_c \leq \min\left\{\frac{1}{\sqrt{TAL}}, \frac{1}{BL}, \frac{\gamma^2}{2CL}\right\},$$

then the expected norms of the gradients satisfy

$$\min_{t\in[T]} \mathbb{E}\left\|\nabla_z f(z^t)\right\|_2 \leq \gamma.$$

In the case A, B or C are equal to zero, the corresponding cases in the upper bound for the learning rate μ_c are ignored and if either $A = 0$ or $C = 0$, the lower bound on T improves by a constant factor.

Proof. We start with the case $A > 0$. By construction μ_c satisfies $\frac{1}{TAL\mu_c^2} \geq 1$, and by Theorem 2.3.7, we have

$$\min_{t\in[T]} \mathbb{E}\left\|\nabla_z f(z^t)\right\|_2^2 \leq A\mu_c f(z^0)\left[1 + \frac{1}{TA\mu_c^2}\right] + CL\mu_c = \frac{2f(z^0)}{T\mu_c} + CL\mu_c.$$

The second term satisfies $CL\mu_c \leq \gamma^2/2$ by the choice of μ_c. Hence, selecting

$$T \geq \frac{4f(z^0)}{\gamma^2\mu_c} \geq \frac{4f(z^0)}{\gamma^2 \min\left\{\frac{1}{\sqrt{TAL}}, \frac{1}{BL}, \frac{\gamma^2}{2CL}\right\}}$$

will provide the desired gradient bound. By splitting the minimum into three separate cases we obtain

$$T \geq \frac{4\sqrt{TAL}f(z^0)}{\gamma^2}, \quad T \geq \frac{4BLf(z^0)}{\gamma^2} \quad \text{and} \quad T \geq \frac{8CLf(z^0)}{\gamma^4},$$

which is equivalent to

$$T \geq \frac{16ALf^2(z^0)}{\gamma^4}, \quad T \geq \frac{4BLf(z^0)}{\gamma^2} \quad \text{and} \quad T \geq \frac{8CLf(z^0)}{\gamma^4}.$$

For the case $A = 0$, the coefficient in front of $f(z^0)$ in Theorem 2.3.7 is better by a factor of 2. For the case $C = 0$, we only need to require $T \geq 2f(z^0)/\mu_c^2\gamma$, which also improves the bound by a factor of 2. □

Comparing the results of Theorem 2.3.4 and Corollary 2.3.8, the two main differences are prominent. Firstly, for gradient descent with $\mu_c = 1/L$ only $T \geq f(z^0)L\gamma^{-2}$ iterations are required to achieve $\min_{t\in[T]} \mathbb{E}\left\|\nabla_z f(z^t)\right\|_2 \leq \gamma$. In comparison, stochastic gradient descent

requires $\mathcal{O}(\gamma^{-4} f^2(z^0))$ iterations, which is significantly slower. Secondly, for stochastic gradient descent the choice of the constant learning rate depends on the number of iterations of the algorithm. However, these two drawbacks are compensated by the reduced computational complexity as the gradients $\nabla_z f_r$ are only evaluated for small subset of $[R]$.
Note that if $g_f(z) = \nabla_z f(z)$, then condition (2.25) is satisfied with $A = C = 0$ and $B = 1$ and Theorem 2.3.7 becomes Theorem 2.3.4. For a random construction of $g_f(z)$ via (2.23), the distribution of v is the determining factor for constants A, B, C. In this thesis, we will only consider a sampling with replacement, however other choices are also possible [53]. That is, the entries of v are given by

$$v_r = \frac{1}{Kp_r} \sum_{k\in[K]} v_r^k, \tag{2.27}$$

where each $v^k \in \mathbb{R}^R$ is independently sampled with replacement from a set of standard basis vectors $\{e_r\}_{r\in[R]}$ with probability $0 < p_r \leq 1$ to pick e_r. The main components in the establishment of (2.25) for this sampling scheme are summarized in the following proposition.

Proposition 2.3.9 (Version of [53, Proposition 3]). *Let $v \in \mathbb{R}^R$ be defined as in (2.27) and assume that for all $r \in [R]$ functions $f_r : \mathbb{C}^d \to [0, +\infty)$ are continuously differentiable. Then, for all $z \in \mathbb{C}^d$ the stochastic gradient $g_f(z)$ given by (2.23) satisfies*

$$\mathbb{E} g_f(z) = \nabla_z f(z) \quad \textit{and} \quad \mathbb{E}\left\|g_f(z)\right\|_2^2 \leq \left[1 - \frac{1}{K}\right] \left\|\nabla_z f(z)\right\|_2^2 + \sum_{r\in[R]} \frac{1}{Kp_r} \left\|\nabla_z f_r(z)\right\|_2^2 .$$

Proof. Let us start by computing the expectation of v_r^k,

$$\mathbb{E} v_r^k = \sum_{r'\in[R]} p_{r'} (e_{r'})_r = \sum_{r'\in[R]} p_{r'} \mathcal{I}_{r'=r} = p_r. \tag{2.28}$$

Therefore, by linearity of the expectation we obtain

$$\begin{aligned}
\mathbb{E} g_f(z) &= \sum_{r\in[R]} \mathbb{E} v_r \nabla_z f_r(z) = \sum_{r\in[R]} \nabla_z f_r(z) \mathbb{E}\left[\frac{1}{Kp_r} \sum_{k\in[K]} v_r^k\right] \\
&= \sum_{r\in[R]} \frac{1}{Kp_r} \nabla_z f_r(z) \sum_{k\in[K]} \mathbb{E} v_r^k = \sum_{r\in[R]} \nabla_z f_r(z) = \nabla_z f(z).
\end{aligned}$$

For the expectation of the squared norm we have,

$$\begin{aligned}
\mathbb{E}\left\|g_f(z)\right\|_2^2 &= \mathbb{E}\left\langle g_f(z), g_f(z)\right\rangle = \mathbb{E}\left\langle \sum_{r_1\in[R]} v_{r_1} \nabla_z f_{r_1}(z), \sum_{r_2\in[R]} v_{r_2} \nabla_z f_{r_2}(z)\right\rangle \\
&= \sum_{r_1, r_2\in[R]} \left\langle \nabla_z f_{r_1}(z), \nabla_z f_{r_2}(z)\right\rangle \mathbb{E}[v_{r_1} v_{r_2}].
\end{aligned} \tag{2.29}$$

We further compute the expectation by substituting (2.27), which yields

$$\mathbb{E}[v_{r_1} v_{r_2}] = \mathbb{E}\left[\frac{1}{K p_{r_1}} \sum_{k_1 \in [K]} v_{r_1}^{k_1} \frac{1}{K p_{r_2}} \sum_{k_2 \in [K]} v_{r_2}^{k_2}\right] = \frac{1}{K^2 p_{r_1} p_{r_2}} \sum_{k_1, k_2 \in [K]} \mathbb{E}[v_{r_1}^{k_1} v_{r_2}^{k_2}]. \tag{2.30}$$

The expectation $\mathbb{E}[v_{r_1}^{k_1} v_{r_2}^{k_2}]$ takes three possible values depending on k_1, k_2, r_1, r_2. Firstly, if $k_1 \neq k_2$ then v^{k_1} and v^{k_2} are independent and, thus, by (2.28) we obtain

$$\mathbb{E}[v_{r_1}^{k_1} v_{r_2}^{k_2}] = \mathbb{E} v_{r_1}^{k_1} \mathbb{E} v_{r_2}^{k_2} = p_{r_1} p_{r_2}.$$

The second case is $k_1 = k_2$, but $r_1 \neq r_2$. Since v^{k_1} is always the standard basis vector e_j for some $j \in [R]$, it implies that for any $r_1, r_2 \in [R]$ such that $r_1 \neq r_2$ either $(e_j)_{r_1} = 0$ or $(e_j)_{r_2} = 0$. Thus, it always holds that $(e_j)_{r_1}(e_j)_{r_2} = 0$ and, consequently, $v_{r_1}^{k_1} v_{r_2}^{k_1} = 0$. Hence, $\mathbb{E}[v_{r_1}^{k_1} v_{r_2}^{k_2}] = 0$.
Lastly, if $k_1 = k_2$ and $r_1 = r_2$, we have

$$\mathbb{E}[v_{r_1}^{k_1} v_{r_2}^{k_2}] = \mathbb{E}[(v_{r_1}^{k_1})^2] = \mathbb{E}[v_{r_1}^{k_1}] = p_{r_1},$$

where we used that the entries of v^{k_1} are either 0 or 1 and satisfy the equalities $0^2 = 0$ and $1^2 = 1$.
Combining all cases together gives

$$\mathbb{E}[v_{r_1}^{k_1} v_{r_2}^{k_2}] = \begin{cases} p_{r_1} p_{r_2}, & k_1 \neq k_2 \\ 0, & k_1 = k_2,\ r_1 \neq r_2 \\ p_{r_1}, & k_1 = k_2,\ r_1 = r_2. \end{cases}$$

Substitution of this equality to (2.30) yields

$$\mathbb{E}[v_{r_1} v_{r_2}] = \frac{1}{K^2} \sum_{\substack{k_1, k_2 \in [K] \\ k_1 \neq k_2}} 1 + \frac{1}{K^2 p_{r_1}} \sum_{k \in [K]} \mathcal{I}_{r_1 = r_2} = 1 - \frac{1}{K} + \frac{1}{K p_{r_1}} \mathcal{I}_{r_1 = r_2},$$

and, consequently, by (2.29) we obtain

$$\begin{aligned} \mathbb{E}\left\|g_f(z)\right\|_2^2 &= \left[1 - \frac{1}{K}\right] \sum_{r_1, r_2 \in [R]} \left\langle \nabla_z f_{r_1}(z), \nabla_z f_{r_2}(z) \right\rangle + \sum_{r \in [R]} \frac{1}{K p_r} \left\langle \nabla_z f_r(z), \nabla_z f_r(z) \right\rangle \\ &= \left[1 - \frac{1}{K}\right] \left\langle \sum_{r_1 \in [R]} \nabla_z f_{r_1}(z), \sum_{r_2 \in [R]} \nabla_z f_{r_2}(z) \right\rangle + \sum_{r \in [R]} \frac{1}{K p_r} \left\|\nabla_z f_r(z)\right\|_2^2 \\ &= \left[1 - \frac{1}{K}\right] \left\|\nabla_z f(z)\right\|_2^2 + \sum_{r \in [R]} \frac{1}{K p_r} \left\|\nabla_z f_r(z)\right\|_2^2. \end{aligned}$$

□

If the sampling is uniform, so that

$$p_r = 1/R, \quad r \in [R], \tag{2.31}$$

we obtain the following convergence guarantees for stochastic gradient descent.

Corollary 2.3.10. *Fix $\gamma > 0$. Assume that for all $r \in [R]$ functions $f, f_r : \mathbb{C}^d \to [0, +\infty)$ are twice continuously differentiable and satisfy the inequality (2.17) with constants $L, L_r > 0$. Let $\{z^t\}_{t\geq 0}$ be a sequence determined by (2.24) with an arbitrary starting point $z^0 \in \mathbb{C}^d$, sampling (2.27) with probabilities (2.31) and constant learning rate $\mu_t = \mu_c$. If the number of iterations T satisfies*

$$T \geq \max\left\{\frac{4LR\max_{r\in[R]} L_r f^2(z^0)}{K\gamma^4}, \frac{2L(K-1)f(z^0)}{K\gamma^2}\right\},$$

and the constant learning rate fulfills

$$\mu_c \leq \min\left\{\sqrt{\frac{K}{TLR\max_{r\in[R]} L_r}}, \frac{K}{L(K-1)}\right\},$$

then the expected norms of the gradient satisfy

$$\min_{t\in[T]} \mathbb{E}\left\|\nabla_z f(z^t)\right\|_2 \leq \gamma.$$

In the case $K = 1$, the learning rate only needs to be $\mu_c \leq (TLR\max_{r\in[R]} L_r)^{-1/2}$.

Proof. Let us show that the stochastic gradient satisfies condition (2.25) with

$$A = R\max_{r\in[R]} L_r/K, \quad B = 1 - 1/K \quad \text{and } C = 0.$$

In view of Proposition 2.3.9, we only need to bound $\sum_{r\in[R]} \frac{1}{Kp_r}\|\nabla_z f_r(z)\|_2^2$. For this, we consider a single step of gradient descent applied to f_r with starting point z and learning rate $\mu = 1/L_r$. Then, the assumptions of Theorem 2.3.4 are satisfied and the inequality (2.18) gives

$$\|\nabla_z f_r(z)\|_2^2 \leq L_r[f_r(z) - f_r(z - \tfrac{1}{L_r}\nabla_z f_r(z))] \leq L_r f_r(z).$$

Combining this inequality with (2.31) leads to

$$\sum_{r\in[R]} \frac{1}{Kp_r}\|\nabla_z f_r(z)\|_2^2 \leq \frac{R}{K}\sum_{r\in[R]} L_r f_r(z) \leq \frac{R}{K}\max_{r\in[R]} L_r f(z),$$

and, hence, Proposition 2.3.9 yields

$$\mathbb{E}\|g_f(z)\|_2^2 \leq \left[1 - \frac{1}{K}\right]\|\nabla_z f(z)\|_2^2 + \frac{R}{K}\max_{r\in[R]} L_r f(z).$$

Thus, Corollary 2.3.8 applies, which concludes the proof. □

Chapter 3

Ptychography

3.1 Continuous ptychographic problem

In this section we discuss the reconstruction of an object from diffraction patterns generated by monochromatic illumination in the ptychographic experiment.
We start with the intensity function of the diffraction patterns given by (1.12),

$$\left|\frac{\nu}{p}\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}]\left(\frac{\nu s}{p}\right)\right|^2, \quad r,s\in\mathbb{R}^2,$$

derived in Chapter 1 and for readability, we introduce a few simplifications. We recall that the functions $\boldsymbol{x},\boldsymbol{w}:\mathbb{R}^2\to\mathbb{C}$ denote the unknown object and the window, respectively, and $\nu,p>0$ are fixed parameters describing the measurement process. Firstly, we introduce a change of variable from s to ps/ν in order to avoid scaling in the argument of the Fourier transform. Additionally, the intensity function is rescaled to remove the multiplicative factor ν/p. Lastly, in this section we will work with the one-dimensional counterpart of the problem.
As a result, the one-dimensional rescaled intensity function of the diffraction patterns is given by

$$\boldsymbol{I}(r,s)=\left|\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}](s)\right|^2, \quad s\in\mathbb{R}, r\in\mathbb{R}. \tag{3.1}$$

From this point onward, we will concentrate on the mathematical aspects of the recovery of $\boldsymbol{x}$ from the measurements $\boldsymbol{I}$.
So far, we did not discuss the requirements on functions and operators involved and, thus, let us make a step back and briefly reintroduce operators and functions rigorously.
We start by defining the $\mathbb{L}_p(\mathbb{R}^k), k\in\mathbb{N}$ spaces with parameter $1\le p<\infty$ as

$$\mathbb{L}_p(\mathbb{R}^k):=\{\boldsymbol{u}:\mathbb{R}^k\to\mathbb{C}:\ \boldsymbol{u}-\text{Lebesque measurable},\ \int_{\mathbb{R}^k}|\boldsymbol{u}(s)|^p ds<+\infty\},$$

with the norm

$$\|u\|_p:=\left(\int_{\mathbb{R}^k}|\boldsymbol{u}(s)|^p ds\right)^{1/p}.$$

Clearly, the space $\mathbb{L}_2(\mathbb{R}^k)$ is a Hilbert space with inner product

$$\langle\boldsymbol{u},\boldsymbol{v}\rangle:=\int_{\mathbb{R}^k}\boldsymbol{u}(s)\overline{\boldsymbol{v}}(s)ds.$$

The translation operator $\mathcal{T}_r$ with shift $r \in \mathbb{R}$ acting on functions $\boldsymbol{u} : \mathbb{R} \to \mathbb{R}$ is defined by

$$\mathcal{T}_r \boldsymbol{u}(s) = \boldsymbol{u}(s - r), \quad s \in \mathbb{R}.$$

Since the Lebesque measure is translation invariant, we have $\mathcal{T}_r \boldsymbol{u} \in \mathbb{L}_p(\mathbb{R})$ and $\|\mathcal{T}_r \boldsymbol{u}\|_p = \|\boldsymbol{u}\|_p$ whenever $\boldsymbol{u} \in \mathbb{L}_p(\mathbb{R})$. Moreover, the inverse of $\mathcal{T}_r$ is the reverse shift $\mathcal{T}_{-r}$, so that

$$\mathcal{T}_{-r}\mathcal{T}_r \boldsymbol{u} = \mathcal{T}_r \mathcal{T}_{-r} \boldsymbol{u} = \boldsymbol{u}.$$

The Fourier transform $\mathcal{F}$ on $\mathbb{L}_1(\mathbb{R}^k)$ is given by

$$\mathcal{F}\boldsymbol{u}(\xi) := \int_{\mathbb{R}^k} \boldsymbol{u}(s) e^{-2\pi i \langle s, \xi \rangle} ds.$$

It is pointwise well-defined, since

$$|\mathcal{F}\boldsymbol{u}(\xi)| = \left| \int_{\mathbb{R}^k} \boldsymbol{u}(s) e^{-2\pi i \langle s, \xi \rangle} ds \right| \leq \int_{\mathbb{R}^k} |\boldsymbol{u}(s)| ds = \|\boldsymbol{u}\|_1 < +\infty.$$

In particular, if we assume that $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{L}_2(\mathbb{R})$, the intensity function $\boldsymbol{I}$ is pointwise well-defined, since by Cauchy-Schwarz inequality

$$|\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}](\xi)| \leq \int_{\mathbb{R}} |\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}|(s) ds \leq \|\boldsymbol{w}\|_2 \|\mathcal{T}_{-r}\boldsymbol{x}\|_2 = \|\boldsymbol{w}\|_2 \|\boldsymbol{x}\|_2 < +\infty, \quad \forall r \in \mathbb{R}.$$

Further properties of the intensity function $\boldsymbol{I}$ can be deduced after a slight rearrangement of its components,

$$\begin{aligned} \boldsymbol{I}(r,s) &= |\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}](s)|^2 = |\mathcal{F}[(\mathcal{T}_{-r}\mathcal{T}_r\boldsymbol{w})\mathcal{T}_{-r}\boldsymbol{x}](s)|^2 = |\mathcal{F}\mathcal{T}_{-r}[\boldsymbol{x}\mathcal{T}_r\boldsymbol{w}](s)|^2 \\ &= \left| \int_{\mathbb{R}} [\boldsymbol{x}\mathcal{T}_r\boldsymbol{w}](q+r) e^{-2\pi i q s} dq \right|^2 = \left| \int_{\mathbb{R}} [\boldsymbol{x}\mathcal{T}_r\boldsymbol{w}](q) e^{-2\pi i (q-r) s} dq \right|^2 \\ &= \left| e^{2\pi i r s} \mathcal{F}[\boldsymbol{x}\mathcal{T}_r\boldsymbol{w}](s) \right|^2 = |\mathcal{F}[\boldsymbol{x}\mathcal{T}_r\boldsymbol{w}](s)|^2 = \left| \mathcal{F}[\boldsymbol{x}\mathcal{T}_r\overline{\boldsymbol{w}}](s) \right|^2. \end{aligned} \tag{3.2}$$

The transform

$$\mathcal{V}_{\boldsymbol{w}}[\boldsymbol{x}](r,s) := \mathcal{F}[\boldsymbol{x}\mathcal{T}_r\overline{\boldsymbol{w}}](s). \tag{3.3}$$

is also known as the *Short-Time Fourier Transform* (STFT) with window $\boldsymbol{w}$. When the window is a Gaussian function $\boldsymbol{w}(s) = e^{-\pi s^2/a}$ for $a > 0$, we refer to $\mathcal{V}_{\boldsymbol{w}}[\boldsymbol{x}]$ as *Gabor transform* of $\boldsymbol{x}$.
For STFT the following properties hold.

Theorem 3.1.1 (Selected properties of STFT)**.** *Let $\boldsymbol{x}, \boldsymbol{x'}, \boldsymbol{w}, \boldsymbol{w'} \in \mathbb{L}_2(\mathbb{R})$. Then, we have:*

1. *$\mathcal{V}_{\boldsymbol{w}}[\boldsymbol{x}]$ is a uniformly continuous function.*
2. *$\mathcal{V}_{\boldsymbol{w}}$ is a linear operator, which maps $\mathbb{L}_2(\mathbb{R})$ to $\mathbb{L}_2(\mathbb{R}^2)$.*
3. *$\langle \mathcal{V}_{\boldsymbol{w}}[\boldsymbol{x}], \mathcal{V}_{\boldsymbol{w'}}[\boldsymbol{x'}] \rangle = \langle \boldsymbol{x}, \boldsymbol{x'} \rangle \langle \boldsymbol{w}, \boldsymbol{w'} \rangle$. In particular, $\|\mathcal{V}_{\boldsymbol{w}}[\boldsymbol{x}]\|_2 = \|\boldsymbol{x}\|_2 \|\boldsymbol{w}\|_2$.*
4. *If $\langle \boldsymbol{w}, \boldsymbol{w'} \rangle \neq 0$, then*

$$\boldsymbol{x}(q) = \frac{1}{\langle \boldsymbol{w'}, \boldsymbol{w} \rangle} \int_{\mathbb{R}} \int_{\mathbb{R}} \mathcal{V}_{\boldsymbol{w}}[\boldsymbol{x}](r,s) [\mathcal{T}_r \boldsymbol{w'}](q) e^{2\pi i q s} dr ds. \tag{3.4}$$

Proof. See Lemma 3.1.1, Theorem 3.2.1, Corollaries 3.2.2 and 3.2.3 in [44]. □

The function

$$\boldsymbol{S}_{\boldsymbol{w}}[\boldsymbol{x}] := |\mathcal{V}_{\boldsymbol{w}}[\boldsymbol{x}]|^2$$

is known as the *spectrogram* of the function $\boldsymbol{x}$ with respect to the window $\boldsymbol{w}$. In view of (3.2), $\boldsymbol{I}$ is the spectrogram of $\boldsymbol{x}$ with respect to $\overline{\boldsymbol{w}}$,

$$\boldsymbol{I} = \left|\mathcal{F}[\boldsymbol{x}\mathcal{T}_r\overline{\boldsymbol{w}}]\,(s)\right|^2 = |\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}]|^2 = \boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}] \in \mathbb{L}_1(\mathbb{R}^2),$$

and the *continuous ptychographic problem* is stated as

$$\boxed{\text{Find } \boldsymbol{x} \in \mathbb{L}_2(\mathbb{R}) \text{ from data } \boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}].}$$

If the phase function of the Fourier transform was known, the object $\boldsymbol{x}$ could be theoretically recovered via (3.4) and, thus, the main obstacle is the reconstruction of the phases of STFT $\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}]$ from $\boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}]$. The next example suggests, that it is impossible without further assumptions.

Example 3.1.2. *Consider two objects*

$$\boldsymbol{x}_+(s) = \mathcal{I}_{s\in[-2,-1]} + \mathcal{I}_{s\in[1,2]} \text{ and } \boldsymbol{x}_-(s) = \mathcal{I}_{s\in[-2,-1]} - \mathcal{I}_{s\in[1,2]}$$

and the window $\boldsymbol{w}(s) = \overline{\boldsymbol{w}}(s) = \mathcal{I}_{s\in[0,1]}$. *We note that* $\boldsymbol{x}_+, \boldsymbol{x}_-, \boldsymbol{w} \in \mathbb{L}_2(\mathbb{R})$ *and*

$$\|\boldsymbol{x}_+\|_2^2 = \|\boldsymbol{x}_-\|_2^2 = 2 \text{ and } \|\boldsymbol{w}\|_2^2 = 1.$$

For the indicator function $\mathcal{I}_{s\in[a,b]}$, $a, b \in \mathbb{R}, a < b$, *its STFT with respect to* $\overline{\boldsymbol{w}}$ *is given by*

$$\begin{aligned}\mathcal{V}_{\overline{\boldsymbol{w}}}[\mathcal{I}_{\cdot\in[a,b]}](r,s) &= \int_{\mathbb{R}} \mathcal{I}_{q\in[a,b]}(q)\mathcal{I}_{q-r\in[0,1]}e^{-2\pi iqs}dq = \int_{[a,b]\cap[r,r+1]} e^{-2\pi iqs}dq \\ &= \begin{cases} \frac{1}{-2\pi is}\left(e^{-2\pi is\min\{b,r+1\}} - e^{-2\pi is\max\{a,r\}}\right), & [a,b]\cap[r,r+1] \neq \varnothing, \\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$

Then, using the obtained formula and the linearity of $\mathcal{V}_{\boldsymbol{w}}[\boldsymbol{x}_\pm]$, *STFT of both* $\boldsymbol{x}_+$ *and* $\boldsymbol{x}_-$ *with respect to* $\overline{\boldsymbol{w}}$ *is given by*

$$\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}_\pm](r,s) = \begin{cases} \frac{1}{-2\pi is}\left(e^{-2\pi is\min\{-1,r+1\}} - e^{-2\pi is\max\{-2,r\}}\right), & r \in [-3,-1], \\ \pm\frac{1}{-2\pi is}\left(e^{-2\pi is\min\{2,r+1\}} - e^{-2\pi is\max\{1,r\}}\right), & r \in [0,2], \\ 0, & \text{otherwise,}\end{cases}$$

and, hence,

$$\boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}_+] = |\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}_+]|^2 = |\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}_-]|^2 = \boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}_-].$$

Therefore, it is not possible to distinguish $\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}_+]$ *and* $\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}_-]$ *from the spectrogram* $\boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}_\pm]$.

In order to achieve the unique determination of the phase of STFT $\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}]$, it is commonly assumed that $\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}]$ belongs to a certain subset of $\mathbb{L}_2(\mathbb{R})$ [56, 57, 58, 1, 59, 60, 61, 62, 63, 64, 65] (we note, that meaning of "unique" is vague, with a proper definition to follow in Section 3.3). However, even if the phases of $\mathcal{V}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}]$ are reconstructed, the numerical evaluation of (3.4) is a hard task.
Another line of research inquires, which assumptions on $\boldsymbol{x}$ and $\boldsymbol{w}$ will allow to determine $\boldsymbol{x}$ uniquely from the spectrogram measurements. For instance, in [66, 67, 68, 69] $\boldsymbol{x}$ has compact support or $\boldsymbol{x}$ belongs to the Paley-Wiener spaces in [70], shift-invariant spaces in [67, 71], modulation spaces in [72, 73], Hardy spaces in [58, 74] and many more. We provide a more detailed overview of these results at the end of Section 3.3.
The reconstruction of $\boldsymbol{x}$ from the spectogram $\boldsymbol{S}_{\overline{\boldsymbol{w}}}$ can be also formulated as a deconvolution problem [75, 76]. For a function $\boldsymbol{u} \in \mathbb{L}_2(\mathbb{R})$, define the Wigner distribution of $\boldsymbol{u}$ as

$$\mathcal{W}_{\boldsymbol{u}}(r,s) := \mathcal{F}[\boldsymbol{u}(r+\cdot/2)\overline{\boldsymbol{u}}(r-\cdot/2)](s) = \int_{\mathbb{R}} \boldsymbol{u}(r+\tau/2)\overline{\boldsymbol{u}}(r-\tau/2)e^{-2\pi i s\tau}d\tau.$$

It possesses many properties, which can be found in [77, 78, 75, 44].

Theorem 3.1.3 (Selected properties of Wigner distribution)**.** *Let $\boldsymbol{x} \in \mathbb{L}_2(\mathbb{R})$. Then, we have:*

1. *$\mathcal{W}_{\boldsymbol{x}}$ is uniformly continuous and $|\mathcal{W}_{\boldsymbol{x}}(r,s)| \leq 2\|\boldsymbol{x}\|_2^2$.*
2. *$\mathcal{W}_{\boldsymbol{x}}(r,s) = 2e^{4\pi i r s}\mathcal{V}_{\overline{\boldsymbol{x}}(-\cdot)}[\boldsymbol{x}](2r,2s)$ and, consequently, $\mathcal{W}_{\boldsymbol{x}} \in \mathbb{L}_2(\mathbb{R}^2)$.*
3. *In particular, the spectrogram $\boldsymbol{S}_{\overline{\boldsymbol{w}}}$ satisfies the identity*

$$\boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}](r,s) = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathcal{W}_{\boldsymbol{x}}(q,\tau)\mathcal{W}_{\overline{\boldsymbol{w}}}(q-r,s-\tau)d\tau dq.$$

Proof. See Lemma 4.3.1, Proposition 4.3.2 in [44] and equation (4.5) in [75]. The fact that $\mathcal{W}_{\boldsymbol{x}} \in \mathbb{L}_2(\mathbb{R}^2)$ is a consequence of Theorem 3.1.1. □

The third property can be recast as a two-dimensional convolution of the Wigner distribution $\mathcal{W}_{\boldsymbol{x}}$ with kernel $\mathcal{K}_{\boldsymbol{w}}(r,s) := \mathcal{W}_{\overline{\boldsymbol{w}}}(-r,s)$,

$$\boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}] = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathcal{W}_{\boldsymbol{x}}(q,\tau)\mathcal{K}_{\boldsymbol{w}}(r-q,s-\tau)d\tau dq =: \mathcal{W}_{\boldsymbol{x}} * \mathcal{K}_{\boldsymbol{w}}.$$

To separate $\boldsymbol{x}$ from effects of the window, we solve the deconvolution problem, e.g. by using the convolution theorem [44, p.8],

$$\mathcal{F}\boldsymbol{S}_{\overline{\boldsymbol{w}}}[\boldsymbol{x}] = \mathcal{F}[\mathcal{W}_{\boldsymbol{x}}] \cdot \mathcal{F}[\mathcal{K}_{\boldsymbol{w}}]. \tag{3.5}$$

under assumption that both $\mathcal{W}_{\boldsymbol{x}}$ and $\mathcal{K}_{\boldsymbol{w}}$ are $\mathbb{L}_1(\mathbb{R}^2)$ functions. If $\mathcal{F}[\mathcal{K}_{\boldsymbol{w}}](r,s) \neq 0$ for all $r,s \in \mathbb{R}$, then $\mathcal{W}_{\boldsymbol{x}}$ is recovered from the measurements and the next step is the reconstruction of $\boldsymbol{x}$ from its Wigner distribution $\mathcal{W}_{\boldsymbol{x}}$. In [79, 5], this technique is called the Wigner distribution deconvolution.
At last, we have to take into account that measurements are usually noisy, i.e.,

$$\boldsymbol{Y} = \boldsymbol{I} + \boldsymbol{N}, \tag{3.6}$$

where $\boldsymbol{N}$ is the noise. It represents the cumulative error resulting from approximations, which are performed in the derivation of the diffraction formula, the non-ideal measurement environment and other unaccounted factors during the ptychographic experiment. Clearly, the noise distorts the quality of the reconstruction and the theoretical analysis of its impact on the performance of the algorithms is a major research topic.

3.2 Discrete ptychographic problem and connection to phase retrieval

In numerical applications, a common approach is to approximate the continuous ptychographic problem by a discrete problem. As all physical objects are contained in a bounded domain, we can without loss of generality assume that both $\boldsymbol{x}$ and $\boldsymbol{w}$ are supported on $[0,1]$.
The transition from continuous to discrete problem is performed by an approximation of the integral in

$$\boldsymbol{I}(r,s) = |\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}]|^2(s) = \left| \int_{\mathbb{R}} [\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}](q) e^{-2\pi i q s} dq \right|^2$$

by a suitable quadrature rule. Generally, any quadrature rule which approximates the integral sufficiently well would work, however there is a benefit of using the partition with $d \in \mathbb{N}$ equidistant nodes $\{0, 1/d, \dots, (d-1)/d\}$. Firstly, the detector stores illumination data as an image, which is a sampled intensity function $\boldsymbol{I}$ at the equidistant points. Secondly, this choice leads to the discrete Fourier transform, which can be efficiently computed. Using the equidistant nodes, the Fourier transform of $\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}$ is approximated by

$$\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}](s) = \int_{\mathbb{R}} [\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}](q) e^{-2\pi i q s} dq \approx \frac{1}{d} \sum_{k\in[d]} [\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}] \left(\frac{k}{d}\right) e^{-\frac{2\pi i k s}{d}}, \quad s \in \mathbb{R}$$

If we sample $\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}](s)$ at points $s \in [d]$ and restrict shifts to points on the lattice $\{r/d, r \in \mathbb{Z}\}$, then the integral is approximated by

$$\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}](s) \approx \frac{1}{d} \sum_{k\in[d]} \boldsymbol{w}\left(\frac{k}{d}\right) \boldsymbol{x}\left(\frac{k+r}{d}\right) e^{-\frac{2\pi i k s}{d}}. \tag{3.7}$$

Further, assuming that all shifts are circular shifts, the intensity measurements (3.1) are approximated (up to scaling in d) by

$$I^d_{j,r} := \left|(F_d[w \circ S_{-r}x])_j\right|^2, \quad j \in [d], r \in \mathcal{R} \subseteq \mathbb{Z},$$

where vectors x and w are defined as

$$x_j = \boldsymbol{x}(j/d), \quad w_j = \boldsymbol{w}(j/d), \quad j \in [d], \tag{3.8}$$

and S_r is the circular shift operator (2.6). Then, the goal of the *discrete ptychographic problem* is to

$$\boxed{\text{Reconstruct } x \in \mathbb{C}^d \text{ from data } I^d_{j,r}.}$$

Remark 3.2.1. *With the circular shift operation, the object x is periodically extended once $j+r \geq d$, which allows to preserve group properties of the STFT transform given by Proposition 2.2.2. The second outcome of the circularity is a restriction of shift positions $\mathcal{R}$ from $\mathbb{Z}$ to subset of $[d]$, since for values $r \notin [d]$ measurements will be identical to $r \bmod d \in [d]$. We note that circularity of the shifts is important for Section 3.6, while results of Section 3.5 are also applicable for non-circulant shifts.*

Notation. *For the rest of the section, all indices will be considered modulo d dropping the* mod d *notation, unless it is necessary. For two indices $j, k \in [d]$, we will use $|j-k|_c$ to denote the distance accounting for the circularity,*

$$|j-k|_c := \min\{|j-k|, d-|j-k|\}.$$

In most ptychographic experiments, the window illuminates only a small region of the object and, therefore, the support of $\boldsymbol{w}$ is commonly smaller than $[0,1]$. Assume that it is given by an interval $[0, \delta/d]$ for parameter $\delta \in \mathbb{N}$, $1 < \delta \leq d$. Then, by the equation (3.8), the corresponding w has only δ non-zero entries.
In this case, each diffraction pattern encodes information about δ complex entries of the object in d real-valued entries. This leads to an oversampling ratio $\frac{d}{2\delta}$, which is high if d is much larger than δ. Consequently, reducing the sampling ratio may be beneficial for practical applications, where the low memory usage is crucial for fast performance.

Example 3.2.2. *Consider the recovery of a two-dimensional object represented by an 1024×1024 image. Let the window be an 1024×1024 image supported on the 64×64 bottom-left pixels. Assume that 128 shifts positions per axis are observed. If the sampling ratio is not reduced, then for each position an 1024×1024 image of the diffraction pattern is stored. Assuming that 4 Bytes per single pixel are used, the total required memory is*

$$4 \cdot 1024^2 \cdot 128^2 \ B \ = 2^{36} \ B \ = 2^{26} \ KB \ = 2^{16} \ MB \ = 2^{6} \ GB \ = 64 \ GB,$$

which does not fit into RAM of a standard modern laptop. On contrary, if for each position only a 128×128 subsampled image of the diffraction pattern is obtained, then the storage requirements are

$$4 \cdot 128^2 \cdot 128^2 \ B \ = 2^{30} \ B \ = 2^{20} \ KB \ = 2^{10} \ MB \ = 1 \ GB.$$

It is 64 times less and would easily fit into RAM.

Example 3.2.2 motivates to consider subsampled intensity measurements obtained by reducing the number of frequency samples of $\boldsymbol{I}$. Instead of $s \in [d]$, let us consider the set

$$\left\{0, \frac{d}{m}, 2\frac{d}{m}, \ldots, (m-1)\frac{d}{m}\right\} = \left\{j\frac{d}{m}, \ j \in [m]\right\}$$

of size $m \in \mathbb{N}, \delta \leq m \leq d$. Then, returning to the approximation (3.7), we obtain

$$\mathcal{F}[\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}]\left(j\frac{d}{m}\right) \approx \frac{1}{d}\sum_{k\in[d]} w_k x_{k+r} e^{-\frac{2\pi i k j d}{dm}} = \frac{1}{d}\sum_{k\in[\delta]} w_k x_{k+r} e^{-\frac{2\pi i k j}{m}},$$

where we used that $\operatorname{supp}(w) \subseteq [\delta]$. Consequently, the subsampled discrete intensity measurements (up to scaling in d) are given by

$$I_{j,r}^m := |(F_m P_m[S_{-r}x \circ w])_j|^2 = \left| \sum_{k\in[\delta]} w_k x_{k+r} e^{-\frac{2\pi i k j}{m}} \right|^2, \quad j \in [m], r \in \mathcal{R} \subseteq [d],$$

with P_m denoting the projection operator (2.10).
Just as in the continuous case (3.6), the measurements may be corrupted by noise $N \in \mathbb{R}^{m\times|\mathcal{R}|}$

$$Y_{j,r} = I_{j,r}^m + N_{j,r}, \quad j \in [m], r \in \mathcal{R} \subseteq [d]. \tag{PTY}$$

When all $N_{j,r} = 0$, we will refer to the measurements as *noiseless* and in all other cases as *noisy*. For the rest of this chapter, we will concentrate on solving the following reconstruction problem:

Reconstruct $x \in \mathbb{C}^d$ from data (PTY).

While some of the recovery methods in the literature are developed specifically for the ptychographic measurements (PTY), many others were developed for measurements of the form

$$y_k = |(Ax)_k|^2 + n_k, \quad k \in [M], \tag{PR}$$

with measurement matrix $A \in \mathbb{C}^{M\times d}$ and noise n. The corresponding reconstruction problem

Reconstruct $x \in \mathbb{C}^d$ from data (PR).

is known as the *discrete phase retrieval problem*. The connection between the two problems is observed by rewriting the Hadamard product as a diagonal matrix for each fixed illumination location r,

$$F_m P_m[S_{-r}x \circ w] = F_m P_m[w \circ S_{-r}x] = F_m P_m \operatorname{diag}(w) S_{-r}x.$$

Combining all locations together as a block matrix, we obtain that in the ptychographic case A is given by

$$A = \begin{bmatrix} F_m P_m \operatorname{diag}(w) S_{-r_1} \\ \vdots \\ F_m P_m \operatorname{diag}(w) S_{-r_R} \end{bmatrix}, \tag{3.9}$$

with $M = m|\mathcal{R}|$. If $\mathcal{R} = [d]$, the matrix A corresponds to the *Discrete Short-Time Fourier Transform* (STFT) with window w. If $\mathcal{R} \neq [d]$, it is a subsampled STFT matrix.
Thus, the discrete ptychographic problem is a special case of the phase retrieval problem and sometimes referred to as the *ptychographic phase retrieval* or *STFT phase retrieval*.

Notation. *The blockwise representation of A allows to treat $Y \in \mathbb{R}^{m\times|\mathcal{R}|}$ as a vector $y \in \mathbb{R}^{m|\mathcal{R}|}$ with entries $y_{(j,r)} := Y_{j,r}$. When working with ptychographic measurements (PTY), we will use the notation $Y, N \in \mathbb{R}^{m\times|\mathcal{R}|}$ to refer to the matrix form of the measurements and the noise, while $y, n \in \mathbb{R}^{m|\mathcal{R}|}$ denotes their vector form.*

Notes and References. *In the derivation of the formula* (PTY)*, it is assumed that both the window* $\boldsymbol{w}$ *and the object* $\boldsymbol{x}$ *are compactly supported on* $[0,1]$*. This assumption also implies that the support of* $\boldsymbol{w}\mathcal{T}_{-r}\boldsymbol{x}$ *is bounded and due to the uncertainty principles [44, Chapter 2.2], the intensity will have an unbounded support. Yet, practically it is impossible to sample an unbounded wave due to the finite nature of the detector and, thus, the high frequencies are left out which inevitably leads to errors. A detailed study on the arising instabilities can be found in [80].*

3.3 Ambiguities, uniqueness and stability of ptychographic phase retrieval

In this section we discuss the concept of unique reconstruction from ptychographic measurements. Since ptychographic recovery is a special case of the phase retrieval problem, some of the results apply directly. In particular, for the phase retrieval problem it is known that unique recovery is generally not possible. Let

$$\alpha \in \mathbb{T} = \{\beta \in \mathbb{C} : |\beta| = 1\}$$

and observe that for all $x \in \mathbb{C}^d$ the measurements (PR) generated by x and αx coincide,

$$|(A[\alpha x])_k|^2 = |\alpha (Ax)_k|^2 = |(Ax)_k|^2. \tag{3.10}$$

Hence, the unique recovery of $x \in \mathbb{C}^d$ is not possible. The equation (3.10) establishes the equivalence relation

$$x \sim x' \;\Leftrightarrow\; x = \alpha x' \text{ for some } \alpha \in \mathbb{T}.$$

Therefore, a reconstruction is called *unique*, if it belongs to the set

$$\{\alpha x : \alpha \in \mathbb{T}\},$$

which is also referred to as a recovery *up to a global phase*. We note that it is equivalent to considering the unique recovery on the quotient space

$$\mathbb{C}^d/\,\mathbb{T} := \{\{\alpha x : \alpha \in \mathbb{T}\} : x \in \mathbb{C}^d\}.$$

To account for the equivalence relation $\sim$ when measuring the distance between two vectors $x, x' \in \mathbb{C}^d$, we use

$$\operatorname{dist}(x, x') := \min_{|\alpha|=1} \|x - \alpha x'\|_2 . \tag{3.11}$$

Lemma 3.3.1. *The mapping* $\operatorname{dist}(\cdot,\cdot)$ *is a metric on* $\mathbb{C}^d/\mathbb{T}$*.*

Proof. Let $x, x' \in \mathbb{C}^d$. Firstly, $\operatorname{dist}(\cdot,\cdot)$ is non-negative,

$$\operatorname{dist}(x, x') = \min_{|\alpha|=1} \|x - \alpha x'\|_2 \geq 0,$$

and if $\operatorname{dist}(x, x') = 0$, then by definition $x = \alpha x'$, so that $x \sim x'$.

Secondly, it is symmetric as

$$\operatorname{dist}(x, x') = \min_{|\alpha|=1} \|x - \alpha x'\|_2 = \min_{|\beta|=1} \|\beta x - x'\|_2 = \operatorname{dist}(x', x).$$

Finally, we show that $\operatorname{dist}(\cdot,\cdot)$ satisfies the triangle inequality. Let $x, u, v \in \mathbb{C}^d$, Using the triangle inequality for $\|\cdot\|_2$, we obtain

$$\operatorname{dist}(x, u) = \min_{|\alpha|=1} \|x - \alpha u\|_2 = \min_{|\alpha|=1} \|x - \beta v + \beta v - \alpha u\|_2 \leq \|x - \beta v\|_2 + \min_{|\alpha|=1} \|\beta v - \alpha u\|_2 ,$$

for any $\beta \in \mathbb{T}$. By a substitution $\alpha' = \alpha\overline{\beta}$ with $|\alpha'| = 1$, we rewrite the second term as

$$\operatorname{dist}(\beta v, u) = \min_{|\alpha|=1} \|\beta v - \alpha u\|_2 = \min_{|\alpha|=1} \left\|v - \alpha\overline{\beta} u\right\|_2 = \min_{|\alpha'|=1} \|v - \alpha' u\|_2 = \operatorname{dist}(v, u).$$

Hence, we have

$$\operatorname{dist}(x, u) \leq \|x - \beta v\|_2 + \operatorname{dist}(v, u).$$

Since $\beta \in \mathbb{T}$ is arbitrary, we can select β as the minimizer of $\min_{|\beta|=1} \|x - \beta v\|_2$, so that

$$\operatorname{dist}(x, u) \leq \operatorname{dist}(x, v) + \operatorname{dist}(v, u).$$

□

The reconstruction up to a global phase factor is the only known ambiguity occurring in the phase retrieval problem for all $A \in \mathbb{C}^{M\times d}$. However, further ambiguities may arise depending on the choice of A.

Example 3.3.2. *Let A be diagonal, so that $A = \operatorname{diag}(a)$ for some $a \in \mathbb{C}^d$. Then, for all $x \in \mathbb{C}^d$ and $v \in \mathbb{C}^d$ such that $|v_j| = 1, j \in [d]$ we have*

$$|(A(x \circ v))_j| = |a_j x_j v_j| = |a_j x_j| = |(Ax)_j|.$$

Moreover, the same problem arises for A of the form

$$A = \begin{bmatrix} \operatorname{diag}(a^1) \\ \vdots \\ \operatorname{diag}(a^k) \end{bmatrix}, \quad \text{for all } k \in \mathbb{N}, \quad a^1, \ldots, a^k \in \mathbb{C}^d.$$

The last construction of A corresponds to the ptychographic measurements with $\delta = 1$, that is why we require that $\delta > 1$.

Example 3.3.3 ([1, Proposition 2.1])**.** *Let $A = F_d$. Then, for all $x \in \mathbb{C}^d$, $r \in [d]$, the objects x, $S_r x$ and $R_d\overline{x}$ generate the same measurements. In the first case Proposition 2.2.2 yields*

$$|(F_d S_r x)_j|^2 = |(M_{-r} F_d x)_j|^2 = |e^{-\frac{2\pi i r j}{d}} (F_d x)_j|^2 = |(F_d x)_j|^2.$$

In the second case, we apply Proposition 2.2.5, which gives us

$$|(F_d R_d \overline{x})_j|^2 = |(R_d F_d \overline{x})_j|^2 = |(R_d R_d \overline{F_d x})_j|^2 = |(\overline{F_d x})_j|^2 = |(F_d x)_j|^2.$$

Therefore, it is crucial to understand whether or not non-trivial ambiguities arise for the ptychographic measurement matrix A as in (3.9). We note that the unique recovery up to a global phase is equivalent the injectivity of the mapping $\{\alpha x : \alpha \in \mathbb{T}\} \mapsto |Ax|^2$. In the following, we show that weaker requirements on A are not sufficient.
For the linear measurements Ax, we know that injectivity of the matrix A is a necessary and sufficient condition for the unique recovery of x. However, as Example 3.3.3 suggests, for the measurements $|Ax|^2$ with injective matrix A equal to F_d, non-trivial ambiguities are present. Hence, the injectivity of A is not sufficient to ensure unique reconstruction. On the other hand, it is a necessary condition.

Lemma 3.3.4. *Let us consider the noiseless phase retrieval measurements* (PR). *If for all $x \in \mathbb{C}^d$ the matrix A admits the unique reconstruction of x up to a global phase, then A is injective.*

Proof. Assume that A is not injective. Let $\mathbb{0}_d \in \mathbb{C}^d$ be the vector with zero entries. Then, there exists $x \in \mathbb{C}^d, x \neq \mathbb{0}_d$ such that $(Ax)_j = 0 = (A\mathbb{0}_d)_j$ for all $j \in [m]$. It implies that the measurements coincide, $|(Ax)_j|^2 = |(A\mathbb{0}_d)_j|^2$. However, A admits the unique reconstruction up to a global phase and, thus, $x = \alpha\mathbb{0}_d = \mathbb{0}_d$, which contradicts $x \neq \mathbb{0}_d$. Therefore, A is injective. □

In the case of ptychographic measurement matrix (3.9), the injectivity of A has the following characterization.

Lemma 3.3.5 (Necessary and sufficient conditions for injectivity of ptychographic measurements)**.** *Consider the ptychographic measurements* (PTY) *and the corresponding measurement matrix A* (3.9)*. Let v be the vector in $\mathbb{R}^d$ with entries*

$$v_j = m \sum_{r \in \mathcal{R}} |(S_r w)_j|^2 \text{ for all } j \in [d],$$

where $\mathcal{R}$ denotes the set of observed shift positions. Then, $A^*A = \operatorname{diag}(v)$ and A is injective if and only if $v_j > 0$ for all $j \in [d]$.

Proof. For the intensity measurements (PTY) we compute the matrix A^*A using the block representation (3.9), which expands the derivation provided in [36, p.6]. More precisely,

$$\begin{aligned} A^*A &= \sum_{r \in \mathcal{R}} (F_m P_m \operatorname{diag}(w) S_{-r})^* F_m P_m \operatorname{diag}(w) S_{-r} \\ &= \sum_{r \in \mathcal{R}} (\operatorname{diag}(w) S_{-r})^* P_m^* F_m^* F_m P_m \operatorname{diag}(w) S_{-r}. \end{aligned}$$

By Proposition 2.2.1, $F_m^* F_m = m I_d$ and the summands simplify to

$$A^*A = m \sum_{r \in \mathcal{R}} (\operatorname{diag}(w) S_{-r})^* P_m^* P_m \operatorname{diag}(w) S_{-r}.$$

Since $\operatorname{supp}(w) = [\delta]$ and in view of (2.12) we obtain

$$A^*A = m \sum_{r \in \mathcal{R}} (\operatorname{diag}(w) S_{-r})^* \operatorname{diag}(w) S_{-r}.$$

For the circular shift operators, the identity

$$(\mathrm{diag}(w)S_{-r}u)_j = w_j u_{j+r} = ((S_r w) \circ u)_{j+r} = (S_{-r}\,\mathrm{diag}(S_r w)u)_j,$$

holds for all $u \in \mathbb{C}^d$ and, thus,

$$\begin{aligned}(\mathrm{diag}(w)S_{-r})^*\,\mathrm{diag}(w)S_{-r} &= (S_{-r}\,\mathrm{diag}(S_r w))^* S_{-r}\,\mathrm{diag}(S_r w)\\ &= \mathrm{diag}(S_r w)^* S_{-r}^* S_{-r}\,\mathrm{diag}(S_r w).\end{aligned}$$

Furthermore, by Proposition 2.2.2 the circular shift operator satisfies $S_{-r}^* S_{-r} = I_d$ and the adjoint of $\mathrm{diag}(S_r w)$ is given by $\mathrm{diag}(\overline{S_r w})$. Hence,

$$\mathrm{diag}(S_r w)^* S_{-r}^* S_{-r}\,\mathrm{diag}(S_r w) = \mathrm{diag}(\overline{S_r w})\,\mathrm{diag}(S_r w) = \mathrm{diag}(|S_r w|^2).$$

Finally, we note that the sum of diagonal matrices is a diagonal matrix of the sum and, therefore,

$$A^*A = m\sum_{r\in\mathcal{R}} \mathrm{diag}(|S_r w|^2) = \mathrm{diag}\left(m\sum_{r\in\mathcal{R}} |S_r w|^2\right) = \mathrm{diag}(v).$$

Turning to the injectivity claim, let us assume that A is injective. Then, $\mathrm{rank}\, A = d \le M$ and $A^*A = \mathrm{diag}(v)$ is invertible, which implies that $v_j > 0$ for all $j \in [d]$.
For the reverse claim, assume that $v_j > 0$ for all $j \in [d]$. By construction of v, the inequality $v_j > 0$ is true if there exists a shift position r, such that $(S_r w)_j \neq 0$. Since $\mathrm{supp}(w) \subseteq [\delta]$, per single shift there are at most δ entries of v which are non-zero. Therefore, at least $\lceil d/\delta \rceil$ shifts are needed if $v_j > 0$ for all $j \in [d]$. Hence, the number of measurements satisfies

$$M = m|\mathcal{R}| \ge m\lceil d/\delta \rceil \ge md/\delta \ge d,$$

and the invertibility of $A^*A = \mathrm{diag}(v)$ implies that A is injective. □

The result of Lemma 3.3.5 has a physical explanation. The entries of the vector v_j can be interpreted as the total intensity of light passing through the part of the object corresponding to the entry x_j. If some part was not illuminated ($v_j = 0$), there is no information about it in the measurements and, consequently, it cannot be identified.
Following our claim that the injectivity of A is not sufficient, the next example shows that an even stronger requirement, namely the injectivity of the mapping
$\{\alpha x : \alpha \in \mathbb{T}\} \mapsto \{\alpha Ax : \alpha \in \mathbb{T}\}$ is not sufficient.

Example 3.3.6. *Consider two vectors $x, x' \in \mathbb{C}^d$ such that $\mathrm{dist}(x, x') > 0$. Assume that $Ax' = Ax \circ v$ with entries of vector v satisfying $|v_j| = 1$. Furthermore, assume that for a pair $j, k \in [M], j \neq k$, the entries of v do not coincide, $v_j \neq v_k$, and the corresponding values $|(Ax)_j|, |(Ax)_k|$ are non-zero. Then $|Ax'| = |Ax| \circ |v| = |Ax|$, while*

$$\begin{aligned}\mathrm{dist}^2(Ax', Ax) &= \min_{|\alpha|=1} \|Ax' - \alpha Ax\|_2^2 \ge \min_{|\alpha|=1} \left[|(Ax')_k - \alpha(Ax)_k|^2 + |(Ax')_j - \alpha(Ax)_j|^2\right]\\ &= \min_{|\alpha|=1} \left[|(Ax)_k|^2|v_k - \alpha|^2 + |(Ax)_j|^2|v_j - \alpha|^2\right]\\ &\ge \min\{|(Ax)_k|^2, |(Ax)_j|^2\} \cdot \min_{|\alpha|=1} \left[|v_k - \alpha|^2 + |v_j - \alpha|^2\right]\\ &\ge \min\{|(Ax)_k|^2, |(Ax)_j|^2\} \cdot |v_k - v_j|^2/2 > 0,\end{aligned}$$

where we have used

$$|\beta - \gamma|^2 \leq (|\beta| + |\gamma|)^2 \leq 2(|\beta|^2 + |\gamma|^2)$$

for $\beta, \gamma \in \mathbb{C}$ *to obtain the last inequality.*

Therefore, the injectivity of the mapping $\{\alpha x : |\alpha| = 1\} \mapsto |Ax|^2$ has to be considered in order to ensure unique reconstruction. Results of [81] established that if the matrix A corresponds to a generic frame with $M \geq 4d - 2$, the mapping $\{\alpha x : |\alpha| = 1\} \mapsto |Ax|^2$ is injective with probability 1. Later, it was conjectured [82] and proved [83] that $M \geq 4d - 4$ measurements will suffice. It also was shown recently that $M \geq 4d$ ptychographic measurements (PTY) are sufficient for the unique recovery, when both x and w are generic, i.e., do not belong to a set of zero measure [84]. This set, however, is non-descriptive.
In the case of ptychographic measurements (PTY), the matrix A or the window w are generally non-random and may belong to an event of probability zero. The next example suggests that for the (non-random) measurements (PTY), a non-generic x can be constructed such that a unique recovery is not possible.

Example 3.3.7. *Consider the ptychographic measurements* (PTY). *Let* $d \geq 2\delta$ *and consider two objects* $x^+, x^- \in \mathbb{C}^d$ *such that*

$$x_0^+ = x_0^- = 1, \quad x_\delta^+ = 1, x_\delta^- = -1, \quad x_j^+ = x_j^- = 0, \ j \in [d] \backslash \{0, \delta\}.$$

Let us show that $\operatorname{dist}(x^+, x^-) > 0$. *By definition, we have*

$$\operatorname{dist}^2(x^+, x^-) = \min_{|\alpha|=1} \left\|x^+ - \alpha x^-\right\|_2^2 = \min_{|\alpha|=1} |1 - \alpha|^2 + |1 + \alpha|^2 = \min_{|\alpha|=1} 2 + 2|\alpha|^2 = 4.$$

The intensity measurements for x^+ *are given by*

$$I_{j,r}^{m,+} = \left| \sum_{k \in [\delta]} w_k x_{k+r}^+ e^{-\frac{2\pi i k j}{m}} \right|^2$$

We note that at most one summand in the sum is non-zero, either x_0^+ *or* x_δ^+. *The distance between two indices (taking into account circularity) is*

$$|0 - \delta|_c = \min\{\delta, d - \delta\} \geq \min\{\delta, 2\delta - \delta\} = \delta,$$

and due to $\operatorname{supp}(w) \subseteq [\delta]$, *they never appear in the sum simultaneously. Therefore, we simplify* $I_{j,r}^{m,+}$ *as*

$$I_{j,r}^{m,+} = \begin{cases} |w_{-r}|^2, & r \in \{-\delta + 1, -\delta + 2, \ldots, 0\} \\ |w_{\delta - r}|^2, & r \in \{1, 2, \ldots, \delta\} \\ 0, & \text{otherwise.} \end{cases}$$

Since $I_{j,r}^{m,+}$ *is independent of the sign of* x_δ^+, *the measurements* $I_{j,r}^{m,-}$ *for* x^- *will be equal to* $I_{j,r}^{m,+}$ *and, thus, the unique recovery from the ptychographic measurements* (PTY) *is not possible.*

Example 3.3.7 shows that due to the local nature of the ptychographic measurements (PTY), i.e., $\operatorname{supp}(w) \subseteq [\delta]$, the non-trivial ambiguities arise when x has some zero entries. The next result provides sufficient conditions for the uniqueness of reconstruction for non-vanishing objects, that is $|x_j| > 0$ for all $j \in [d]$.

Theorem 3.3.8 ([85, Theorem 2.4], also [86, Proposition III.4]). *Consider the ptychographic measurements* (PTY) *and the corresponding measurement matrix* A (3.9). *Let* $m = d$ *and consider the set of shifts* $\mathcal{R} = [d]$. *Assume that*

$$[F_d(w \circ S_r \overline{w})]_j \neq 0 \text{ for } r = 0, 1, \ j \in [d].$$

Then, the mapping $\{\alpha x : \alpha \in \mathbb{T}\} \mapsto |Ax|^2$ *is injective for all non-vanishing* $x \in \mathbb{C}^d$.

Theorem 3.3.8 only requires minor assumptions on the window and, in payoff, significantly restricts x. The latest result, [87, Corollary 2.5], provides a tradeoff between the number of allowed consequent zeros in x and the number of shifts r such that $[F_d(w \circ S_r \overline{w})]_j$ are non-zero. Moreover, the notion of uniqueness may be extended to account for ambiguities arising from the structure of $\operatorname{supp}(x)$ [87].
While the uniqueness of ptychographic reconstruction allows to determine the set $\{\alpha x : \alpha \in \mathbb{T}\}$ from the measurements, it is also desirable that the inverse mapping $|Ax|^2 \mapsto \{\alpha x : \alpha \in \mathbb{T}\}$ is continuous. It implies that for two sets of measurements which are alike, the corresponding objects should be similar. This notion is known as the *stability* of phase retrieval. The problem is called *stable* with respect to a norm $\|\cdot\|$ if there exists a constant $C > 0$ such that for any $x, x' \in \mathbb{C}^d$ the inequality

$$\operatorname{dist}(x, x') \leq C \, \| |Ax| - |Ax'| \| \tag{3.12}$$

holds. If the constant C depends on x, the problem is said to be *locally stable*. Note that the inequality (3.12) implies the uniqueness of reconstruction, if $|Ax| = |Ax'|$. The converse also holds for the discrete version of the phase retrieval problem. By [88, Proposition 1.4], if the problem has a unique solution for every $x \in \mathbb{C}^d$, then there exists a constant $C > 0$ such that the inequality (3.12) holds. However, in view of Example 3.3.7, this result does not apply to the discrete ptychographic problem. Nevertheless, the stability constant C can be estimated for subsets of $\mathbb{C}^d$ [89, 87].

Notes and References. *The stable reconstruction is often studied in the context of random matrices. Besides generic* A *[81, 83], a common choice is the matrix* A *with rows drawn at random from a certain distribution, for which it suffices to have* $M = \mathcal{O}(d)$ *measurements to achieve stability for real-valued objects [90, 91] as well as for complex-valued objects [92]. For more structured random measurements the required sampling complexity is near-optimal,* $M = \mathcal{O}(d \log d)$ *[93, 94].*
The uniqueness of reconstruction from the ptychographic measurements (PTY) *was addressed by an earlier result [95], already mentioned [85, 86, 87] and graph-theoretical analysis [96]. The authors of [97, 98, 99] also consider an alternative measurement scenario with non-circular shifts. An overview of these results can be found in [100].*
The concept of uniqueness and stability is broadly studied for the continuous ptychographic problem (3.1). *It is well-known that injectivity is guaranteed, when the ambiguity function of the window* $\boldsymbol{w}$ *is almost everywhere non-vanishing [71, 70]. Further relaxed requirements were derived under additional assumptions on the object* $\boldsymbol{x}$ *belonging to shift invariant subspaces [71] or Paley-Wiener spaces [70]. In [67], the authors derive injectivity for subsampled continuous measurements in the special case of the Gabor transform.*
While in the discrete case the uniqueness of reconstruction implies stability, the continuous problem is known to be unstable. Several works derived its instability in general Hilbert

[88] and Banach spaces [60] and specifically for the Gabor phase retrieval [101]. Moreover, in [88] it was shown that for the discrete approximation of the continuous problem, the stability constant degrades exponentially in the dimension of the approximation.
In order to achieve stability, the regions in which the measurements are non-vanishing has to be considered [62]. Similar ideas are used in [72, 73, 102] to establish the stability of the Gabor phase retrieval in local sense, i.e., the stability constant C is depending on $\boldsymbol{x}$.
For an overview of results on injectivity and stability of the ptychographic phase retrieval we refer the reader to [68].

3.4 Overview of recovery algorithms

Since the introduction of the phase retrieval problem, many recovery methods were developed and studied in the literature. All these algorithms can be classified into several categories.
The first major group consists of projection methods, which consider the recovery from the phase retrieval measurements as a problem of finding a point in an intersection of two or more sets. One of the sets imposes that the measurements are satisfied and the other provide additional constraints on the object. The most known among these methods is Error Reduction [37, 103, 26, 104, 38, 105], which will be discussed in greater detail in Section 3.5.2. Other prominent algorithms are Hybrid Input-Output [27, 103], Difference Map [28], Averaged Successive Reflections [29], and Relaxed Averaged Alternating Reflectors [30] and many more. Some of the above-mentioned methods can be linked to the Douglas-Rachford splitting [106, 107, 108, 29, 109], an optimization-based approach towards finding the intersection of two sets as a minimizer of a sum of two functions.
Furthermore, the recovery from phase retrieval and ptychographic measurements is often posed as an optimization of a certain (in general non-convex) loss function and, thus, the reconstruction approaches are categorized based on the used optimization method.
The gradient methods for phase retrieval are mainly represented by Wirtinger and Amplitude flow algorithms [31, 32], which minimize the squared loss based on the magnitude of the measurements (see Section 3.5.1). Later studies introduced variants of these algorithms [110, 32, 111, 112, 113] with modified loss functions, that allowed then to avoid stagnation and obtain faster convergence. An overview on many recent developments for first order methods can be found in [114]. Some of the algorithms derived specifically for ptychography [26, 39, 33, 115], such as Ptychographic Iterative Engine [39], belong to the class of gradient-based techniques. We explain, why this inclusion is reasonable in Section 3.5.3. While first order optimization methods are dominant in the literature, second order approaches for phase retrieval were also considered, see [116, 117].
The third group is alternating minimization methods [118, 119, 120, 117, 121]. They introduce supplementary variables to the optimization problem, such that the minimization with respect to a single variable is much easier than the simultaneous minimization of all variables. Then, for each iteration, the algorithm selects a single variable and performs optimization with respect to it. The two most prominent subgroups are Alternating Direction Method of Multipliers frequently applied to the augmented Lagrangian function [118, 119] and Majorization-Minimization algorithms also known as Iteratively Reweighted Least Squares [117, 120, 121]. The latter computes weights and then solves

a weighted least squares problem.
The next group of algorithms is based on the relaxation of the original optimization problem [34, 122, 123, 124, 125, 126, 92, 91, 127, 128, 129]. The relaxed problem is convex, hence, easier to solve and its solution can be computed in polynomial time. If the relaxation is tight, the solution of the relaxation coincides with the solution of the original problem. A well-known instance of such techniques is PhaseLift [34, 122], which poses the phase retrieval problem as a recovery of the rank-one matrix via rank minimization. The rank of a matrix is a non-convex function and by replacing it with the nuclear norm, a computationally feasible convex relaxation is obtained. Another example is the PhaseMax algorithm [127], which relaxes recovery to linear programming.
With the rise of deep learning, neural networks were also applied to solve ptychography and the phase retrieval problem [130, 131, 132, 133, 134, 135]. These applications include a pure reconstruction via neural network [132], mixed approaches [133], a usage of generative priors [131] and unrolling techniques [135].
Lastly, we would like to mention some methods, which do not belong to any of the categories above. Phase retrieval via polarization [136, 38, 137] recovers the unknown phases of the measurements using synchronization and then inverts the Short-Time Fourier transform. Another approach is the Block Phase Retrieval algorithm [35, 138, 139, 40, 140, 141, 142], which aims to reconstruct the object via the direct (non-iterative) inversion of the measurements. Section 3.6 of this chapter is dedicated to Block Phase Retrieval.

3.5 Iterative Methods

3.5.1 Amplitude Flow

In this section we discuss the Amplitude Flow algorithm (AF), which performs gradient-based minimization of the amplitude-based loss

$$\mathcal{L}_2(z) := \mathcal{L}_2(z; \mathcal{Q}) = \sum_{k\in[M]} |\sqrt{z^* Q_k z} - \sqrt{y_k}|^2, \quad z \in \mathbb{C}^d, \tag{3.13}$$

where $\mathcal{Q} = \{Q_k\}_{k\in[M]} \subset \mathbb{H}^d$ is a family of Hermitian positive semidefinite measurement matrices and $y \in \mathbb{R}^M$ is the vector containing noisy measurements $x^* Q_k x$, $k \in [M]$. In the case of the phase retrieval problem, the family corresponding to the measurement matrix A is given by rank-one matrices $Q_k = a_k a_k^*$ with a_k being the conjugate of the rows of the matrix A, so that $(Az)_k = a_k^* z$, $z^* Q_k z = |(Az)_k|^2$, and y is given by (PR). Therefore, when working with the phase retrieval measurements, we will either use a short notation $\mathcal{L}_2$ or $\mathcal{L}_2(z; A)$ if matrix specification is necessary. While the introduction of $\mathcal{Q}$ may seem artificial to the reader at this point, it allows to generalize the original proofs for AF derived in [36] to other settings, such as polychromatic ptychography discussed in Chapter 5.
The function $\mathcal{L}_2$ is non-negative and in absence of noise we have $\mathcal{L}_2(x) = 0$. If the mapping $\{\alpha z : |\alpha| = 1\} \mapsto |Az|^2$ is injective, the set $\{\alpha x : |\alpha| = 1\}$ contains all global minimizers of $\mathcal{L}_2$ and, therefore, the object x can be recovered by minimizing the function $\mathcal{L}_2$. Furthermore, if the problem is stable (3.12), even local minimizers of $\mathcal{L}_2$ with a small function value provide a good approximation to x.

The Amplitude Flow algorithm is a gradient descent scheme of the form (2.16) based on the Wirtinger derivatives, which are reviewed in Section 2.3. That is, given an initial guess $z^0 \in \mathbb{C}^d$, AF constructs a sequence of iterates $\{z^t\}_{t\geq 0}$ via

$$z^{t+1} = z^t - \mu_t \nabla_z \mathcal{L}_2(z^t), \quad t \geq 0, \tag{AF}$$

where $\mu_t > 0$ denotes the so-called learning rate and $\nabla_z \mathcal{L}_2$ is the generalized Wirtinger gradient of $\mathcal{L}_2$. The iteration process is continued until the gradient $\nabla_z \mathcal{L}_2(z^t)$ vanishes, which is equivalent to reaching a fixed point $z^{t+1} = z^t$.

We note that the function $\mathcal{L}_2$ is not differentiable at points where $z^* Q_k z = 0$ for some $k \in [M]$ and a workaround is necessary to adapt the notion of the gradient for $\mathcal{L}_2$. For this reason we consider a smoothed square loss for general quadratic measurements

$$\mathcal{L}_{2,\varepsilon}(z) := \mathcal{L}_{2,\varepsilon}(z; \mathcal{Q}) = \sum_{k\in[M]} \left| \sqrt{z^* Q_k z + \varepsilon} - \sqrt{y_k + \varepsilon} \right|^2, \tag{3.14}$$

where $\varepsilon \geq 0$ is a smoothing parameter.

The function $\mathcal{L}_{2,\varepsilon}$ possesses some useful properties. Firstly, $\mathcal{L}_{2,\varepsilon}$ is continuous in ε and, in particular, it is right-continuous at $\varepsilon = 0$,

$$\mathcal{L}_2(z) = \mathcal{L}_{2,0}(z) = \lim_{\varepsilon \to 0+} \mathcal{L}_{2,\varepsilon}(z).$$

Secondly, if $\varepsilon > 0$ we can compute the gradient of $\mathcal{L}_{2,\varepsilon}$ everywhere and properly define the generalized gradient of $\mathcal{L}_2$ as the limit of the gradients as ε tends to zero.

Lemma 3.5.1. *Let $\varepsilon > 0$. The function $\mathcal{L}_{2,\varepsilon}$ is continuously differentiable with the gradient*

$$\nabla_z \mathcal{L}_{2,\varepsilon}(z) = \sum_{k\in[M]} \left(1 - \frac{\sqrt{y_k + \varepsilon}}{\sqrt{z^* Q_k z + \varepsilon}}\right) Q_k z.$$

Furthermore, the generalized gradient of $\mathcal{L}_2$ is defined as the pointwise limit

$$\nabla_z \mathcal{L}_2(z) := \lim_{\varepsilon \to 0+} \nabla_z \mathcal{L}_{2,\varepsilon}(z) = \sum_{k\in[M]} \left(1 - \frac{\sqrt{y_k}}{\sqrt{z^* Q_k z}}\right) Q_k z,$$

where division $0/0$ is set to 0, when $Q_k z = 0$.

Proof. A single summand of $\mathcal{L}_{2,\varepsilon}$ is given by

$$f_k(z) := \left| \sqrt{\bar{z}^T Q_k z + \varepsilon} - \sqrt{y_k + \varepsilon} \right|^2 = \left| \sqrt{z^T Q_k^T \bar{z} + \varepsilon} - \sqrt{y_k + \varepsilon} \right|^2, \quad k \in [M].$$

The gradient of f_k can be evaluated by the chain rule (Theorem 2.3.2). We get

$$\begin{aligned}\nabla_z f_k(z) = \left[\frac{\partial f_k}{\partial \bar{z}}(z)\right]^T &= \left[\frac{\partial|\sqrt{z^TQ_k^T\bar{z}+\varepsilon}-\sqrt{y_k+\varepsilon}|^2}{\partial(\sqrt{z^TQ_k^T\bar{z}+\varepsilon}-\sqrt{y_k+\varepsilon})}\right. \\ &\qquad \left.\cdot\frac{\partial\sqrt{z^TQ_k^T\bar{z}+\varepsilon}-\sqrt{y_k+\varepsilon}}{\partial z^TQ_k^T\bar{z}+\varepsilon}\cdot\frac{\partial z^TQ_k^T\bar{z}+\varepsilon}{\partial\bar{z}}\right]^T \\ &= 2\left(\sqrt{z^TQ^T\bar{z}+\varepsilon}-\sqrt{y_k+\varepsilon}\right)\frac{1}{2\sqrt{z^TQ_k^T\bar{z}+\varepsilon}}\left[z^TQ^T\right]^T \\ &= \left(1-\frac{\sqrt{y_k+\varepsilon}}{\sqrt{z^TQ_k^T\bar{z}+\varepsilon}}\right)Q_kz = \left(1-\frac{\sqrt{y_k+\varepsilon}}{\sqrt{z^*Q_kz+\varepsilon}}\right)Q_kz.\end{aligned}$$

Then, by the linearity of the derivatives,

$$\nabla_z\mathcal{L}_{2,\varepsilon}(z) = \sum_{k\in[M]}\nabla_z f_k(z) = \sum_{k\in[M]}\left(1-\frac{\sqrt{y_k+\varepsilon}}{\sqrt{z^*Q_kz+\varepsilon}}\right)Q_kz.$$

For the generalized gradient of $\mathcal{L}_2$ we consider two cases. If $Q_kz\neq 0$ for all $k\in[M]$, then $\sqrt{y_k+\varepsilon}/\sqrt{z^*Q_kz+\varepsilon}\to\sqrt{y_k}/\sqrt{z^*Q_kz}$ as $\varepsilon\to 0+$. Note that in this case, $\mathcal{L}_2$ is differentiable at z and its gradient coincides with the limit of $\nabla\mathcal{L}_{2,\varepsilon}(z), \varepsilon\to 0+$. On the other hand, if $Q_kz=0$ for some $k\in[M]$, then we have

$$\frac{\sqrt{y_k+\varepsilon}}{\sqrt{z^*Q_kz+\varepsilon}}Q_kz = \frac{\sqrt{y_k+\varepsilon}}{\sqrt{0+\varepsilon}}\cdot 0 = 0\to 0 = \frac{y_k}{\sqrt{z^*Q_kz}}Q_kz, \quad \varepsilon\to 0+,$$

with ambiguity $0/0$ resolved as 0. □

For the case of phase retrieval we get the following gradient formulas.

Corollary 3.5.2 (Gradient formulas for phase retrieval)**.** *Let $\varepsilon>0$. For the phase retrieval measurements (PR), the gradients of $\mathcal{L}_{2,\varepsilon}$ and $\mathcal{L}_2$ are given by*

$$\nabla_z\mathcal{L}_{2,\varepsilon}(z;A) = A^*\left[I_M-\operatorname{diag}\left(\frac{\sqrt{y+\varepsilon}}{\sqrt{|Az|^2+\varepsilon}}\right)\right]Az$$

and

$$\nabla_z\mathcal{L}_2(z;A) = A^*\left[Az-\operatorname{sgn}_0(Az)\circ\sqrt{y}\right],$$

respectively.

Proof. Using that $Q_k=a_ka_k^*$ and the results of Lemma 3.5.1, we have

$$\begin{aligned}\nabla_z\mathcal{L}_{2,\varepsilon}(z) &= \sum_{k\in[M]}\left[1-\frac{\sqrt{y_k+\varepsilon}}{\sqrt{z^*a_ka_k^*z+\varepsilon}}\right]a_ka_k^*z \\ &= \sum_{k\in[M]}\left[1-\frac{\sqrt{y_k+\varepsilon}}{\sqrt{|(Az)_k|^2+\varepsilon}}\right]a_k(Az)_k = A^*\left[I_M-\operatorname{diag}\left(\frac{\sqrt{y+\varepsilon}}{\sqrt{|Az|^2+\varepsilon}}\right)\right]Az.\end{aligned}$$

The formula for $\nabla_z \mathcal{L}_2(z)$ is obtained by taking the limit $\varepsilon \to 0$, so that

$$\begin{aligned}\nabla_z \mathcal{L}_2(z) &= \sum_{k\in[M]} \left[1 - \frac{\sqrt{y_k}}{|(Az)_k|}\right] a_k (Az)_k \\ &= \sum_{k\in[M]} a_k \left[(Az)_k - \sqrt{y_k}\frac{(Az)_k}{|(Az)_k|}\right] = A^* \left[Az - \operatorname{sgn}_0(Az) \circ \sqrt{y}\right],\end{aligned}$$

where the mapping of 0 to 0 in sgn_0 is due to $0/0$ being set to 0. □

In order to analyze the gradient descent for the non-smooth function $\mathcal{L}_2$, we first consider the gradient descent for the function $\mathcal{L}_{2,\varepsilon}$ with iterations given by

$$z^{t+1} = z^t - \mu_t \nabla_z \mathcal{L}_{2,\varepsilon}(z^t), \quad t \geq 0. \tag{AF$_\varepsilon$}$$

If the learning rate μ_t is chosen as a constant or via Algorithm 1, Theorem 2.3.6 guarantees the convergence of AF$_\varepsilon$ when the Hessian matrix satisfies (2.17), which is shown to be fulfilled in the next lemma.

Lemma 3.5.3. *Let $\varepsilon > 0$. The function $\mathcal{L}_{2,\varepsilon}$ is twice continuously differentiable and its Hessian matrix satisfies*

$$\begin{bmatrix} v \\ \bar{v} \end{bmatrix}^* \nabla^2 \mathcal{L}_{2,\varepsilon}(z) \begin{bmatrix} v \\ \bar{v} \end{bmatrix} \leq 2v^* \sum_{k\in[M]} Q_k v \leq \left\| \sum_{k\in[M]} Q_k \right\|_\infty \left\| \begin{bmatrix} v \\ \bar{v} \end{bmatrix} \right\|_2^2 \quad \text{for all } z, v \in \mathbb{C}^d.$$

Proof. First, we compute the second order derivatives $\nabla^2_{z,z}\mathcal{L}_{2,\varepsilon}$ and $\nabla^2_{\bar{z},z}\mathcal{L}_{2,\varepsilon}$. Using that $Q_k = Q_k^*$, we obtain

$$\begin{aligned}\nabla_{z,z}\mathcal{L}_{2,\varepsilon}(z) &= \frac{\partial}{\partial z} \sum_{k\in[M]} \left[Q_k z - \frac{\sqrt{y_k} Q_k z}{\sqrt{z^* Q_k z + \varepsilon}}\right] \\ &= \sum_{k\in[M]} \left[Q_k - \frac{\sqrt{y_k} Q_k}{\sqrt{z^* Q_k z + \varepsilon}} + \frac{\sqrt{y_k} Q_k z z^* Q_k}{2(z^* Q_k z + \varepsilon)^{3/2}}\right] \\ &= \sum_{k\in[M]} \left[Q_k - \frac{\sqrt{y_k} Q_k}{\sqrt{z^* Q_k z + \varepsilon}} + \frac{\sqrt{y_k} Q_k z z^* Q_k^*}{2(z^* Q_k z + \varepsilon)^{3/2}}\right], \\ \nabla_{\bar{z},z}\mathcal{L}_{2,\varepsilon}(z) &= \frac{\partial}{\partial \bar{z}} \sum_{k\in[M]} \left[Q_k z - \frac{\sqrt{y_k} Q_k z}{\sqrt{z^T Q_k^T \bar{z} + \varepsilon}}\right] = \sum_{k\in[M]} \frac{\sqrt{y_k} Q_k z z^T Q_k^T}{2(z^* Q_k z + \varepsilon)^{3/2}}.\end{aligned}$$

Moreover, as $\mathcal{L}_{2,\varepsilon}$ is real-valued, (2.22) leads to

$$\begin{aligned}&\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{L}_{2,\varepsilon}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} = 2 \operatorname{Re}\left(u^* \nabla^2_{z,z}\mathcal{L}_{2,\varepsilon} u + u^* \nabla^2_{\bar{z},z}\mathcal{L}_{2,\varepsilon} \bar{u}\right) \\ &= 2 \sum_{k\in[M]} \left[u^* Q_k u - \frac{\sqrt{y_k} u^* Q_k u}{\sqrt{z^* Q_k z + \varepsilon}} + \frac{\sqrt{y_k} |u^* Q_k z|^2}{2(z^* Q_k z + \varepsilon)^{3/2}} + \frac{\sqrt{y_k}\, \Re(u^* Q_k z)^2}{2(z^* Q_k z + \varepsilon)^{3/2}}\right]\end{aligned}$$

Furthermore, $\mathrm{Re}(\alpha) \leq |\alpha|$ yields

$$\begin{aligned}\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{L}_{2,\varepsilon}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} &\leq 2 \sum_{k\in[M]} \left[u^* Q_k u - \frac{\sqrt{y_k} u^* Q_k u}{\sqrt{z^* Q_k z + \varepsilon}} + \frac{\sqrt{y_k} |u^* Q_k z|^2}{(z^* Q_k z + \varepsilon)^{3/2}} \right] \\ &\leq 2 \sum_{k\in[M]} \left[u^* Q_k u - \varepsilon \frac{\sqrt{y_k} u^* Q_k u}{(z^* Q_k z + \varepsilon)^{3/2}} + \frac{\sqrt{y_k} (|u^* Q_k z|^2 - u^* Q_k u \cdot z^* Q_k z)}{(z^* Q_k z + \varepsilon)^{3/2}} \right].\end{aligned}$$

Since Q_k is a positive semidefinite matrix, the second summand is bounded from above by zero. The third term is also non-positive, as Q_k can be written as $Q_k = R_k^* R_k$ and, by the Cauchy-Schwartz inequality we have

$$|u^* Q_k z|^2 \leq \|R_k z\|_2^2 \|R_k u\|_2^2 = z^* R_k^* R_k z \cdot u^* R_k^* R_k u = z^* Q_k z \cdot u^* Q_k u.$$

Thus, we arrive at

$$\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{L}_{2,\varepsilon}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \leq 2 \sum_{k\in[M]} u^* Q_k u = 2 v^* \sum_{k\in[M]} Q_k v \leq \left\| \sum_{k\in[M]} Q_k \right\|_\infty \left\| \begin{bmatrix} v \\ \bar{v} \end{bmatrix} \right\|_2^2.$$

□

Consequently, by applying Theorem 2.3.6 we obtain convergence guarantees for the gradient descent of the smoothed loss and by extension for AF.

Theorem 3.5.4. *Let $\varepsilon > 0$. Set $0 < \mu_c \leq 1/\left\|\sum_{k\in[M]} Q_k\right\|_\infty$. Let $\{z^t\}_{t\geq 0}$ be a sequence defined by AF_ε with arbitrary starting point $z^0 \in \mathbb{C}^d$ and learning rates $\mu_t = \mu_t(\mathcal{L}_{2,\varepsilon}, z^{t-1}, \tau, \mu_c, N)$ determined by Algorithm 1. Then, we have*

$$\mathcal{L}_{2,\varepsilon}(z^t) - \mathcal{L}_{2,\varepsilon}(z^{t-1}) \leq -\mu_t \left\|\nabla_z \mathcal{L}_{2,\varepsilon}(z^t)\right\|_2^2,$$

for all $t \geq 1$. In particular,

$$\lim_{t\to\infty} \left\|z^{t+1} - z^t\right\|_2^2 = 0 \quad \text{and} \quad \min_{t\in[T]} \left\|z^{t+1} - z^t\right\|_2^2 \leq \frac{\mathcal{L}_{2,\varepsilon}(z^0)}{T \left\|\sum_{k\in[M]} Q_k\right\|_\infty}$$

for all $T \geq 1$. Furthermore, if the sequence $\{z^t\}_{t\geq 0}$ is instead defined by AF and the learning rates $\mu_t = \mu_t(\mathcal{L}_2, z^{t-1}, \tau, \mu_c, N)$ are determined by Algorithm 1, the inequalities above hold with $\mathcal{L}_{2,\varepsilon}$ and $\nabla_z \mathcal{L}_{2,\varepsilon}$ replaced by $\mathcal{L}_2$ and $\nabla_z \mathcal{L}_2$, respectively.

Proof. The results for AF_ε follow directly from Theorem 2.3.6 and Lemma 3.5.3. We note that $z^{t+1} - z^t = -\mu_c \nabla_z \mathcal{L}_{2,\varepsilon}(z^t)$ and, thus,

$$\min_{t\in[T]} \left\|z^{t+1} - z^t\right\|_2^2 = \mu_c^2 \min_{t\in[T]} \left\|\nabla_z \mathcal{L}_{2,\varepsilon}(z^t)\right\|_2^2 \leq \frac{\mu_c^2 \mathcal{L}_{2,\varepsilon}(z^0)}{\mu_c T} \leq \frac{\mathcal{L}_{2,\varepsilon}(z^0)}{T \left\|\sum_{k\in[M]} Q_k\right\|_\infty}.$$

For AF, let us consider a single iteration for the smoothed loss first,

$$z_\varepsilon^+ = z - \mu_c \nabla_z \mathcal{L}_{2,\varepsilon}(z),$$

for an arbitrary $z \in \mathbb{C}^d$. Then, by what was shown above we have

$$\mathcal{L}_{2,\varepsilon}(z_\varepsilon^+) - \mathcal{L}_{2,\varepsilon}(z) \leq -\mu_c \left\|\nabla_z \mathcal{L}_{2,\varepsilon}(z)\right\|_2^2.$$

Since pointwise $\mathcal{L}_{2,\varepsilon}$ converges to $\mathcal{L}_2$ as ε tends to zero from above, and $\nabla_z \mathcal{L}_2 = \lim_{\varepsilon \to 0+} \nabla_z \mathcal{L}_{2,\varepsilon}$, taking the limit $\varepsilon \to 0+$ gives us

$$\mathcal{L}_2(z^+) - \mathcal{L}_2(z) \leq -\mu_c \left\|\nabla_z \mathcal{L}_2(z)\right\|_2^2,$$

with $z^+ := z - \mu_c \nabla_z \mathcal{L}_2(z)$. Setting $z = z^{t-1}$, we obtain

$$\mathcal{L}_2(z^t) - \mathcal{L}_2(z^{t-1}) \leq -\mu_c \left\|\nabla_z \mathcal{L}_2(z^{t-1})\right\|_2^2$$

for all $t \geq 1$. Therefore, the constant learning rate provides the desired decrease of the loss function. Consequently, $\mu_t = \mu_t(\mathcal{L}_2, z^{t-1}, \tau, \mu_c, N)$ determined by Algorithm 1 will satisfy

$$\mathcal{L}_2(z^t) - \mathcal{L}_2(z^{t-1}) \leq -\mu_t \left\|\nabla_z \mathcal{L}_2(z^{t-1})\right\|_2^2$$

by construction. The rest of the proof repeats the arguments of the proof of Theorem 2.3.6. □

We restate the results of Theorem 3.5.4 specifically for the phase retrieval problem.

Theorem 3.5.5 (Version of [36, Theorem 1])**.** *Consider measurements y of the form (PR). Let $0 < \mu_c \leq \|A\|_\infty^{-2}$ and $z^0 \in \mathbb{C}^d$ be arbitrary. Then, for a sequence $\{z^t\}_{t \geq 0}$ defined by AF and learning rates $\mu_t = \mu_t(\mathcal{L}_2, z^{t-1}, \tau, \mu_c, N)$ determined by Algorithm 1, we have*

$$\mathcal{L}_2(z^t) \geq \mathcal{L}_2(z^{t+1}) \text{ for all } t \geq 0,$$

$$\lim_{t \to \infty} \left\|z^{t+1} - z^t\right\|_2 = 0, \text{ and } \min_{t \in [T]} \left\|z^{t+1} - z^t\right\|_2^2 \leq \frac{\mathcal{L}_2(z^0)}{T \|A\|_\infty^2},$$

for all $T \geq 1$.

Proof. This result follows from Theorem 3.5.4 by observing that in the case of phase retrieval measurements we have

$$\sum_{k \in [M]} Q_k = \sum_{k \in [M]} a_k a_k^* = A^* A, \text{ so that } \left\|\sum_{k \in [M]} Q_k\right\|_\infty = \|A^* A\|_\infty = \|A\|_\infty^2. \tag{3.15}$$

□

Theorem 3.5.5 only guarantees convergence to a fixed point with a sublinear rate and the fixed point is not necessarily a global minimizer of $\mathcal{L}_2$. Therefore, the initialization z^0 is crucial for the convergence to the global minimum. In the case of ptychography, an outcome of a non-iterative method, e.g. the Block Phase Retrieval algorithm discussed in Section 3.6, is a good starting point. Furthermore, with a sufficiently good initialization AF can achieve a linear convergence rate [86].

The computational complexity of AF for $T \in \mathbb{N}$ iterations is given by $\mathcal{O}(TMd)$. If the learning rate is chosen to be $\mu_t = \|A\|^{-2}$, the computation of the spectral norm can be done with additional $\mathcal{O}(Md)$ operations by performing a fixed number of power method iterations. Moreover, in the case of the ptychographic matrix A as in (3.9), the spectral norm can be computed by Lemma 3.3.5 in $\mathcal{O}(|\mathcal{R}|d\delta) \leq \mathcal{O}(Md)$ operations and the computational cost of a single iteration is $\mathcal{O}(|\mathcal{R}|d + m|\mathcal{R}| \log m)$.

Notes and References. *In the literature, gradient-based methods for phase retrieval have been intensively studied in recent years. In the initial work [31], a first order minimization known as the Wirtinger Flow algorithm was analyzed for the intensity loss* $\sum(|(Az)_k|^2 - y_k)^2$ *under the assumption that the entries of* A *are independent complex Gaussian random variables. Later, it was empirically observed that the amplitude-based loss* $\mathcal{L}_2$ *exhibits smaller reconstruction errors [32], which caused many consequent works on the alternations of the loss function* $\mathcal{L}_2$ *[110, 32, 111, 112, 113]. For a survey of gradient based solvers for phase retrieval see [114]. Recently, two algorithms exploring second order approaches for phase retrieval were introduced in [116, 117].*
Originally AF was derived and analyzed for random Gaussian measurements without noise [32]. For such A*, it is possible to construct a good starting point* z^0 *via spectral initialization [31] or null initialization [143], such that AF admits a linear convergence rate to the set of true solutions* $\{\alpha x : |\alpha| = 1\}$*. Weaker convergence guarantees summarized in Theorem 3.5.5 and applicable for any choice of the measurement matrix* A *were later established in [36]. In [86] authors show that for the ptychographic measurements AF will achieve linear convergence in a small neighborhood of the true solution.*
Our main contribution in this section is Theorem 3.5.4, a generalization of the results of [36] for quadratic measurements of the form z^*Q_kz*. While it does not provide any new results for the phase retrieval problem, it extends AF to many other optical scenarios. In particular, we will later use it in Chapter 5. An additional contribution is the analysis of AF for the learning rates determined by Algorithm 1, which are used, for instance, in the phasepack library for MATLAB [144].*

3.5.2 Error Reduction

Error Reduction (ER) is an iterative algorithm for the phase retrieval problem. It considers an initial guess $z^0 \in \mathbb{C}^d$ and performs iterations

$$z^{t+1} = A^\dagger \operatorname{diag}\left(\frac{\sqrt{y}}{|Az^t|}\right) Az^t, \quad t \geq 0. \tag{ER}$$

The iterations are repeated until a fixed point is reached, i.e., $z^{t+1} = z^t$. For $T \in \mathbb{N}$ iterations of ER, $\mathcal{O}(Md^2 + TMd)$ operations are required, where $\mathcal{O}(Md^2)$ operations are needed to compute the pseudoinverse and $\mathcal{O}(Md)$ operations are performed per iteration. Let us consider $u^t := Az^t, t \geq 0$. Then, the iteration of ER reads as

$$u^{t+1} = Az^{t+1} = AA^\dagger \operatorname{diag}\left(\frac{\sqrt{y}}{|Az^t|}\right) Az^t = AA^\dagger \operatorname{diag}\left(\frac{\sqrt{y}}{|u^t|}\right) u^t.$$

In this form, a single iteration of ER is explained as two consecutive projections of u^t.

Lemma 3.5.6 ([38, Lemma 3.15])**.**

1. *Consider the sets*

$$\mathcal{M}_k := \{\alpha \in \mathbb{C} : |\alpha| = \sqrt{y_k}\}, \quad k \in [M].$$

The projection of $\alpha \in \mathbb{C}\backslash\{0\}$ *onto* $\mathcal{M}_k$ *is given by* $\sqrt{y_k} \cdot \operatorname{sgn}(\alpha)$*.*

2. *Consider the set*
$$\mathcal{M} := \{u \in \mathbb{C}^M : |u| = \sqrt{y}\},$$
which is the product of the one-dimensional sets $\mathcal{M}_k$. *The projection of non-vanishing* $u \in \mathbb{C}^M$ *onto* $\mathcal{M}$ *is given by* $\sqrt{y} \circ \operatorname{sgn} u$.

3. *Consider the set* $\operatorname{im}(A)$. *The projection of* $u \in \mathbb{C}^M$ *onto* $\operatorname{im}(A)$ *is given by* $AA^\dagger u$.

Therefore, ER first projects onto $\mathcal{M}$, the set of all vectors $u \in \mathbb{C}^M$ with modulus equal to the measured values. Secondly, it is projected onto $\operatorname{im}(A)$. The sequential projections onto $\mathcal{M}$ and $\operatorname{im}(A)$ allows for interpretation of ER as an alternating projection scheme. If $\mathcal{M}$ was a convex set, then ER would converge to the intersection of two sets [145]. However, due to the non-convexity of $\mathcal{M}$, the convergence of u^t to the intersection of the two sets is not guaranteed, which is a known problem of the ER algorithm. We note that, if A allows for unique recovery and noise is absent, the intersection of $\mathcal{M}$ and $\operatorname{im}(A)$ is given by $\{\alpha x : |\alpha| = 1\}$ [38].
Another complication arising from the non-convexity of $\mathcal{M}$ is the non-uniqueness of the projection onto $\mathcal{M}$. Let $y_k \neq 0$ and consider the projection of $\alpha \in \mathbb{C}$ onto $\mathcal{M}_k$. If α is nonzero, by Lemma 3.5.6, the closest point in $\mathcal{M}_k$ is given by $\sqrt{y_k} \cdot \operatorname{sgn} \alpha$. If $\alpha = 0$, all points in $\mathcal{M}_k$ have the same distance to zero and each of them could be used as the projection. In the literature, it is resolved by setting the projection either to $\sqrt{y_k}$ or $\sqrt{y_k}e^{i\varphi}$ for a randomly selected angle $\varphi \in [0, 2\pi)$. However, we will instead map zero to zero, which is not precisely a projection, but can be interpreted as an average of all possible projections,
$$0 = \frac{1}{2\pi} \int_0^{2\pi} \sqrt{y_k} e^{i\varphi} d\varphi.$$
Therefore, whenever $(Az^t)_k = 0$, we set $(Az^t)_k/|(Az^t)_k| = 0$, which corresponds to the iterations
$$z^{t+1} = A^\dagger[\sqrt{y} \circ \operatorname{sgn}_0(Az^t)], \quad t \geq 0.$$
It is known that if an initial guess z^0 is chosen sufficiently close to the set $\{\alpha x : |\alpha| = 1\}$, the ER algorithm will converge to a point in this set in absence of noise [38, Theorem 3.16]. In general, ER does not converge globally to $\{\alpha x : |\alpha| = 1\}$ [38, p. 830] and, moreover, its convergence to a fixed point is not guaranteed, except of special cases $A = F_d$ [103] or A corresponding to the noncirculant equivalent of (3.9) [26].
For the initialization z^0 of ER, the polarization method can be used [136, 38]. It constructs a matrix with entries approximating $\operatorname{sgn}_0(Ax_k)\overline{\operatorname{sgn}_0(Ax_\ell)}$, $k, \ell \in [M]$, from the measurements and recovers $\operatorname{sgn}_0(Ax_k)$ by solving the phase synchronization problem discussed in Section 3.6.4.
The ER algorithm can also be interpreted as a projected gradient method [38, Section 3.8] applied to the minimization problem
$$\min_{u \in \operatorname{im}(A)} \| |u| - \sqrt{y} \|_2^2 . \tag{3.16}$$

We note that substituting Az for u, $z \in \mathbb{C}^d$ leads to an unconstrained minimization of the amplitude-based loss (3.13), which suggests that ER can be interpreted as a gradient method applied to the function $\mathcal{L}_2$ discussed in Section 3.5.1. We formalize this intuition in the next lemma.

Lemma 3.5.7. *Let A be injective and consider the amplitude-based loss function $\mathcal{L}_2 = \mathcal{L}_2(\cdot\,; A)$ as in (3.13). Then, ER is a scaled gradient method with iterations given by*

$$z^{t+1} = z^t - (A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t), \quad t \geq 0.$$

Proof. Due to the injectivity of A, the identities (2.1) and (2.2) hold. Therefore, the iteration of ER can be rewritten as

$$\begin{aligned} z^{t+1} &= A^\dagger[\sqrt{y} \circ \operatorname{sgn}_0(Az^t)] = A^\dagger Az^t - A^\dagger Az^t + A^\dagger[\sqrt{y} \circ \operatorname{sgn}_0(Az^t)] \\ &= z^t - A^\dagger[Az^t - \sqrt{y} \circ \operatorname{sgn}_0(Az^t)] = z^t - (A^*A)^{-1}A^*[Az^t - \sqrt{y} \circ \operatorname{sgn}_0(Az^t)] \\ &= z^t - (A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t; A), \end{aligned}$$

where in the last line we applied Corollary 3.5.2. □

We emphasize that the result of Lemma 3.5.7 is only true if the ambiguity $0/0$ in the iteration of ER is defined as 0.
The reinterpretation of ER as a scaled gradient method allows to analyze the convergence of the algorithm similarly to AF, which leads to an analogue of Theorem 3.5.5.

Theorem 3.5.8. *Consider measurements y of the form (PR) with an injective measurement matrix A. Then, for the sequence $\{z^t\}_{t\geq 0}$ defined by ER with an arbitrary starting point $z^0 \in \mathbb{C}^d$ we have*

$$\mathcal{L}_2(z^{t+1}) - \mathcal{L}_2(z^t) \leq -\left\|(A^\dagger)^*\nabla_z \mathcal{L}_2(z^t)\right\|_2^2 \text{ for all } t \geq 0,$$

$$\lim_{t\to\infty} \left\|z^{t+1} - z^t\right\|_2 = 0, \text{ and } \min_{t\in[T]} \left\|z^{t+1} - z^t\right\|_2^2 \leq \frac{\mathcal{L}_2(z^0)}{T\sigma_d^2(A)},$$

for all $T \geq 1$, where $\sigma_d(A)$ denotes the smallest singular value of A.

Proof. In view of Lemma 3.5.7, let us consider a smoothed step of the ER algorithm,

$$z_\varepsilon^+ := z - (A^*A)^{-1}\nabla_z \mathcal{L}_{2,\varepsilon}(z).$$

Note that $(A^*A)^{-1}$ exists due to the injectivity of A. Similarly to the proof of Theorem 3.5.4, we first show that a single step of the smoothed Error Reduction does not increase the loss function $\mathcal{L}_{2,\varepsilon}$. Then we take the pointwise limit to obtain the desired result for $\mathcal{L}_2$. In order to derive that for each iteration the loss function does not increase, we apply the Taylor's theorem (2.15) with an arbitrary $z \in \mathbb{C}^d$ and $v = -(A^*A)^{-1}\nabla_z \mathcal{L}_{2,\varepsilon}(z)$. We note that by Lemma 3.5.3, the integral in (2.15) is bounded, as

$$\int_0^1 (1-s) \begin{bmatrix} v \\ \bar{v} \end{bmatrix}^* \nabla^2 \mathcal{L}_{2,\varepsilon}(z + sv) \begin{bmatrix} v \\ \bar{v} \end{bmatrix} ds \leq 2v^*A^*Av \int_0^1 (1-s)ds = v^*A^*Av.$$

Hence, by (2.15), we have

$$\begin{aligned} \mathcal{L}_{2,\varepsilon}(z_\varepsilon^+) &\leq \mathcal{L}_{2,\varepsilon}(z) - 2[\nabla_z \mathcal{L}_{2,\varepsilon}(z)]^*(A^*A)^{-1}\nabla_z \mathcal{L}_{2,\varepsilon}(z) \\ &\quad + [\nabla_z \mathcal{L}_{2,\varepsilon}(z)]^*((A^*A)^{-1})^*(A^*A)(A^*A)^{-1}\nabla_z \mathcal{L}_{2,\varepsilon}(z) \\ &= \mathcal{L}_{2,\varepsilon}(z) - [\nabla_z \mathcal{L}_{2,\varepsilon}(z)]^*(A^*A)^{-1}\nabla_z \mathcal{L}_{2,\varepsilon}(z), \end{aligned}$$

where we used $((A^*A)^{-1})^* = ((A^*A)^*)^{-1} = (A^*A)^{-1}$. Lemma 3.5.1 gives

$$z_\varepsilon^+ \to z^+ := z - (A^*A)^{-1}\nabla_z \mathcal{L}_2(z), \quad \text{as } \varepsilon \to 0+,$$

and, thus, taking the limit $\varepsilon \to 0+$ yields

$$\mathcal{L}_2(z^+) \leq \mathcal{L}_2(z) - [\nabla_z \mathcal{L}_2(z)]^*(A^*A)^{-1}\nabla_z \mathcal{L}_2(z).$$

If z equals z^t of ER, we obtain

$$\begin{aligned} \mathcal{L}_2(z^{t+1}) &\leq \mathcal{L}_2(z^t) - [\nabla_z \mathcal{L}_2(z^t)]^*(A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t) \\ &= \mathcal{L}_2(z^t) - [\nabla_z \mathcal{L}_2(z^t)]^*(A^*A)^{-1}A^*A(A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t) \\ &= \mathcal{L}_2(z^t) - [\nabla_z \mathcal{L}_2(z^t)]^*A^\dagger(A^\dagger)^*\nabla_z \mathcal{L}_2(z^t) \\ &= \mathcal{L}_2(z^t) - \left\|(A^\dagger)^*\nabla_z \mathcal{L}_2(z^t)\right\|_2^2, \end{aligned} \tag{3.17}$$

where we used the identity (2.1).
In order to prove the remaining statements of Theorem 3.5.8, we need to link the decay of the loss function to the iterates. By Lemma 3.5.7, we have that

$$\left\|z^{t+1} - z^t\right\|_2^2 = \left\|(A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t)\right\|_2^2 = [\nabla_z \mathcal{L}_2(z^t)]^*(A^*A)^{-1}(A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t).$$

Since A is injective, its singular value decomposition is given by $A = U\Sigma V^*$ with an orthogonal $U \in \mathbb{C}^{m\times d}$, a unitary $V \in \mathbb{C}^{d\times d}$ and an invertible diagonal matrix $\Sigma \in \mathbb{C}^{d\times d}$. Then,

$$(A^*A)^{-1} = (V\Sigma^2V^*)^{-1} = (V^*)^{-1}\Sigma^{-2}V^{-1} = V\Sigma^{-2}V^* = (V\Sigma^{-1})(V\Sigma^{-1})^*$$

is the singular value decomposition of $(A^*A)^{-1}$. By the definition of the spectral norm, the squared distance between the iterates is bounded from above, as

$$\begin{aligned} \left\|z^{t+1} - z^t\right\|_2^2 &= (\Sigma^{-1}V^*\nabla_z \mathcal{L}_2(z^t))^*\Sigma^{-2}(\Sigma^{-1}V^*\nabla_z \mathcal{L}_2(z^t)) \\ &= \left\|\Sigma^{-1}(\Sigma^{-1}V^*\nabla_z \mathcal{L}_2(z^t))\right\|_2^2 \\ &\leq \left\|\Sigma^{-1}\right\|^2 \left\|\Sigma^{-1}V^*\nabla_z \mathcal{L}_2(z^t)\right\|_2^2 \\ &= \sigma_1^2(\Sigma^{-1})[\nabla_z \mathcal{L}_2(z^t)]^*V\Sigma^{-1}\Sigma^{-1}V^*\nabla_z \mathcal{L}_2(z^t) \\ &= \sigma_d^{-2}(A)[\nabla_z \mathcal{L}_2(z^t)]^*(A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t). \end{aligned}$$

Next, we sum up the norms for $T \geq 1$ iterations of ER and apply (3.17) to obtain

$$\begin{aligned} \sum_{t=0}^{T-1} \left\|z^{t+1} - z^t\right\|_2^2 &\leq \sigma_d^{-2}(A) \sum_{t=0}^{T-1} [\nabla_z \mathcal{L}_2(z^t)]^*(A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t) \\ &\leq \sigma_d^{-2}(A) \sum_{t=0}^{T-1} \left[\mathcal{L}_2(z^t) - \mathcal{L}_2(z^{t+1})\right] \\ &= \sigma_d^{-2}(A) \left[\mathcal{L}_2(z^0) - \mathcal{L}_2(z^T)\right] \leq \sigma_d^{-2}(A)\mathcal{L}_2(z^0), \end{aligned}$$

where in the last line we used that $\mathcal{L}_2(z) \geq 0$ for all $z \in \mathbb{C}^d$. This implies that the partial sum of the series $\sum_{t=0}^\infty \left\|z^{t+1} - z^t\right\|_2^2$ is bounded and, thus, the series converges.

Consequently, the summands $\|z^{t+1} - z^t\|_2^2$ are tending to zero as $t \to \infty$. Furthermore, we have

$$\min_{t\in[T]} \left\|z^{t+1} - z^t\right\|_2^2 \leq \frac{1}{T}\sum_{t=0}^{T-1} \left\|z^{t+1} - z^t\right\|_2^2 \leq \frac{\mathcal{L}_2(z^0)}{T\sigma_d^2(A)},$$

which concludes the proof. □

Theorem 3.5.8 guarantees that, no matter how noisy the measurements are, ER will always converge to a fixed point and the convergence rate is sublinear. However, even in the absence of noise, it does not guarantee global convergence to a point in the set $\{\alpha x : |\alpha| = 1\}$. Comparing Theorem 3.5.8 to Theorem 3.5.5, we observe that the constant in the convergence rate of ER is worse by a factor $\sigma_1^2(A)/\sigma_d^2(A)$ compared to AF.
A further consequence of Lemma 3.5.7 is the equality of the fixed-point sets of both algorithms.

Corollary 3.5.9. *Let A be injective. Then, $z \in \mathbb{C}^d$ is a fixed point of ER if and only if z is the fixed point of AF.*

Proof. Let $z \in \mathbb{C}^d$ be a fixed point of ER. By Lemma 3.5.7, we have that

$$z = z - (A^*A)^{-1}\nabla_z \mathcal{L}_2(z),$$

which is equivalent to

$$(A^*A)^{-1}\nabla_z \mathcal{L}_2(z) = 0.$$

Since A is injective and $(A^*A)^{-1}$ exists, this equality holds if and only if $\nabla_z \mathcal{L}_2(z) = 0$, so that z is the fixed point of AF. □

However, Corollary 3.5.9 does not imply that given the same initial guess z^0, both algorithms will necessarily converge to the same fixed point.
By Theorem 3.5.5 and Theorem 3.5.8, both algorithms seem to be comparable in terms of convergence rate, and, by Corollary 3.5.9, in terms of critical points. However, for $T \in \mathbb{N}$ iterations of ER, $\mathcal{O}(Md^2 + TMd)$ operations are required, while AF only needs $\mathcal{O}(TMd)$ operations. Thus, in general, ER is considerably slower in terms of computation complexity. The next corollary shows that this difference is less significant in cases where the columns of A are orthogonal.

Corollary 3.5.10. *Let*

$$A^*A = \operatorname{diag}(v) \text{ for some } v \in \mathbb{R}^d,\ v_\ell > 0, \quad \ell \in [d]. \tag{3.18}$$

Then, for $T \in \mathbb{N}$ iterations both algorithms ER and AF require $\mathcal{O}(TMd)$ operations. Furthermore, if $A^*A = cI_d$ for some $c > 0$, then the iteration of ER coincides with the iteration of AF for the constant learning rate $\mu_t = \|A\|^{-2}$.

Proof. Using the condition (3.18), we obtain $(A^*A)^{-1} = \operatorname{diag}(1/v)$. Consequently, by Lemma 3.5.7, the iteration of ER is given by

$$z^{t+1} = z^t - (A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t) = z^t - \operatorname{diag}(1/v)\nabla_z \mathcal{L}_2(z^t).$$

The computation of the gradient requires $\mathcal{O}(Md)$ operations. Both the multiplication with $\operatorname{diag}(1/v)$ and the difference can be done in $\mathcal{O}(d)$ operations. Therefore, the total number of operations for a single iteration of ER is given by $\mathcal{O}(Md + d) = \mathcal{O}(Md)$, which is the same order of operations as for a single iteration of AF. We also note that $\|A\|^2 = \max_{\ell\in[d]} |v_\ell|$ and using the power method for the learning rate μ_c is not required. If $A^*A = cI$, then

$$\|A\|^2 = \|A^*A\| = \|cI_d\| = c \text{ and } (A^*A)^{-1} = c^{-1}I_d = \|A\|^{-2} I_d.$$

Hence, using Lemma 3.5.7 for the iteration of ER, we have

$$z^{t+1} = z^t - (A^*A)^{-1}\nabla_z \mathcal{L}_2(z^t) = z^t - \|A\|^{-2} \nabla_z \mathcal{L}_2(z^t),$$

which is precisely the iteration of AF with $\mu_t = \|A\|^{-2}$. □

While the condition (3.18) may seem restrictive, in many practical applications it is satisfied. For instance, the equivalence of both algorithms was observed for the recovery from Fourier magnitudes i.e., $A = F_d$ in [103]. The next corollary shows that the condition (3.18) and, consequently, the results of Corollary 3.5.10 also hold for ptychographic measurements.

Corollary 3.5.11. *Consider the ptychographic measurements (PTY) with the ptychographic measurement matrix A as in (3.9). Then, A satisfies $A^*A = \operatorname{diag}(v)$ with*

$$v_\ell = m \sum_{r\in\mathcal{R}} |(S_r w)_\ell|^2.$$

Furthermore, if $\mathcal{R} = [d]$, we have $v_\ell = m\|w\|_2^2$ for all $\ell \in [d]$. Consequently, the results of Corollary 3.5.10 apply and the computation cost of one ER iteration is given by $\mathcal{O}(|\mathcal{R}|d + m|\mathcal{R}|\log m)$.

Proof. By Lemma 3.3.5, the matrix A satisfies (3.18) with

$$v_\ell = m \sum_{r\in\mathcal{R}} |(S_r w)_\ell|^2.$$

If $\mathcal{R} = [d]$, the entries of the vector v further simplify to

$$v_\ell = m \sum_{r\in\mathcal{R}} |(S_r w)_\ell|^2 = m \sum_{r\in[d]} |w_{\ell-r}|^2, \quad \ell \in [d].$$

Changing the order of summation yields

$$v_\ell = m \sum_{j\in[d]} |w_j|^2 = m\|w\|_2^2$$

for all $\ell \in [d]$. Therefore, ER coincides with AF.
The computation of v requires $\mathcal{O}(|\mathcal{R}|d)$ operations and the gradient is computed in $\mathcal{O}(|\mathcal{R}|d+m|\mathcal{R}|\log m)$ operations. Hence, one iteration of ER requires $\mathcal{O}(|\mathcal{R}|d+m|\mathcal{R}|\log m)$ operations in total, which concludes the proof. □

Notes and References. *ER was one of the first algorithms for phase retrieval introduced in 1972 by Gerchberg and Saxton [37]. Later contributions [103], [104] and [38] classified ER as an alternating projections technique and supplemented it with the detailed analysis on the convergence and also provided an interpretation of the algorithm as a projected gradient method.*
Our contribution is the establishment of the connection between AF and ER by representing the latter as a scaled gradient method for the minimization of the amplitude-based squared loss $\mathcal{L}_2$. It allows to guarantee the convergence of ER under the mild assumption that A is injective, which was previously only available for cases $A = F_d$ [103] and A corresponding to non-circulant ptychographic measurements [26]. While the linear convergence rate of ER under additional assumptions was derived in [146], the general sublinear convergence rate, to our knowledge, has never been observed in the literature. These results were outlined in our conference paper [147] and consequent publication [148].
We also note that the connection between projection methods and gradient methods was previously established for the discrete case in [149] and in the continuous setting in [105].

3.5.3 Ptychographic Iterative Engine

The Ptychographic Iterative Engine (PIE) is an iterative algorithm designed for ptychography. It is based on the idea that only measurements for a single illumination are used in one iteration step. For an initial guess $z^0 \in \mathbb{C}^d$ the t-th iterate z^t, $t \geq 0$ of the PIE algorithm is constructed by performing the following steps.

Algorithm 2: PIE iteration, version of [39, 33]

Input : Ptychographic measurements Y as in (PTY), previous object iterate $z^t \in \mathbb{C}^d$, parameter $\alpha > 0$.
Output: $z^{t+1} \in \mathbb{C}^d$.
1. Select a shift position $r^t \in \mathcal{R}$
2. Construct an exit wave $\psi = S_{-r^t} z^t \circ w$.
3. Compute its Fourier transform $\Psi = F_m P_m \psi$.
4. Correct the magnitudes of Ψ as $\Psi' = \sqrt{Y^{(r^t)}} \circ \operatorname{sgn}_0 \Psi$.
5. Find an exit wave ψ' corresponding to Ψ' via $\psi' = P_m^* F_m^{-1} \Psi'$.
6. Return $z^{t+1} = z^t + \frac{\alpha}{\|w\|_\infty^2} S_{r^t} \operatorname{diag}(\overline{w})[\psi' - \psi]$.

Note that the division by 0 occurs whenever $|\Psi_j| = 0$ for some $j \in [m]$, which is resolved by setting $0/0$ as 0 (corresponding to sgn_0) analogously to AF and ER.
In the literature two ways of choosing the shift positions r^t are considered. Originally, in [39], the shift r^t was selected such that regions corresponding to r^{t-1} and r^t overlap, that is $|r^{t-1} - r^t|_c < \delta$. Later, in [33, 115] indices r^t are looping through the set $\mathcal{R}$, which is randomly shuffled every loop.
In terms of computational complexity, a single iteration requires $\mathcal{O}(d + m \log m)$ operations, governed by the complexity of the shift and the fast Fourier transform. Therefore, for $T \in \mathbb{N}$ iterations, $\mathcal{O}(Td + Tm \log m)$ operations are performed.
There are several interpretations of the PIE iteration. The first states that the iteration of PIE computes the measurements corresponding to the current shift position r, and adjusts the magnitudes if they do not agree with the measurements $Y^{(r)}$. Then, the algorithm

moves from the previous position z^t in the direction of the object corresponding to the corrected measurements, which gives the new iterate z^{t+1}. Therefore, α plays the role of the learning rate and controls the step size. All steps of Algorithm 2 combined into one gives the following update rule

$$z^{t+1} = z^t + \frac{\alpha S_{r^t} \operatorname{diag}(\overline{w})}{\|w\|_\infty^2} \left[P_m^* F_m^{-1} \operatorname{diag}\left(\frac{\sqrt{Y^{(r^t)}}}{|F_m P_m [S_{-r} z^t \circ w]|} \right) F_m P_m - I_d \right] (S_{-r^t} z^t \circ w). \tag{PIE}$$

In [150], the authors note that in this form the new iterate of PIE is the global minimizer of the function

$$z^{t+1} = \arg\min_{z \in \mathbb{C}^d} \|S_{-r^t} z \circ w - \psi'\|_2^2 + \left\| \left[\frac{\|w\|_\infty^2}{\alpha} I_d - \operatorname{diag}(|S_{r^t} w|^2) \right] (z - z^t) \right\|_2^2.$$

Alternatively, the iteration of the PIE algorithm is the gradient descent step [115] for the loss function

$$\|S_{-r^t} z \circ w - \psi'\|_2^2$$

with learning rate $\mu = \alpha / \|w\|_\infty^2$.

In this section, we establish a novel understanding of the PIE algorithm as stochastic gradient descent applied to the amplitude-based loss (3.13).

To this end, let us recall that the ptychographic measurements (PTY) are the phase retrieval measurements of the form (PR) with the measurement matrix A given by (3.9). By construction, the matrix A is the row-block matrix with blocks

$$A_r := F_m P_m \operatorname{diag}(w) S_{-r}, \quad r \in \mathcal{R}, \tag{3.19}$$

and, thus, the function $\mathcal{L}_2(\,\cdot\,; A)$ also splits into the sum of errors corresponding to separate blocks. More precisely, we have

$$\mathcal{L}_2(z; A) = \sum_{r \in \mathcal{R}} \mathcal{L}_2(z; A_r), \tag{3.20}$$

and we use this summation representation for the construction of the stochastic gradient $g_{\mathcal{L}_2}$ as in (2.23).

Theorem 3.5.12. *Let A be the measurements matrix (3.9) corresponding to the ptychographic measurements (PTY). Consider the amplitude-based loss function $\mathcal{L}_2 = \mathcal{L}_2(\,\cdot\,; A)$ as in (3.13) and its decomposition (3.20). If for each iteration $t \geq 1$, the shift position r^t is sampled uniformly at random from the set $\mathcal{R}$, the iteration of PIE is given by*

$$z^{t+1} = z^t - \mu_c g_{\mathcal{L}_2}(z^t), \tag{3.21}$$

where $\mu_c = \frac{\alpha}{m|\mathcal{R}|\|w\|_\infty^2}$ is the constant learning rate and $g_{\mathcal{L}_2}$ is the stochastic gradient of $\mathcal{L}_2$ given by (2.23). The sampling variables v_r in $g_{\mathcal{L}_2}$ correspond to the sampling with replacement (2.27) for $K = 1$ and probabilities $1/|\mathcal{R}|$ as in (2.31).

Proof. The construction of A_r yields

$$A_r z = F_m P_m [S_{-r} z \circ w].$$

Furthermore, by Proposition 2.2.1 and Proposition 2.2.2, we have $F_m^{-1} = \frac{1}{m} F_m^*$ and $S_r = S_{-r}^*$, so that

$$S_r \operatorname{diag}(\overline{w}) P_m^* F_m^{-1} = \frac{1}{m} S_{-r}^* (\operatorname{diag}(w))^* P_m^* F_m^* = \frac{1}{m} A_r^*.$$

Since $S_{-r} z \circ w$ is supported on $[\delta] \subseteq [m]$, by (2.11) we obtain

$$S_{-r} z \circ w = P_m^* P_m [S_{-r} z \circ w] = \frac{1}{m} P_m^* F_m^* F_m P_m [S_{-r} z \circ w] = \frac{1}{m} P_m^* F_m^* A_r z.$$

Using these equalities, we rewrite the iteration of PIE as

$$\begin{aligned}
z^{t+1} &= z^t + \frac{\alpha}{\|w\|_\infty^2} \left[\frac{1}{m} A_{r^t}^* \operatorname{diag}\left(\frac{\sqrt{Y^{(r^t)}}}{|A_{r^t} z^t|} \right) F_m P_m - S_{r^t} \operatorname{diag}(\overline{w}) \right] (S_{-r^t} z^t \circ w) \\
&= z^t + \frac{\alpha}{\|w\|_\infty^2} \left[\frac{1}{m} A_{r^t}^* \operatorname{diag}\left(\frac{\sqrt{Y^{(r^t)}}}{|A_{r^t} z^t|} \right) A_{r^t} z^t - S_{-r^t}^* (\operatorname{diag}(w))^* \frac{1}{m} P_m^* F_m^* A_{r^t} z^t \right] \\
&= z^t + \frac{\alpha}{\|w\|_\infty^2} \left[\frac{1}{m} A_{r^t}^* \operatorname{diag}\left(\frac{\sqrt{Y^{(r^t)}}}{|A_{r^t} z^t|} \right) A_{r^t} z^t - \frac{1}{m} A_{r^t}^* A_{r^t} z^t \right] \\
&= z^t - \frac{\alpha}{m \|w\|_\infty^2} A_{r^t}^* \left[I_m - \operatorname{diag}\left(\frac{\sqrt{Y^{(r^t)}}}{|A_{r^t} z^t|} \right) \right] A_{r^t} z^t \\
&= z^t - \frac{\alpha}{m \|w\|_\infty^2} \nabla_z \mathcal{L}_2(z^t; A_{r^t}),
\end{aligned} \tag{3.22}$$

where in the last line Corollary 3.5.2 was used. Let us consider a random vector v corresponding to the sampling with replacement (2.27) with $K = 1$ and probabilities $1/|\mathcal{R}|$. Then, v is given by

$$v = |\mathcal{R}| \hat{v},$$

where $\hat{v}$ is sampled uniformly at random from a set of standard basis vectors $\{e_r\}_{r \in \mathcal{R}}$, which is equivalent to sampling index $\hat{r}$ from the set $\mathcal{R}$. Therefore, the entries of v can be written as

$$v_r = |\mathcal{R}| \hat{v}_r = |\mathcal{R}| (e_{\hat{r}})_r = |\mathcal{R}| \mathcal{I}_{\hat{r}=r},$$

and the stochastic gradient of $\mathcal{L}_2$ is equal to

$$g_{\mathcal{L}_2}(z) = \sum_{r \in \mathcal{R}} v_r \nabla_z \mathcal{L}_2(z; A_r) = |\mathcal{R}| \sum_{r \in \mathcal{R}} \mathcal{I}_{\hat{r}=r} \nabla_z \mathcal{L}_2(z; A_r) = |\mathcal{R}| \nabla_z \mathcal{L}_2(z; A_{\hat{r}}). \tag{3.23}$$

Returning to (3.22), we recall that r^t is selected uniformly at random from $\mathcal{R}$ and, thus, follows the same distribution as $\hat{r}$. Hence, we obtain

$$z^{t+1} = z^t - \frac{\alpha}{m \|w\|_\infty^2} \nabla_z \mathcal{L}_2(z^t; A_{r^t}) = z^t - \frac{\alpha}{m|\mathcal{R}| \|w\|_\infty^2} g_{\mathcal{L}_2}(z^t) = z^t - \mu_c g_{\mathcal{L}_2}(z^t).$$

□

The stochastic gradient descent representation of the PIE algorithm allows us to derive the convergence of the method if the parameter α is chosen appropriately.

Theorem 3.5.13. *Let $\{z^t\}_{t\geq 0}$ be a sequence determined by PIE with an arbitrary starting point $z^0 \in \mathbb{C}^d$. Fix $\gamma > 0$. If the number of iterations T satisfies*

$$T \geq 4m^2\gamma^{-4}|\mathcal{R}|\,\|w\|_\infty^2 \left\|\sum_{r\in\mathcal{R}} |S_r w|^2\right\|_\infty \mathcal{L}_2^2(z^0),$$

and the parameter α fulfills

$$\alpha \leq \frac{\|w\|_\infty \sqrt{|\mathcal{R}|}}{\sqrt{T\left\|\sum_{r\in\mathcal{R}} |S_r w|^2\right\|_\infty}},$$

then the expected norms of the gradients satisfy

$$\min_{t\in[T]} \mathbb{E}\left\|\nabla_z \mathcal{L}_2(z^t)\right\|_2 \leq \gamma.$$

Proof. In Theorem 3.5.12 we established that PIE is the stochastic gradient descent applied to the function $\mathcal{L}_2(\,\cdot\,; A)$ with A given by (3.9). Recall that the function $\mathcal{L}_2$ is not differentiable everywhere and in order to apply Theorem 2.3.7 or its Corollary 2.3.10 providing the convergence guarantees for the stochastic gradient descent, we need to repeat the smoothing argument used in the proofs of Theorem 3.5.4 and Theorem 3.5.8. That is, consider a stochastic gradient descent applied to the smoothed amplitude-based loss $\mathcal{L}_{2,\varepsilon}(\,\cdot\,; A)$ given by (3.14) with parameter $\varepsilon > 0$. The function $\mathcal{L}_{2,\varepsilon}$ is twice continuously differentiable. The constants L and L_r, $r \in \mathcal{R}$, in (2.17) are given by

$$L = \|A^*A\|_\infty = m \max_{\ell\in[d]} \sum_{r\in\mathcal{R}} |S_r w|^2 = m\left\|\sum_{r\in\mathcal{R}} |S_r w|^2\right\|_\infty \text{ and } L_r = \|A_r^* A_r\|_\infty = m\,\|w\|_\infty^2, \tag{3.24}$$

where we used Lemma 3.5.3, the equations (3.15), and Lemma 3.3.5. If we repeat the steps of the proof of Theorem 2.3.7 with $f = \mathcal{L}_{2,\varepsilon}$ until inequality (2.26), we obtain

$$\mathcal{L}_{2,\varepsilon}(z^{t+1}) \leq \mathcal{L}_{2,\varepsilon}(z^t) - 2\mu_c \operatorname{Re}\left\langle g_{\mathcal{L}_{2,\varepsilon}}(z^t), \nabla_z \mathcal{L}_{2,\varepsilon}(z^t)\right\rangle + L\mu_c^2 \left\|g_{\mathcal{L}_{2,\varepsilon}}(z^t)\right\|_2^2.$$

As ε tends to zero from above, $\mathcal{L}_{2,\varepsilon}(z) \to \mathcal{L}_2(z)$ for all $z \in \mathbb{C}^d$. Recall that the generalized gradient $\nabla_z \mathcal{L}_2(z)$ is defined as $\lim_{\varepsilon\to 0+} \nabla_z \mathcal{L}_{2,\varepsilon}(z)$. Thus, the equality $g_{\mathcal{L}_2}(z) = \lim_{\varepsilon\to 0+} g_{\mathcal{L}_{2,\varepsilon}}(z)$ also holds. Consequently, taking the limit $\varepsilon \to 0+$ yields

$$\mathcal{L}_2(z^{t+1}) \leq \mathcal{L}_2(z^t) - 2\mu_c \operatorname{Re}\left\langle g_{\mathcal{L}_2}(z^t), \nabla_z \mathcal{L}_2(z^t)\right\rangle + L\mu_c^2 \left\|g_{\mathcal{L}_2}(z^t)\right\|_2^2.$$

Then, we can repeat the rest of the proof of Theorem 2.3.7 and guarantee that the results of Theorem 2.3.7 apply for $\mathcal{L}_2$. For the analogue of Corollary 2.3.10 we would also require the inequality

$$\|\nabla_z \mathcal{L}_2(z; A_r)\|_2^2 \leq L_r \mathcal{L}_2(z; A_r),$$

which is obtained by considering a single step of the gradient descent for $\mathcal{L}_{2,\varepsilon}(\,\cdot\,; A_r)$ with $\mu = 1/L_r$. Then, by Theorem 2.3.4 and the inequality (2.18) in particular, we get

$$\|\nabla_z \mathcal{L}_{2,\varepsilon}(z; A_r)\|_2^2 \leq L_r \mathcal{L}_{2,\varepsilon}(z; A_r),$$

and taking $\varepsilon \to 0+$ grants the desired inequality.
Therefore, in order to establish the convergence of PIE in the sense that

$$\min_{t\in[T]} \mathbb{E}\left\|\nabla_z \mathcal{L}_2(z^t)\right\|_2 \leq \gamma$$

for arbitrary $\gamma > 0$, we apply Corollary 2.3.10. For this we observe that by Theorem 3.5.12 the iteration of PIE is the stochastic gradient step which uses sampling with replacement (2.27) with $K = 1$ and probabilities $1/|\mathcal{R}|$. Thus, by the requirements of Corollary 2.3.10, the number of iterations T has to satisfy

$$T \geq \frac{4L|\mathcal{R}| \max_{r\in\mathcal{R}} L_r \mathcal{L}_2^2(z^0)}{K\gamma^4} = 4m^2\gamma^{-4}|\mathcal{R}|\,\|w\|_\infty^2 \left\|\sum_{r\in\mathcal{R}} |S_r w|^2\right\|_\infty \mathcal{L}_2^2(z^0),$$

and the learning rate μ_c is chosen such that

$$\mu_c \leq \frac{\sqrt{K}}{\sqrt{TL|\mathcal{R}| \max_{r\in\mathcal{R}} L_r}} = \frac{1}{m\,\|w\|_\infty \sqrt{T|\mathcal{R}|\left\|\sum_{r\in\mathcal{R}} |S_r w|^2\right\|_\infty}}.$$

Since by Theorem 3.5.12 the learning rate is given by $\mu_c = \frac{\alpha}{m|\mathcal{R}|\|w\|_\infty^2}$, it implies that the parameter α has to satisfy

$$\alpha \leq \frac{\|w\|_\infty \sqrt{|\mathcal{R}|}}{\sqrt{T\left\|\sum_{r\in\mathcal{R}} |S_r w|^2\right\|_\infty}},$$

which concludes the proof. □

We also consider a version of PIE, where the shift position r^t is selected non-uniformly, but based on the norms of the gradients. In this case, the probability that r^t is equal to $r \in \mathcal{R}$ is given by

$$p_r^t = \frac{\|\mathcal{L}_2(z^t; A_r)\|_2}{\sum_{r'\in\mathcal{R}} \|\mathcal{L}_2(z^t; A_{r'})\|_2}. \tag{3.25}$$

We note that at least one p_r^t is larger than zero and division by 0 does not occur if and only if $\|\mathcal{L}_2(z^t; A)\|_2 > 0$. Thus, $\sum_{r'\in\mathcal{R}} \|\mathcal{L}_2(z^t; A_{r^t})\|_2$ will always be non-zero unless the fixed point $\|\mathcal{L}_2(z^t; A)\|_2 = 0$ is reached. Furthermore, we can conclude that the following holds.

Lemma 3.5.14. *The fixed-point sets of AF, ER and PIE with sampling probabilities (3.25) coincide.*

Proof. Follows from Lemma 3.5.7 and the considerations above. □

Using norm-based probabilities requires a slight adjustment of the iteration of PIE, which leads to the following algorithm

$$z^{t+1} = z^t + \frac{\alpha S_{r^t} \operatorname{diag}(\overline{w})}{p_r^t \|w\|_\infty^2} \left[P_m^* F_m^{-1} \operatorname{diag}\left(\frac{\sqrt{Y^{(r^t)}}}{|F_m P_m [S_{-r} z^t \circ w]|} \right) F_m P_m - I_d \right] (S_{-r^t} z^t \circ w).$$

The computation of the norms requires additional operations, however, due to the locality of the measurements, this increase is not high. Note that $\mathcal{L}_2(z; A_r)$ only depends on the entries of z in the set $\mathcal{J}_r := \{r, r+1, \dots, r+\delta-1\}$. Therefore, after a single step of the modified PIE, the gradients corresponding to shifts r such that $\mathcal{J}_r \cap \mathcal{J}_{r^t} = \varnothing$ remain the same and for the rest of the indices the gradients have to be computed again. These are the shifts satisfying the condition $|r - r^t|_c < \delta$ and there are at most $2\delta - 1$ such indices. For the initialization, we have to compute all gradients once, which results in $\mathcal{O}(|\mathcal{R}|d + m|\mathcal{R}| \log m)$ operations. Then, for each iteration at most $2\delta - 1$ gradients are computed, which costs $\mathcal{O}(\delta d + m\delta \log m)$ operations. Hence, the total complexity for $T \in \mathbb{N}$ iterations is $\mathcal{O}\left((|\mathcal{R}| + T\delta)(d + m \log m)\right)$.
However, we are not able to provide convergence guarantees and step size selection for PIE with probabilities (3.25)

Notes and References. *The PIE algorithm is a popular method among practitioners. It was first introduced in [39] and later it was slightly adjusted and extended for the simultaneous estimation of the object and the window [33, 115]. In this section, we have only considered the object estimation, while the joint estimation will be discussed in Chapter 4. However, despite its popularity, no convergence analysis can be found in the literature. Thus, our main contribution is the reinterpretation of PIE as stochastic gradient descent and the derivation of its convergence guarantees summarized in Theorem 3.5.13.*
Note that the stochastic version of AF was separately studied in [151, 152] under the assumption that the entries of A are independent standard complex Gaussian random variables, but the connection to PIE was not made. Another stochastic variant of AF which randomly selects a single measurement is the randomized Kaczmarz algorithm for phase retrieval [153, 154, 155, 156, 157].
Further details on the convergence of stochastic gradient descent for phase retrieval and ptychography can be found in our recent preprint [158].

3.6 Block Phase Retrieval and its extensions

3.6.1 The idea of Block Phase Retrieval algorithm

In this section, we study the Block Phase Retrieval algorithm (BPR) [35, 40] developed specifically for ptychographic measurements (PTY) with $\mathcal{R} = [d]$. It is based on the idea of lifting, where the measurements are presented in a high-dimensional space. More precisely, define masks $w^j, j \in [m]$ as

$$w_k^j := (P_m^* M_j P_m \overline{w})_k = \overline{w}_k e^{\frac{2\pi i k j}{m}} = \begin{cases} \overline{w}_k e^{\frac{2\pi i k j}{m}}, & k \in [\delta], \\ 0, & \text{otherwise}, \end{cases} \tag{3.26}$$

where P_m is the projection operator (2.10) and M_j is the modulation operator (2.7). Then, measurements are presented as linear operation applied to the rank-one matrix xx^*,

$$I_{j,r}^m = \left| \sum_{k\in[\delta]} w_k x_{k+r} e^{-\frac{2\pi i k j}{m}} \right|^2 = \left|\langle S_{-r}x, w^j\rangle\right|^2 = \left|\langle x, S_r w^j\rangle\right|^2 = x^* S_r w^j (S_r w^j)^* x$$
$$= \operatorname{tr}(x^* S_r w^j (S_r w^j)^* x) = \operatorname{tr}(S_r w^j (S_r w^j)^* x x^*) = \langle xx^*, S_r w^j (S_r w^j)^*\rangle_F. \tag{3.27}$$

Now, define a linear operator $\mathcal{A}$ acting on the space $\mathbb{H}^d$ of $d \times d$ Hermitian matrices as

$$\mathcal{A}(Z)_{j,r} := \langle Z, S_r w^j (S_r w^j)^*\rangle_F. \tag{3.28}$$

The intensity measurements can now be written as

$$I_{j,r}^m = \mathcal{A}(xx^*)_{j,r}, \quad j \in [m], r \in \mathcal{R}. \tag{3.29}$$

In general $M = md < d^2$ and the obtained linear system (3.29) is underdetermined. Thus, a direct recovery of xx^* is not possible. Due to the condition that $\operatorname{supp}(w) = [\delta]$, the space $\operatorname{span}\{S_r w^j (S_r w^j)^*,\ j \in [m], r \in [d]\}$ is a subspace of $\mathbb{T}_\delta$ given by

$$\mathbb{T}_\delta := \{U \in \mathbb{H}^d : U_{k,\ell} = 0 \text{ for all } k, \ell \in [d] \text{ such that } |\ell - k|_c \geq \delta\} \subseteq \mathbb{H}^d.$$

Hence, $\mathcal{A}(Z)$ only depends on a part of the entries in Z.
Let us denote by T_δ the projection operator onto $\mathbb{T}_\delta$, it is given as

$$T_\delta(U)_{k,\ell} = U_{k,\ell}\mathcal{I}_{|k-\ell|_c<\delta} = \begin{cases} U_{k,\ell}, & |k-\ell|_c < \delta, \\ 0, & \text{otherwise}, \end{cases} \tag{3.30}$$

and visualized in Figure 3.1. Since $xx^* - T_\delta(xx^*)$ is orthogonal to $\mathbb{T}_\delta$ and $S_r w^j (S_r w^j)^*$ we get

$$I_{j,r}^m = \langle xx^*, S_r w^j (S_r w^j)^*\rangle_F = \langle T_\delta(xx^*), S_r w^j (S_r w^j)^*\rangle_F + \langle xx^* - T_\delta(xx^*), S_r w^j (S_r w^j)^*\rangle_F$$
$$= \langle T_\delta(xx^*), S_r w^j (S_r w^j)^*\rangle_F + 0 = \mathcal{A}(T_\delta(xx^*))_{j,r}.$$

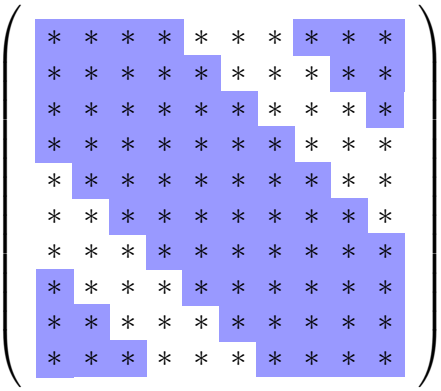


Figure 3.1: Highlighted entries of the matrix form the space $\mathbb{T}_\delta$ for $d = 10$, $\delta = 4$.

This shows that the intensity measurements only depend on the projection

$$X := T_\delta(xx^*). \tag{3.31}$$

Hence, assuming that $\operatorname{span}\{S_r w^j (S_r w^j)^*,\ j \in [m], r \in [d]\}$ coincides with $\mathbb{T}_\delta$, the matrix X is recovered from the measurements by applying the pseudoinverse of $\mathcal{A}$ restricted to $\mathbb{T}_\delta$, that is

$$X = \mathcal{A}\big|^\dagger_{\mathbb{T}_\delta}(I^m).$$

We note that for vectorized X, the operator $\mathcal{A}\big|_{\mathbb{T}_\delta}$ is a matrix and its pseudoinverse is obtained via the singular value decomposition. Since X is a Hermitian matrix in $\mathbb{T}_\delta$, it is characterized by d real values on the main diagonal and $(\delta-1)d$ complex values on the lower triangular part. Hence, in total $(2\delta-1)d$ real unknowns have to be recovered from the measurements, which implies that the condition $m \geq 2\delta - 1$ has to be satisfied.

Remark 3.6.1. *When $d \leq 2\delta - 1$, for $k, \ell \in [d]$ the distance between indices is bounded from above, as*

$$|k-\ell|_c = \min\{|k-\ell|, d - |k-\ell|\} \leq \min\{|k-\ell|, 2\delta - 1 - |k-\ell|\}.$$

If $|k-\ell| < \delta$, then $|k-\ell|_c < \delta$. Otherwise, if $|k-\ell| \geq \delta$, we have

$$2\delta - 1 - |k-\ell| \leq 2\delta - 1 - \delta = \delta - 1,$$

and again $|k-\ell|_c < \delta$. Therefore, by (3.30), the equality $T_\delta(U)_{k,\ell} = U_{k,\ell}$ holds for all indices $k, \ell \in [d]$, so that $T_\delta = I_d$ and $\mathbb{T}_\delta = \mathbb{H}^d$. It implies that $X = xx^*$, which significantly simplifies the further recovery of x. Hence, we will concentrate on the case $d > 2\delta - 1$ which means $\mathbb{T}_\delta \neq \mathbb{H}^d$.

The next step is to recover x from X by employing the rank-one properties of xx^* partially preserved in X. By the definition of $\mathbb{T}_\delta$, the main diagonal of X is always recovered from the measurements and it is given by

$$X_{k,k} = |x_k|^2.$$

Therefore, the magnitudes of x are reconstructed from the main diagonal.
The phases of x are obtained by computing the top eigenvector $\tilde{x}$ of the phase difference matrix

$$\operatorname{sgn}_0(X_{k,\ell}) = \begin{cases} \operatorname{sgn} x_k \overline{\operatorname{sgn} x}_\ell, & X_{k,\ell} \neq 0, \\ 0, & \text{otherwise}, \end{cases}$$

and setting $\operatorname{sgn} x = \operatorname{sgn} \tilde{x}$. It is linked to the angular synchronization problem with more detailed explanation to follow in Section 3.6.4. For now we only provide an intuition of this step. Let x be non-vanishing, so that $\sum_{\ell=0}^{d-1} \mathcal{I}_{X_{k,\ell}\neq 0} = 2\delta - 1$ for all $k \in [d]$. For the true vector of phases $\operatorname{sgn} x$ we have

$$\operatorname{sgn} x^* \operatorname{sgn}_0(X) \operatorname{sgn} x = \sum_{k,\ell=0}^{d-1} \overline{\operatorname{sgn} x}_k \operatorname{sgn} x_k \overline{\operatorname{sgn} x}_\ell \operatorname{sgn} x_\ell \mathcal{I}_{X_{k,\ell}\neq 0} = \sum_{k,\ell=0}^{d-1} \mathcal{I}_{X_{k,\ell}\neq 0} = (2\delta-1)d,$$

and for any other vector v with $\|v\|_2 = \sqrt{d}$ we obtain

$$|v^* \operatorname{sgn}_0(X) v| \le \sum_{k,\ell=0}^{d-1} |v_k||v_j| \mathcal{I}_{X_{k,\ell}\neq 0} \le \sum_{k,\ell=0}^{d-1} \frac{|v_k|^2 + |v_j|^2}{2} \mathcal{I}_{X_{k,\ell}\neq 0} = \sum_{k,\ell=0}^{d-1} |v_k|^2 \mathcal{I}_{X_{k,\ell}\neq 0}$$
$$= \sum_{k=0}^{d-1} |v_k|^2 \sum_{\ell=0}^{d-1} \mathcal{I}_{X_{k,\ell}\neq 0} = (2\delta - 1) \sum_{k=0}^{d-1} |v_k|^2 = (2\delta - 1) \|v\|_2^2 = (2\delta - 1) d,$$

where we used that X is Hermitian. Thus, $\operatorname{sgn} x$ is the top eigenvector of matrix $\operatorname{sgn}_0(X)$. Combining all steps together, BPR can be summarized in Algorithm 3.

Algorithm 3: Block Phase Retrieval [40]

Input : Ptychographic measurements $Y \in \mathbb{R}^{m\times d}$ as in (PTY) with $\mathcal{R} = [d]$.
Output: $z \in \mathbb{C}^d$ with $z \approx e^{-i\theta} x$ for some $\theta \in [0, 2\pi)$.
1. Compute $Z = \mathcal{A}|^{\dagger}_{\mathbb{T}_\delta}(Y) \in \mathbb{T}_\delta$ as an Hermitian estimate of X.
2. Form the matrix of phase differences $\operatorname{sgn}_0(Z) \in \mathbb{T}_\delta$.
3. Compute the top eigenvector of $\operatorname{sgn}_0(Z)$, denoted by $\tilde{z} \in \mathbb{C}^d$ with $\|\tilde{z}\|_2 = \sqrt{d}$.
4. Set $z_j = \sqrt{Z_{j,j}} \cdot \operatorname{sgn} \tilde{z}_j$ for all $j \in [d]$ to form $z \in \mathbb{C}^d$.

Compared to AF, ER and PIE discussed in the previous sections, Algorithm 3 has three main advantages and two disadvantages.
Firstly, it is non-iterative and, therefore, extremely fast [40]. Thanks to a special structure of $\mathcal{A}$, the action of the pseudoinverse in Step 1 can be efficiently computed and requires $\mathcal{O}(\delta d \log d)$ operations. We provide details on the computational complexity of Step 1 at the end of Section 3.6.2.1. The search of the top eigenvector uses the sparsity of matrix $\operatorname{sgn}_0(Z)$ which grants a total complexity of $\mathcal{O}(\delta^2 d \log d)$.
Secondly, Algorithm 3 possess recovery guarantees as stated in the next theorem.

Theorem 3.6.2. *[138, 40, Theorem 1] Consider the noisy ptychographic measurements (PTY) with all shifts observed $\mathcal{R} = [d]$. Let $\delta > 2$, $d \ge 4\delta$, $m = 2\delta - 1$ and the δd-th singular value of the operator $\mathcal{A}$ denoted by $\sigma_{\delta d}(\mathcal{A})$ be non-zero. If $x \in \mathbb{C}^d$ is non-vanishing with $|x|_{\min} := \min_{k\in[d]} |x_k|$, then the estimate $z \in \mathbb{C}^d$ produced by Algorithm 3 satisfies*

$$\operatorname{dist}(x, z) \le 24 \frac{\|x\|_\infty}{|x|^2_{\min}} \cdot \frac{d^2}{\delta^{5/2}} \cdot \sigma^{-1}_{\delta d}(\mathcal{A}) \|N\|_F + d^{1/4} \sqrt{\sigma^{-1}_{\delta d}(\mathcal{A}) \|N\|_F}.$$

In particular, by Theorem 3.6.2, in the noiseless scenario Algorithm 3 uniquely identifies non-vanishing vectors x from the ptychographic measurements. To our knowledge, there are no recovery guarantees available for any other method for the ptychographic measurements (PTY) in the literature.
Finally, BPR consists of three steps: the inversion, the magnitude and the phase estimation. The structure of Algorithm 3 is modular and each step only depends on the outcome of the previous steps. Thus, we can easily change procedures used for single steps of Algorithm 3. Throughout the following sections, new methods are introduced with separate recovery guarantees for each step, which are of the form

$$\|Z - X\|_F \le f_1(\|N\|_F),$$
$$\||z| - |x|\|_2 \le f_2(\|Z - X\|_F),$$
$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn} z) \le f_3(\|Z - X\|_F),$$

where f_1, f_2 and f_3 are some non-decreasing functions from $\mathbb{R}$ to $\mathbb{R}$ possibly depending on x. The recovery guarantees for Algorithm 3 are then obtained by combining these bounds as

$$\begin{aligned}\operatorname{dist}(x,z) &= \min_{|\alpha|=1} \|x-\alpha z\|_2 = \min_{|\alpha|=1} \| |x|\circ\operatorname{sgn} x \pm \alpha|x|\circ\operatorname{sgn} z - \alpha|z|\circ\operatorname{sgn} z\|_2 \\ &\le \min_{|\alpha|=|} \| |x|\circ\operatorname{sgn} x - \alpha|x|\circ\operatorname{sgn} z\| + \||x|-|z|\|_2 \\ &\le \|x\|_\infty \operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn} z) + \||x|-|z|\|_2 \le \|x\|_\infty f_3(f_1(\|N\|_F)) + f_2(f_1(\|N\|_F)).\end{aligned} \tag{3.32}$$

Returning to Theorem 3.6.2, we observe that it also highlights the weaknesses of BPR. The singular value $\sigma_{\delta d}(\mathcal{A})$ appearing in the bound might be small, which leads to sensitivity of the algorithm to noise. Depending on the choice of the window w, the singular value $\sigma_{\delta d}(\mathcal{A})$ may be equal to 0, which means that the assumption

$$\operatorname{span}\{S_r w^j (S_r w^j)^*,\ j\in[m], r\in[d]\} = \mathbb{T}_\delta$$

does no longer hold and the reconstruction process therefore fails.
Another problem of Algorithm 3 is the requirement $\mathcal{R} = [d]$. In practice, it would be preferable to work with a subsampled set of shifts. A compromise between theoretical results on one hand and application demands on the other hand can be reached by extensions of BPR [138, 139, 141, 159] to equidistant shifts $\mathcal{R} = \{0, s, \dots, d-s\}$ with a shift length $s < \delta$ and divisor of d. Further relaxations to arbitrary sets $\mathcal{R}$ might be possible, however the computational complexity and recovery guarantees are expected to degrade due to the loss of the structure of $\mathcal{R}$.
Furthermore, BPR is based on the circularity of shifts. For aperiodic objects, the circularity can be introduced by padding an object with a proper amount of dummy entries and artificial measurements. As a result, the window never overlaps with two disjoint ends of the padded object simultaneously.
The rest of the section is structured in the following way. In Section 3.6.2.1, the inversion step is discussed in detail and in subsequent Section 3.6.2.2 instabilities arising from a certain class of windows are treated by regularization of the BPR algorithm. Section 3.6.3 provides an overview of more advanced magnitude estimation techniques in the literature and in Section 3.6.4 the phase reconstruction step and its alternations are explained. The extension of Algorithm 3 to equidistant shifts is the topic of Section 3.6.5. Finally, we briefly discuss the implications of the recovery guarantees for BPR in the context of uniqueness and stability of the reconstruction in Section 3.6.6.

Notes and References. *The initial version of BPR was introduced in [35] with a greedy phase reconstruction procedure. The subsequent work [40] replaced it with the eigenvalue-based estimation and brought Algorithm 3 to its presented above form. Interestingly, a similar algorithm was independently derived in [96]. Later work [160] extended BPR for a two-dimensional setting for windows formed by a rank-one matrix. In [141], the authors introduced a reinterpretation of the inversion step via Wigner distribution deconvolution, which allowed to extend BPR for the two-dimensional reconstruction for an arbitrary window [142]. This connection also provides an understanding of BPR as a discrete version of the Wigner distribution deconvolution algorithm [79, 5, 161]. Yet, other discretizations are possible [162, 163].*

The various alternations of steps in Algorithm 3 are present in the literature [138, 140, 159, 164] and we will discuss each contribution in the following sections. Similarly, the extensions of BPR to equidistant shifts [138, 139, 141, 159] will be covered in the corresponding section.
We note that BPR can be linked to the analysis of uniqueness of reconstruction from the ptychographic measurements (PTY) *provided in [85, 70].*
Finally, the lifting trick (3.27) *is a foundation for another algorithm known as PhaseLift [34, 122, 165, 124, 166, 92, 125, 126]. While BPR uses lifting to obtain the matrix X and only then benefits from the inherited rank-one structure, PhaseLift recovers xx^* by solving a convex relaxation of the rank minimization problem.*

3.6.2 Inversion step

In this section, we discuss the inversion step in detail. In the first half of the section we show that the inversion step coincides with Wigner Distribution Deconvolution and in the second half we address singularities and instabilities arising for certain classes of windows.

3.6.2.1 Inversion step as Wigner Distribution Deconvolution

In the previous section, the inversion step is viewed as an application of the pseudoinverse matrix to the measurement vector in order to obtain the vectorized matrix X defined in (3.31). It is a valid approach established in [35], but it requires careful manipulations with vectorization, construction of matrices and their further decomposition. However, the resulting formula for the pseudoinverse operator suggests that there is a connection between the inversion and the time-frequency analysis.
The first major step towards the understanding of this connection was the discrete analogue of the relation between intensity measurements and the Wigner distribution (3.5), which is provided by the next theorem.

Theorem 3.6.3 (Wigner distribution deconvolution, version of [141, Theorem 4])**.** *Consider the noiseless ptychographic measurements* (PTY) *with all shifts, i.e., $\mathcal{R} = [d]$, and $m \geq 2\delta - 1$. For $j \in [m]$ define the transform*

$$\rho(j) := \begin{cases} j, & j \leq \lfloor m/2 \rfloor, \\ j - m, & j > \lfloor m/2 \rfloor. \end{cases}$$

Then, for all $j \in [m]$ the j-th row of the matrix $F_m^{-1} I^m F_d^T$ is given by

$$(F_m^{-1} I^m F_d)_{(j)} = F_d[x \circ S_{\rho(j)}\overline{x}] \circ \overline{F_d[\overline{w} \circ S_{\rho(j)} w]}.$$

Furthermore, for $\delta \leq j \leq m - \delta$ the coefficients $F_d[\overline{w} \circ S_{\rho(j)} w]_k$ are zero.

Proof. It follows from the more general Theorem 3.6.4 applied to X. See also our discussion below for further clarifications. □

In order to make a connection between Theorem 3.6.3 and the inversion step, let us first consider a matrix U in $\mathbb{T}_\delta$. It has $2\delta - 1$ non-zero diagonals

$$d^{\rho(j)}(U)_k = U_{k,k-\rho(j)} \quad \text{for } k \in [d], j \in [m], |\rho(j)| < \delta,$$

and due to the inclusion $\mathbb{T}_\delta \subseteq \mathbb{H}^d$ the matrix U is completely identified by its lower-triangular diagonals $d^j(U)$, $j \in [\delta]$. Specifically for X, the diagonals $d^{\rho(j)}(X)$ satisfy

$$d^{\rho(j)}(X)_k = X_{k,k-\rho(j)} = (xx^*)_{k,k-\rho(j)} = x_k \overline{x}_{k-\rho(j)} = (x \circ S_{\rho(j)}\overline{x})_k, \tag{3.33}$$

and, thus, Theorem 3.6.3 determines a relation between the Fourier coefficients of the diagonals $d^{\rho(j)}(X)$ and the measurements,

$$(F_m^{-1}\mathcal{A}(X)F_d)_{(j)} = (F_m^{-1}I^m F_d)_{(j)} = F_d[d^{\rho(j)}(X)] \circ \overline{F_d[\overline{w} \circ S_{\rho(j)}w]}, \quad j \in [m],$$

This is a linear system with respect to diagonals. Furthermore, the result of Theorem 3.6.3 can be extended for any matrix $U \in \mathbb{T}_\delta$.

Theorem 3.6.4. *Consider the setup of Theorem 3.6.3. Let $\mathcal{A}$ be the measurement operator defined in (3.28). Then, for all $U \in \mathbb{T}_\delta$ the equality*

$$(F_m^{-1}\mathcal{A}(U)F_d)_{(j)} = F_d[d^{\rho(j)}(U)] \circ \overline{F_d[\overline{w} \circ S_{\rho(j)}w]}, \quad j \in [m].$$

holds. Furthermore, for $\delta \le j \le m - \delta$ the coefficients $F_d[\overline{w} \circ S_{\rho(j)}w]_k$ are zero.

Proof. By the definition of the measurements operator (3.28), we have

$$\mathcal{A}(U)_{\ell,r} = \langle U, S_r w^\ell (w^\ell)^* S_r^* \rangle_F.$$

An application of the inverse Fourier transform gives

$$\begin{aligned} F_m^{-1}\mathcal{A}(U)_{j,r} &= \frac{1}{m}\sum_{\ell\in[m]} \langle U, S_r w^\ell (w^\ell)^* S_r^* \rangle_F e^{\frac{2\pi i j\ell}{m}} \\ &= \left\langle U, S_r \left[\frac{1}{m}\sum_{\ell\in[m]} w^\ell (w^\ell)^* e^{-\frac{2\pi i j\ell}{m}}\right] S_r^* \right\rangle_F. \end{aligned} \tag{3.34}$$

Using the circularity of the complex exponential, we replace j with $\rho(j)$,

$$e^{-\frac{2\pi i j\ell}{m}} = \left\{\begin{array}{ll} e^{-\frac{2\pi i j\ell}{m}}, & j \le \lfloor m/2 \rfloor \\ e^{-\frac{2\pi i (j-m)\ell}{m}}, & j > \lfloor m/2 \rfloor \end{array}\right\} = e^{-\frac{2\pi i \rho(j)\ell}{m}}.$$

The definition (3.26) of vectors w^ℓ for $k, s \in [d]$ yields

$$\begin{aligned} \left[\frac{1}{m}\sum_{\ell\in[m]} w^\ell (w^\ell)^* e^{-\frac{2\pi i \rho(j)\ell}{m}}\right]_{k,s} &= \overline{w}_k w_s \frac{1}{m}\sum_{\ell\in[m]} e^{\frac{2\pi i \ell k}{m}} e^{-\frac{2\pi i s\ell}{m}} e^{-\frac{2\pi i \rho(j)\ell}{m}} \\ &= \overline{w}_k w_s \frac{1}{m}\sum_{\ell\in[m]} e^{\frac{2\pi i \ell(k-s-\rho(j))}{m}} = \overline{w}_k w_s \mathcal{I}_{0=k-s-\rho(j) \bmod m}. \end{aligned} \tag{3.35}$$

If either $k \ge \delta$ or $s \ge \delta$, the factor $\overline{w}_k w_s = 0$ and we can write

$$\overline{w}_k w_s \mathcal{I}_{0=k-s-\rho(j) \bmod m} = 0 = \overline{w}_k w_s \mathcal{I}_{0=k-s-\rho(j)}$$

Let $k, s \in [\delta]$. By the definition of $\rho(j)$, the inequality $|\rho(j)| \leq \lfloor m/2 \rfloor$ holds. Thus, using the assumption $m \geq 2\delta - 1$, we have $|k - s - \rho(j)| < \delta + \lfloor m/2 \rfloor \leq m$. It implies that the modulo operation can be discarded, so that for all $k, s \in [d]$

$$\overline{w}_k w_s \mathcal{I}_{0=k-s-\rho(j) \bmod m} = \overline{w}_k w_s \mathcal{I}_{0=k-s-\rho(j)}. \tag{3.36}$$

Substituting (3.35) and (3.36) in (3.34) gives

$$\begin{aligned} F_m^{-1}\mathcal{A}(U)_{j,r} &= \sum_{k,s\in[d]} U_{k,s} \left[\overline{\frac{1}{m} \sum_{\ell\in[m]} w^\ell (w^\ell)^* e^{-\frac{2\pi i \rho(j) \ell}{m}}} \right]_{k-r,s-r} \\ &= \sum_{k,s\in[d]} U_{k,s} \overline{\overline{w}_{k-r} w_{s-r} \mathcal{I}_{0=k-r-s+r-\rho(j)}} = \sum_{k\in[d]} U_{k,k-\rho(j)} w_{k-r} \overline{w}_{k-\rho(j)-r} \\ &= \sum_{k\in[d]} d^{\rho(j)}(U)_k (w \circ S_{\rho(j)} \overline{w})_{k-r} = \sum_{k\in[d]} d^{\rho(j)}(U)_k (R_d[w \circ S_{\rho(j)} \overline{w}])_{r-k} \\ &= \left(d^{\rho(j)}(U) *_d (R_d[w \circ S_{\rho(j)} \overline{w}]) \right)_r, \end{aligned} \tag{3.37}$$

where R_d denotes the time reversal operator (2.8) and $*_d$ is the circular convolution (2.9). By the circular convolution theorem (Theorem 2.2.3), the application of the Fourier transform with respect to the variable r leads to

$$(F_m^{-1}\mathcal{A}(U)F_d)_{(j)} = (F_m^{-1}\mathcal{A}(U)F_d^T)_{(j)} = (F_m^{-1}\mathcal{A}(U))_{(j)} F_d^T = F_d[d^{\rho(j)}(U)] \circ F_d R_d[w \circ S_{\rho(j)} \overline{w}].$$

By Proposition 2.2.5, the second term transforms into

$$F_d R_d[w \circ S_{\rho(j)} \overline{w}]] = R_d F_d \overline{[w \circ S_{\rho(j)} \overline{w}]} = \overline{F_d [w \circ S_{\rho(j)} \overline{w}]} = \overline{F_d [\overline{w} \circ S_{\rho(j)} w]}.$$

Finally, we note that for $\delta \leq j \leq m - \delta$ the supports of w and $S_{\rho(j)} w$ do not overlap and, consequently, $F_d[\overline{w} \circ S_{\rho(j)} w] = 0$. □

Remark 3.6.5. *From the equation (3.37) it can be observed that m has no impact on the number of equations with a right-hand side different from zero. The right-hand side of (3.37) is non-zero if $j < \delta$ or $j > m - \delta$ for all choices of m, which is always $2\delta - 1$ values of j. Moreover, for each $r \in \mathcal{R}$ the left-hand side $\mathcal{A}(U)_{(r)}$ is a real vector in $\mathbb{R}^m$ and, hence, $F_m^{-1}\mathcal{A}(U)_{(r)}$ can also be described by $2\delta - 1$ real values. Thus, the dimension of $\operatorname{span}\{S_r w^\ell (w^\ell)^* S_r^*, j \in [m], r \in \mathcal{R}\}$ over the field $\mathbb{R}$ is at most $(2\delta - 1)|\mathcal{R}|$ and in the absence of noise the choice of m has no significance for the dimension. However, in the presence of noise, the entries of $F_m^{-1}(\mathcal{A}(U) + N)_{j,r}$, $\delta \leq j \leq m - \delta$ will only contain noise. This allows to discard a part of the noise in the reconstruction process. Hence, $m/(2\delta - 1)$ has a meaning of an oversampling ratio and its higher values provide better noise robustness of the inversion step.*

Firstly, Theorem 3.6.4 highlights that recovery of the diagonals outside $\mathbb{T}_\delta$ is not possible and we can restrict $\mathcal{A}$ to the subspace $\mathbb{T}_\delta$.
Secondly, since the operator $\mathcal{A}$ is real-valued for all $U \in \mathbb{T}_\delta \subseteq \mathbb{H}^d$,

$$\mathcal{A}(U)_{j,r} = \langle U, S_r w^\ell (S_r w^\ell)^* \rangle_F = (S_r w^\ell)^* U S_r w^\ell \in \mathbb{R},$$

the result of Theorem 3.6.4 is redundant for values of $j > \lfloor m/2 \rfloor$. More precisely, by the symmetry of the discrete Fourier transform for real-valued vectors, we have

$$(F_m^{-1}\mathcal{A}(U)F_d)_{(j)} = (F_m^{-1}\mathcal{A}(U))_{(j)}F_d = \overline{(F_m^{-1}\mathcal{A}(U))_{(m-j)}}F_d = \overline{R_d(F_m^{-1}\mathcal{A}(U)F_d)_{(m-j)}}.$$

This also can be viewed as a consequence of U being a Hermitian matrix with diagonals satisfying

$$d^{-j}(U)_k = U_{k,k+j} = \overline{U}_{k+j,k} = \overline{U}_{k+j,k-j+j} = \overline{d^j(U)}_{k+j}.$$

Thirdly, by rescaling the Fourier matrices,

$$\left[\sqrt{m}F_m^{-1}\mathcal{A}(U)\frac{1}{\sqrt{d}}F_d\right]_{(j)} = \frac{1}{\sqrt{d}}F_d[d^j(U)] \circ \sqrt{m}\overline{F_d[\overline{w} \circ S_j w]}, \quad j \in [\delta],$$

we obtain that all transformations are unitary except for multipliers $\sqrt{m}\overline{F_d[\overline{w} \circ S_j w]}_k$. Hence, the set $\{\sqrt{m}|F_d[\overline{w} \circ S_j w]_k|\}_{j\in[\delta],k\in[d]}$ contains the singular values of the operator $\mathcal{A}\big|_{\mathbb{T}_\delta}$ and the linear system is not underdetermined whenever

$$\sigma_{\delta d}(\mathcal{A}\big|_{\mathbb{T}_\delta}) = \min_{j\in[\delta],k\in[d]} \sqrt{m}|F_d[\overline{w} \circ S_j w]_k| > 0. \tag{3.38}$$

This is equivalent to the assumption

$$\mathbb{T}_\delta = \operatorname{span}\{S_r w^j (S_r w^j)^*,\ j \in [m], r \in [d]\}.$$

Finally, for a matrix $V \in \mathbb{R}^{m\times d}$ the result of $U = \mathcal{A}\big|_{\mathbb{T}_\delta}^{\dagger}(V)$ is obtained by recovering the diagonals via

$$d^j(U) = F_d^{-1}\left[\frac{(F_m^{-1}VF_d)_{(j)}}{\overline{F_d[\overline{w} \circ S_j w]}}\right], \quad j \in [\delta],$$

and the construction of U from its diagonals. It justifies that the pseudoinverse operator $\mathcal{A}\big|_{\mathbb{T}_\delta}^{\dagger}$ acts as the discrete Wigner distribution deconvolution applied to the given data V. In particular, for noisy measurements $Y = \mathcal{A}(X) + N$, the matrix Z obtained in Step 1 of Algorithm 3 satisfies

$$d^j(Z) = F_d^{-1}\left[\frac{(F_m^{-1}YF_d)_{(j)}}{\overline{F_d[\overline{w} \circ S_j w]}}\right] = d^j(X) + F_d^{-1}\left[\frac{(F_m^{-1}NF_d)_{(j)}}{\overline{F_d[\overline{w} \circ S_j w]}}\right], \quad j \in [\delta]. \tag{3.39}$$

A reconstruction error of the inversion step is quantified by the next corollary.

Corollary 3.6.6. *Consider the noisy ptychographic measurements (PTY) with all shifts, i.e., $\mathcal{R} = [d]$, and $m \geq 2\delta - 1$. Assume that the inequality (3.38) holds. Let X be defined as in (3.31) and Z be the matrix obtained in the Step 1 of Algorithm 3 by the reconstruction of its diagonals via (3.39). Then*

$$\|Z - X\|_F \leq \sigma_{\delta d}^{-1}(\mathcal{A}\big|_{\mathbb{T}_\delta})\,\|N\|_F\,.$$

Proof. Since both X and Z are in $\mathbb{T}_\delta$, they are Hermitian. Then, we compute the Frobenius norm by summing up the entries of $Z - X$ along the diagonals,

$$\begin{aligned}\|Z-X\|_F^2 &= \left\|d^0(Z)-d^0(X)\right\|_2^2 + \sum_{j=1}^{\delta} \left\|d^j(Z)-d^j(X)\right\|_2^2 + \left\|d^{-j}(Z)-d^{-j}(X)\right\|_2^2 \\ &= \left\|d^0(Z)-d^0(X)\right\|_2^2 + \sum_{j=1}^{\delta} 2\left\|d^j(Z)-d^j(X)\right\|_2^2 . \end{aligned} \tag{3.40}$$

By the equation (3.39) we have

$$\begin{aligned}\left\|d^j(Z)-d^j(X)\right\|_2 &= \left\| F_d^{-1}\left[\frac{(F_m^{-1}NF_d)_{(j)}}{\overline{F_d[\overline{w}\circ S_{\rho(j)}w]}}\right]\right\|_2 \\ &= \left\| \sqrt{d}F_d^{-1}\,\mathrm{diag}\left(\frac{1}{\sqrt{m}\overline{F_d[\overline{w}\circ S_{\rho(j)}w]}}\right)\frac{1}{\sqrt{d}}F_d(\sqrt{m}F_m^{-1}N)_{(j)}^T\right\|_2 \\ &\le \sigma_{\delta d}^{-1}(\mathcal{A}|_{\mathbb{T}_\delta})\left\|(\sqrt{m}F_m^{-1}N)_{(j)}\right\|_2 .\end{aligned}$$

Then, the symmetry of the discrete Fourier transform for real vectors yields

$$\begin{aligned}\|Z-X\|_F^2 &\le \sigma_{\delta d}^{-2}(\mathcal{A}|_{\mathbb{T}_\delta})\left[\left\|(\sqrt{m}F_m^{-1}N)_{(0)}\right\|_2^2 + \sum_{j=1}^{\delta} 2\left\|(\sqrt{m}F_m^{-1}N)_{(j)}\right\|_2^2\right] \\ &= \sigma_{\delta d}^{-2}(\mathcal{A}|_{\mathbb{T}_\delta})\left[\left\|(\sqrt{m}F_m^{-1}N)_{(0)}\right\|_2^2 + \sum_{j=1}^{\delta}\left\|(\sqrt{m}F_m^{-1}N)_{(j)}\right\|_2^2 + \left\|\overline{(\sqrt{m}F_m^{-1}N)_{(m-j)}}\right\|_2^2\right] \\ &\le \sigma_{\delta d}^{-2}(\mathcal{A}|_{\mathbb{T}_\delta})\left\|\sqrt{m}F_m^{-1}N\right\|_F^2 = \sigma_{\delta d}^{-2}(\mathcal{A}|_{\mathbb{T}_\delta})\|N\|_F^2 .\end{aligned} \tag{3.41}$$

□

Notes and References. *The matrix form of the pseudoinverse operator for vectorized matrices $U \in \mathbb{T}_\delta$ was established in the one-dimensional case in [35, 140] for $m = 2\delta - 1$. Its construction is heavily based on the computations of matrix multiplications. The results of [35] were extended to the two-dimensional case [160] for special windows given by rank-one matrices. This construction exploits the tensor decomposition and leads to a reduction to the one-dimensional case.*

In [141], an analogue of Theorem 3.6.3 is established for the case $m \ge 2\delta - 1$ and m being a divisor of d, which connects the inversion step to the discrete version of the Wigner distribution deconvolution. Based on the relation between the diagonals of X and the intensity measurements given in Theorem 3.6.3, the authors proposed recovery by diagonals (3.39). However, the fact that it coincides with an application of the pseudoinverse $\mathcal{A}|_{\mathbb{T}_\delta}^{\dagger}$ was only observed in numerical trials. One of the strong points of the results in [141] is its matrix multiplication-free proof. This was later generalized to the two-dimensional scenario in [142].

The first step towards the connection between two methods was made in Theorem 3 of our earlier work [140]. The main contribution of this section is Theorem 3.6.4, which provides a complete theoretical justification of the fact that the pseudoinverse- and the Wigner

distribution deconvolution-based approaches coincide. Just as the proof of Theorem 3.6.3 in [141], the derivation of Theorem 3.6.4 does not require matrix multiplication. Thus, it can be easily generalized to the two-dimensional scenario.
A benefit of our measurement setup is that the subsampling of intensities is performed during the transition from the continuous to the discrete model, which allows to avoid the additional assumption in [141] that m is a divisor of d.
We note that the recovery by diagonals in (3.39) consists only of the discrete Fourier transforms and entrywise operations, which grants the computation complexity $\mathcal{O}(d \log d)$ per diagonal and total $\mathcal{O}(\delta d \log d)$ for the inversion step.

3.6.2.2 Instabilities of inversion step and subspace completion

The crucial part of the inversion step is the assumption that the operator $\mathcal{A}\big|_{\mathbb{T}_\delta}$ is injective on $\mathbb{T}_\delta$, so that inequality (3.38) holds. In the original work on BPR [35], the authors used the following window.

Proposition 3.6.7 ([35])**.** *Let $m = 2\delta - 1$ and consider an exponential window of the form*

$$w_k = \begin{cases} \frac{1}{(2\delta-1)^{1/4}} e^{-\frac{\mathrm{i}}{\alpha}}, & k \in [\delta], \\ 0, & k \notin [\delta], \end{cases}$$

with $\alpha = \max\{4, (\delta-1)/2\}$. Then, the minimal singular value satisfies

$$\sigma_{\delta d}(\mathcal{A}\big|_{\mathbb{T}_\delta}) > \frac{7}{20\alpha} e^{-\frac{\delta-1}{\alpha}} > c\delta^{-1},$$

for a constant $c > 0$.

Later, similar bounds on the smallest singular value were established for near-flat windows [159], i.e.,

$$w_k = \begin{cases} \alpha + 1, & k = 0, \\ 1, & k \in [\delta]\backslash\{0\}, \\ 0, & k \notin [\delta]. \end{cases}$$

A more general description of windows, which lead to an invertible operator $\mathcal{A}\big|_{\mathbb{T}_\delta}$ was given in [141].

Proposition 3.6.8 ([141, Proposition A1])**.** *Consider a window $w \in \mathbb{C}^d$ with $\operatorname{supp}(w) = [\delta]$. If*

$$|w_0| > (\delta - 1)|w_1| \text{ and } |w_k| \geq |w_{k+1}|, \quad k \in [\delta - 1],$$

then $\sigma_{\delta d}(\mathcal{A}\big|_{\mathbb{T}_\delta}) > 0$.

However, there are also examples for which $\mathcal{A}\big|_{\mathbb{T}_\delta}$ is not invertible.

Example 3.6.9. *Let d be even. Consider a window w which satisfies the following symmetry condition*

$$w_k = \overline{w}_{\delta-k-1}, \quad k \in [\delta].$$

Since $\operatorname{supp}(w) = [\delta]$, *we obtain*

$$\begin{aligned} F_d[\overline{w} \circ S_j w]_{d/2} &= \sum_{k\in[d]} \overline{w}_k w_{k-j} e^{-\frac{2\pi i k d}{2d}} = \sum_{k=j}^{\delta-1} \overline{w}_k w_{k-j}(-1)^k \\ &= \sum_{k=j}^{\delta-1} \overline{w}_{\delta+j-1-k} w_{\delta+j-1-k-j}(-1)^{\delta+j-1+k} = \sum_{k=j}^{\delta-1} w_{k-j}\overline{w}_k(-1)^{\delta+j-1+k}, \end{aligned}$$

where in the second line we changed the summation order and applied the symmetry of w. *Using a combination of representations above provides*

$$\begin{aligned} F_d[\overline{w} \circ S_j w]_{d/2} &= \frac{1}{2}\sum_{k=j}^{\delta-1} \overline{w}_k w_{k-j}(-1)^k + \frac{1}{2}\sum_{k=j}^{\delta-1} w_{k-j}\overline{w}_k(-1)^{\delta+j-1+k} \\ &= \frac{1}{2}\sum_{k=j}^{\delta-1} \overline{w}_k w_{k-j}(-1)^k(1+(-1)^{\delta+j-1}). \end{aligned}$$

If $\delta + j$ *is even, then* $(-1)^{\delta+j-1} = -1$ *and, hence, all summands are zero. Thus, for* $j \in [\delta]$ *such that* $\delta + j$ *is even, we get*

$$F_d[\overline{w} \circ S_j w]_{d/2} = 0.$$

In particular, Example 3.6.9 implies that for the discrete Gaussian window corresponding to a discrete analogue of the Gabor transform

$$w_k = \begin{cases} e^{-\frac{(j-(\delta-1)/2)^2}{a}}, & k \in [\delta], \\ 0, & k \notin [\delta], \end{cases} \tag{3.42}$$

the inversion step of BPR fails.

Even if $\mathcal{A}\big|_{\mathbb{T}_\delta}$ has no zero singular values, the reconstruction of the diagonals via (3.39) is sensitive to small values $|F_d[\overline{w} \circ S_j w]|$ as noise is being amplified by $|F_d[\overline{w} \circ S_j w]|^{-1}$. Therefore, a regularization procedure is necessary. This can be achieved by ignoring the reconstructed coefficients corresponding to the singular values of $\mathcal{A}\big|_{\mathbb{T}_\delta}$ below a threshold $\varepsilon \geq 0$.

More precisely, let us consider the set

$$\mathcal{S}_\varepsilon := \{(j,k) : j \in [\delta], k \in [d], |F_d[\overline{w} \circ S_j w]_k| \leq \varepsilon\},$$

and denote the complement of $\mathcal{S}_\varepsilon$ by $\mathcal{S}_\varepsilon^c := ([\delta] \times [d]) \backslash \mathcal{S}_\varepsilon$. Then, the truncation procedure replaces the Fourier coefficients corresponding to $\mathcal{S}_\varepsilon$ with zero, i.e.,

$$F_d[d^j(Z)]_k = \frac{(F_m^{-1} Y F_d)_{j,k}}{F_d[\overline{w} \circ S_j w]_k}, \cdot \mathcal{I}_{(j,k)\in\mathcal{S}_\varepsilon^c} = \begin{cases} \frac{(F_m^{-1} Y F_d)_{j,k}}{\overline{F_d[\overline{w} \circ S_j w]_k}}, & (j,k) \in \mathcal{S}_\varepsilon^c, \\ 0, & (j,k) \in \mathcal{S}_\varepsilon. \end{cases} \tag{3.43}$$

The resulting recovery procedure is summarized in Algorithm 4.

Algorithm 4: Block Phase Retrieval with Truncation (BPR+TR$_\varepsilon$)

Input : Ptychographic measurements $Y \in \mathbb{R}^{m\times d}$ as in (PTY) with $\mathcal{R} = [d]$, truncation parameter $\varepsilon \geq 0$.

Output: $z \in \mathbb{C}^d$ with $z \approx e^{-i\theta}x$ for some $\theta \in [0, 2\pi)$.

1. Construct $F_d[d^j(Z)]_k$ via (3.43).
2. Construct the matrix Z by its diagonals.
3. Form the matrix of phase differences $\operatorname{sgn}_0(Z) \in \mathbb{T}_\delta$.
4. Compute the top eigenvector of $\operatorname{sgn}_0(Z)$, denoted by $\tilde{z} \in \mathbb{C}^d$ with $\|\tilde{z}\|_2 = \sqrt{d}$.
5. Set $z_j = \sqrt{Z_{j,j}} \cdot \operatorname{sgn} \tilde{z}_j$ for all $j \in [d]$ to form $z \in \mathbb{C}^d$.

Clearly, in the noiseless case the diagonal $d^j(Z)$ may not contain some frequencies of $d^j(X)$. Thus, the exact reconstruction of X is not guaranteed. However, by this regularization we achieve better robustness against noise as the next lemma suggests.

Lemma 3.6.10. *Consider the noisy ptychographic measurements (PTY) with all shifts, i.e., $\mathcal{R} = [d]$, and assume that $m \geq 2\delta - 1$. Denote by*

$$\sigma_\varepsilon(\mathcal{A}|_{\mathbb{T}_\delta}) := \sqrt{m} \min_{(j,k)\in\mathcal{S}_\varepsilon^c} |F_d[\overline{w} \circ S_j w]_k| > 0$$

the smallest non-zero singular value above the threshold ε and let

$$s_\varepsilon := |\{k \in [d] : (0,k) \in \mathcal{S}_\varepsilon\}| + 2|\{(j,k) \in \mathcal{S}_\varepsilon : j \neq 0\}|. \tag{3.44}$$

Let X be defined as in (3.31) and Z be the matrix obtained by the reconstruction of its diagonals via (3.43). Then,

$$\|Z - X\|_F \leq \sigma_\varepsilon^{-1}(\mathcal{A}|_{\mathbb{T}_\delta}) \|n\|_2 + \sqrt{\frac{s_\varepsilon}{d}} \|x\|_2^2 .$$

Proof. We aim to use (3.40), which splits the total error $\|Z - X\|_F$ as a sum of errors on the diagonals. Then, we bound the error for each diagonal separately. Using Plancherel's identity (Proposition 2.2.1), the error further splits as

$$\begin{aligned}\left\|d^j(Z) - d^j(X)\right\|_2^2 &= d^{-1} \left\|F_d d^j(Z) - F_d d^j(X)\right\|_2^2 = d^{-1} \sum_{k\in[d]} |F_d[d^j(Z)]_k - F_d[d^j(X)]_k|^2 \\ &= d^{-1} \sum_{k\in[d]} \left| \frac{(F_m^{-1} Y F_d)_{j,k}}{F_d[\overline{w} \circ S_j w]_k} \cdot \mathcal{I}_{(j,k)\in\mathcal{S}_\varepsilon^c} - F_d[d^j(X)]_k \right|^2 .\end{aligned}$$

If $(j,k) \in \mathcal{S}_\varepsilon^c$, we proceed by expanding Y via (PTY) and applying Theorem 3.6.4, similarly to the proof of Corollary 3.6.6,

$$\begin{aligned}\left| \frac{(F_m^{-1} Y F_d)_{j,k}}{F_d[\overline{w} \circ S_j w]_k} \cdot \mathcal{I}_{(j,k)\in\mathcal{S}_\varepsilon^c} - F_d[d^j(X)]_k \right|^2 &= \left| F_d[d^j(X)]_k + \frac{(F_m^{-1} N F_d)_{j,k}}{F_d[\overline{w} \circ S_j w]_k} - F_d[d^j(X)]_k \right|^2 \\ = \left| \frac{(F_m^{-1} N F_d)_{j,k}}{F_d[\overline{w} \circ S_j w]_k} \right|^2 &\leq \sigma_\varepsilon^{-2}(\mathcal{A}|_{\mathbb{T}_\delta}) \left|(\sqrt{m} F_m^{-1} N F_d)_{j,k}\right|^2 .\end{aligned} \tag{3.45}$$

In case $(j,k) \in \mathcal{S}_\varepsilon$, we have

$$\left| \frac{(F_m^{-1} Y F_d)_{j,k}}{\overline{F_d[\overline{w} \circ S_j w]_k}} \cdot \mathcal{I}_{(j,k)\in\mathcal{S}_\varepsilon^c} - F_d[d^j(X)]_k \right|^2 = \left| F_d[d^j(X)]_k \right|^2 . \tag{3.46}$$

The Fourier coefficient is bounded as

$$\begin{aligned} \left| F_d[d^j(X)]_k \right| &= \left| \sum_{\ell\in[d]} d^j(X)_\ell e^{-\frac{2\pi i k\ell}{d}} \right| \le \sum_{\ell\in[d]} |d^j(X)_\ell| = \sum_{\ell\in[d]} |(x \circ S_j\overline{x})_\ell| \\ &= \sum_{\ell\in[d]} |x_\ell| |(S_j\overline{x})_\ell| \le \|x\|_2 \|S_j\overline{x}\|_2 = \|x\|_2^2 , \end{aligned} \tag{3.47}$$

where we used equation (3.33) and Cauchy-Schwartz inequality. Combining bounds (3.45), (3.46) and (3.47), we arrive at

$$\left\| d^j(Z) - d^j(X) \right\|_2^2 \le \sigma_\varepsilon^{-2}(\mathcal{A}|_{\mathbb{T}_\delta}) \sum_{k:(j,k)\in\mathcal{S}_\varepsilon^c} \left| (\sqrt{m} F_m^{-1} N \frac{1}{\sqrt{d}} F_d)_{j,k} \right|^2 + \frac{1}{d} \sum_{k:(j,k)\in\mathcal{S}_\varepsilon} \|x\|_2^4 .$$

Since $\frac{1}{\sqrt{d}} F_d$ is a unitary matrix, the first sum is bounded by the squared norm of j-th row of the matrix $\sqrt{m} F_m^{-1} N$ as

$$\begin{aligned} \sum_{k:(j,k)\in\mathcal{S}_\varepsilon^c} \left| (\sqrt{m} F_m^{-1} N \frac{1}{\sqrt{d}} F_d)_{j,k} \right|^2 &\le \sum_{k\in[d]} \left| (\sqrt{m} F_m^{-1} N \frac{1}{\sqrt{d}} F_d)_{j,k} \right|^2 = \left\| (\sqrt{m} F_m^{-1} N \frac{1}{\sqrt{d}} F_d)_{(j)} \right\|_2^2 \\ &= \left\| (\sqrt{m} F_m^{-1} N)_{(j)} \frac{1}{\sqrt{d}} F_d \right\|_2^2 = \left\| (\sqrt{m} F_m^{-1} N)_{(j)} \right\|_2^2 , \end{aligned}$$

and, thus,

$$\left\| d^j(Z) - d^j(X) \right\|_2^2 \le \sigma_\varepsilon^{-2}(\mathcal{A}|_{\mathbb{T}_\delta}) \left\| (\sqrt{m} F_m^{-1} N)_{(j)} \right\|_2^2 + \frac{1}{d} \sum_{k:(j,k)\in\mathcal{S}_\varepsilon} \|x\|_2^4 .$$

Now, we use (3.40) split the total error $\|Z - X\|_F$ into the errors on diagonals and apply the obtained bound for each diagonal to get

$$\begin{aligned} \|Z - X\|_F^2 \le \sigma_\varepsilon^{-2}(\mathcal{A}|_{\mathbb{T}_\delta}) &\left[\left\| (\sqrt{m} F_m^{-1} N)_{(0)} \right\|_2^2 + 2 \sum_{j=1}^{\delta-1} \left\| (\sqrt{m} F_m^{-1} N)_{(j)} \right\|_2^2 \right] \\ &+ \frac{\|x\|_2^4}{d} \left[\sum_{k:(0,k)\in\mathcal{S}_\varepsilon} 1 + 2 \sum_{j=1}^{\delta-1} \sum_{k:(j,k)\in\mathcal{S}_\varepsilon} 1 \right] . \end{aligned}$$

The second term is exactly s_ε and the first term can be bounded analogously to (3.41). Hence, we have

$$\|Z - X\|_F^2 \le \sigma_\varepsilon^{-2}(\mathcal{A}|_{\mathbb{T}_\delta}) \|n\|_2^2 + \frac{s_\varepsilon}{d} \|x\|_2^4 \le \left[\sigma_\varepsilon^{-1}(\mathcal{A}|_{\mathbb{T}_\delta}) \|n\|_2 + \sqrt{\frac{s_\varepsilon}{d}} \|x\|_2^2 \right]^2 ,$$

which concludes the proof. □

Lemma 3.6.10 highlights a tradeoff between a noise-induced error and an error obtained by discarding information in the measurements via truncation. If the truncation parameter ε is chosen too big, the value $\sigma_\varepsilon^{-1}(\mathcal{A}\big|_{\mathbb{T}_\delta})$ is rather small and the impact of noise is minimized. In contrast, s_ε given by (3.44) may be large and the truncation error is big. On the other hand, for small ε noise becomes a dominating factor in the bound.
It is important to note that if the set $\mathcal{S}_\varepsilon$ is not empty, the exact reconstruction via (3.43) for all $x \in \mathbb{C}^d$ is not possible. In particular, this implies that for the discrete Gabor transform (3.42), even in the noiseless case, the truncation procedure alone is not sufficient.
Since the inversion step is based on the direct inversion of the linear measurements in the space $\mathbb{T}_\delta$, it completely ignores the rank-one structure of $X = T_\delta(xx^*)$. Therefore, δd Fourier coefficients of the diagonals $F_d[d^j(X)]_k$ are recovered, while there are only d unknowns x_k. Hence, the set $\{F_d[d^j(X)]_k : j \in [\delta],\ k \in [d]\}$ contains redundant information about $\{x_k,\ k \in [d]\}$. This implies that there exists a relation between the Fourier coefficients, which allows to express $F_d[d^j(X)]_k$ as a combination of the remaining coefficients. In particular, it is given by

$$d^j(X) \circ S_\ell \overline{d^j(X)} = d^\ell(X) \circ S_j \overline{d^\ell(X)} \text{ for all } j, \ell \in [\delta], \tag{3.48}$$

which is a consequence of (3.33) as

$$d^j(X) \circ S_\ell \overline{d^j(X)} = (x \circ S_j \overline{x}) \circ S_\ell \overline{(x \circ S_j \overline{x})} = (x \circ S_\ell \overline{x}) \circ S_j \overline{(x \circ S_\ell \overline{x})} = d^\ell(X) \circ S_j \overline{d^\ell(X)}.$$

If the coefficients $F_d[d^j(X)]_k$, $(j,k) \in \mathcal{S}_\varepsilon$, are lost due to the truncation procedure, using the surplus information in $F_d[d^j(X)]_k$, $(j,k) \in \mathcal{S}_\varepsilon^c$, their values can be potentially recovered via (3.48). We note that (3.48) is a quadratic relation between the diagonals. Thus, the recovery of the Fourier coefficients in $\mathcal{S}_\varepsilon$ via (3.48) is equivalent to solving a system of quadratic equations, which is in general as hard to solve as the ptychographic recovery itself. However, in specific cases (3.48) reduces to a linear system with respect to the unknown Fourier coefficients. The next theorem provides sufficient conditions for the linear recovery of the lost coefficients motivated by Example 3.6.9.

Theorem 3.6.11 (Subspace completion)**.** *Let X be defined as in (3.31). Assume that*

1. *There is a diagonal $d^\ell(X)$, $\ell \in [\delta]$, such that all Fourier coefficients are recovered, that is $(\ell, k) \in \mathcal{S}_\varepsilon^c$ for all $k \in [d]$.*
2. *For each diagonal $d^j(X)$, $j \in [\delta]$, at most one Fourier coefficient is lost, $|\{k \in [d] : (j,k) \in \mathcal{S}_\varepsilon\}| \leq 1$.*

Then, (3.48) can be expressed as a linear system with respect to the unknown Fourier coefficients and can be solved as a linear regression problem.

Proof. Let $F_d[d^j(X)]_{k_0}$ be the unknown Fourier coefficient of j-th diagonal. First, let us rewrite (3.48) in terms of Fourier coefficients $F_d[d^j(X)]$ via Theorem 2.2.3,

$$v := dF_d[d^\ell(X) \circ S_j \overline{d^\ell(X)}] = dF_d[d^j(X) \circ S_\ell \overline{d^j(X)}] = F_d[d^j(X)] *_d F_d[S_\ell \overline{d^j(X)}].$$

Using Propositions 2.2.2 and 2.2.5, we further rewrite the latter equality as

$$\begin{aligned} v &= F_d[d^j(X)] *_d F_d[S_\ell \overline{d^j(X)}] \\ &= F_d[d^j(X)] *_d M_{-\ell} F_d[\overline{d^j(X)}] = F_d[d^j(X)] *_d M_{-\ell} R_d \overline{F_d[d^j(X)]}. \end{aligned}$$

The expansion of the circular convolution for an index $s \in [d]$ gives us

$$v_s = \sum_{k\in[d]} F_d[d^j(X)]_k (M_{-\ell} R_d \overline{F_d[d^j(X)]})_{s-k} = \sum_{k\in[d]} e^{\frac{2\pi i \ell(k-s)}{d}} F_d[d^j(X)]_k \overline{F_d[d^j(X)]}_{k-s}.$$

Note that for $s = 0$, the right-hand side contains only $|F_d[d^j(X)]_k|^2$ and the phase information about $F_d[d^j(X)]_k$ is lost. Therefore, we will only consider indices $s \geq 1$. The unknown coefficient $F_d[d^j(X)]_{k_0}$ appears in the sum twice, for $k = k_0$ and $k - s \bmod d = k_0$, which are distinct indices for $s \geq 1$. Hence, we separate the unknowns from the rest of the summands,

$$\begin{aligned} u_s &:= v_s - \sum_{\substack{k\in[d] \\ k\neq k_0, k-s \bmod d \neq k_0}} e^{\frac{2\pi i \ell(k-s)}{d}} F_d[d^j(X)]_k \overline{F_d[d^j(X)]}_{k-s} \\ &= e^{\frac{2\pi i \ell(k_0-s)}{d}} F_d[d^j(X)]_{k_0} \overline{F_d[d^j(X)]}_{k_0-s} + e^{\frac{2\pi i \ell k_0}{d}} F_d[d^j(X)]_{k_0+s} \overline{F_d[d^j(X)]}_{k_0}. \end{aligned}$$

Now, separating real and imaginary parts of the unknown coefficient $F_d[d^j(X)]_{k_0}$ leads to

$$\begin{aligned} u_s &= \left[e^{\frac{2\pi i \ell(k_0-s)}{d}} \overline{F_d[d^j(X)]}_{k_0-s} + e^{\frac{2\pi i \ell k_0}{d}} F_d[d^j(X)]_{k_0+s} \right] \operatorname{Re} F_d[d^j(X)]_{k_0} \\ &\quad + i \left[e^{\frac{2\pi i \ell(k_0-s)}{d}} \overline{F_d[d^j(X)]}_{k_0-s} - e^{\frac{2\pi i \ell k_0}{d}} F_d[d^j(X)]_{k_0+s} \right] \operatorname{Im} F_d[d^j(X)]_{k_0} \\ &=: a_s \operatorname{Re} F_d[d^j(X)]_{k_0} + i b_s \operatorname{Im} F_d[d^j(X)]_{k_0}. \end{aligned}$$

In the matrix form this reads as

$$\begin{bmatrix} \operatorname{Re} u_s \\ \operatorname{Im} u_s \end{bmatrix} = \begin{bmatrix} \operatorname{Re} a_s & -\operatorname{Im} b_s \\ \operatorname{Im} a_s & \operatorname{Re} b_s \end{bmatrix} \begin{bmatrix} \operatorname{Re} F_d[d^j(X)]_{k_0} \\ \operatorname{Im} F_d[d^j(X)]_{k_0} \end{bmatrix} =: Q_s \begin{bmatrix} \operatorname{Re} F_d[d^j(X)]_{k_0} \\ \operatorname{Im} F_d[d^j(X)]_{k_0} \end{bmatrix}.$$

Combining all $s \geq 1$, we obtain

$$\begin{bmatrix} \operatorname{Re} u_1 \\ \operatorname{Im} u_1 \\ \vdots \\ \operatorname{Re} u_{d-1} \\ \operatorname{Im} u_{d-1} \end{bmatrix} = \begin{bmatrix} Q_1 \\ \vdots \\ Q_{d-1} \end{bmatrix} \begin{bmatrix} \operatorname{Re} F_d[d^j(X)]_{k_0} \\ \operatorname{Im} F_d[d^j(X)]_{k_0} \end{bmatrix}. \tag{3.49}$$

Finally, this linear system can be solved to obtain $F_d[d^j(X)]_{k_0}$. □

Remark 3.6.12. *We note that even if it is possible to establish the linear system (3.49), it can be ill-conditioned. For instance, if all coefficients $F_d[d^j(X)]_k = 0$, $k \neq k_0$, the matrices Q_s, $s \geq 1$, are zero matrices and the linear system (3.49) cannot be solved.*

If the assumptions of Theorem 3.6.11 hold, it is possible to recover lost Fourier coefficients by solving the linear system (3.49). Otherwise, the same linear system can be used as a heuristic procedure to improve the performance of the inversion step. In this case, all the lost Fourier coefficients are initially set to zero and then they can be estimated one by one by solving (3.49).
The revised version of BPR with subspace completion is summarized in Algorithm 5.

Algorithm 5: Block Phase Retrieval with Subspace Completion (BPR+SC$_\varepsilon$)

Input : Ptychographic measurements $Y \in \mathbb{R}^{m\times d}$ as in (PTY) with $\mathcal{R} = [d]$, truncation parameter $\varepsilon \geq 0$.

Output: $z \in \mathbb{C}^d$ with $z \approx e^{-i\theta}x$ for some $\theta \in [0, 2\pi)$.

1. Construct $F_d[d^j(Z)]_k$ via (3.43).
2. One by one reconstruct $F_d[d^j(Z)]_k$ for $(j,k) \in \mathcal{S}_\varepsilon$ via (3.49).
3. Construct the matrix Z by its diagonals.
4. Form the matrix of phase differences $\operatorname{sgn}_0(Z) \in \mathbb{T}_\delta$.
5. Compute the top eigenvector of $\operatorname{sgn}_0(Z)$, denoted by $\tilde{z} \in \mathbb{C}^d$ with $\|\tilde{z}\|_2 = \sqrt{d}$.
6. Set $z_j = \sqrt{Z_{j,j}} \cdot \operatorname{sgn} \tilde{z}_j$ for all $j \in [d]$ to form $z \in \mathbb{C}^d$.

We do not provide theoretical guarantees for Algorithm 5.

Notes and References. *The material in this section is based on our publication [140]. We did not derive a reconstruction error bound for the inversion with subspace completion analogous to Corollary 3.6.6 or Lemma 3.6.10. While an error bound is achievable, it would quadratically depend on the noise norm, which limits its use unless the noise level is small.*

3.6.3 Magnitude estimation

In this section, we study different techniques to recover the magnitudes $|x|$ from the matrix $X = T_\delta(xx^*)$.

3.6.3.1 Diagonal Magnitude Estimation

The first recovery method is the reconstruction from the main diagonal mentioned in Section 3.6.1, which is based on the equality

$$\sqrt{X_{j,j}} = \sqrt{d^0(X)_j} = \sqrt{(x \circ \overline{x})_j} = \sqrt{|x_j|^2} = |x_j|.$$

It is a fast way to construct an estimate of the magnitudes and the reconstruction error bound is given by the next lemma.

Lemma 3.6.13 ([138, Lemma 7]). *Let $X, Z \in \mathbb{H}^d$ and set $|x_j| = \sqrt{X_{j,j}}$, $|z_j| = \sqrt{|Z_{j,j}|}$. Then,*

$$\||x| - |z|\|_2 \leq d^{1/4}\sqrt{\|X - Z\|_F}.$$

Note that the bound in Lemma 3.6.13 depends on the dimension d. In the next subsection we will state a dimensional-independent alternative to Lemma 3.6.13.
The major drawback of the Diagonal Estimation technique is the fact, that it ignores the entries of X, which do not belong to the main diagonal. In the rest of Section 3.6.3, we provide two alternatives, which use information beyond the main diagonal.

3.6.3.2 Block Magnitude Estimation

In principle, inclusion of the off-diagonal entries should improve noise robustness of the magnitude estimation. This motivated the authors in [40] and [138, 159] to explore empirically and, respectively, theoretically a new method called Blockwise Magnitude Estimation involving more entries of X. The main idea of the method is to extract blocks from X, to estimate the magnitudes of x corresponding to each block separately and then to combine the estimates together. More precisely, let us consider index sets $\mathcal{J}_p \subseteq [d]$, $p \in [P]$ for some $P \in \mathbb{N}$ and vectors

$$(\mathcal{I}_{\mathcal{J}_p})_k = \begin{cases} 1, & k \in \mathcal{J}_p, \\ 0, & \text{otherwise.} \end{cases}$$

Then, construct a block X^p, $p \in [P]$ by preserving the entries with indices in $\mathcal{J}_p$ and nullifying the rest,

$$(X^p)_{k,\ell} := (\operatorname{diag}(\mathcal{I}_{\mathcal{J}_p}) X \operatorname{diag}(\mathcal{I}_{\mathcal{J}_p}))_{k,\ell} = X_{k,\ell}(\mathcal{I}_{J_p})_k(\mathcal{I}_{J_p})_\ell = \begin{cases} X_{k,\ell}, & k,\ell \in \mathcal{J}_p, \\ 0, & \text{otherwise.} \end{cases}$$

We require that

$$|\ell - k|_c < \delta \quad \text{for all } j,k \in \mathcal{J}_p \tag{3.50}$$

to ensure that all entries of X^p are recovered by the inversion step, i.e., supported withing $\mathbb{T}_\delta$ as shown in Figure 3.2. That is, by (3.50), for all $k,\ell \in \mathcal{J}_p$ the corresponding diagonal $x \circ S_{k-\ell}\overline{x}$ is recovered via Theorem 3.6.4. Then, each block X^p is a rank-one matrix

$$X^p = x^p(x^p)^* \text{ and } |X^p| = |x^p||x^p|^*, \quad p \in [P],$$

where we use the notation

$$x^p := \mathcal{I}_{\mathcal{J}_p} \circ x. \tag{3.51}$$

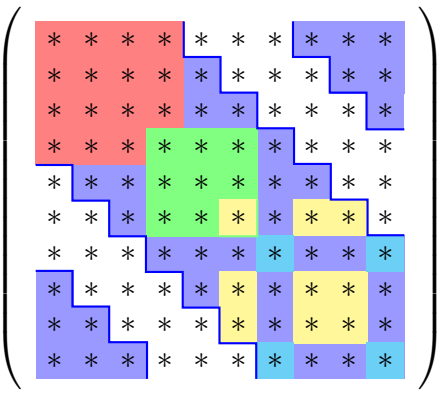


Figure 3.2: Highlighted in blue are the entries, which form the space $\mathbb{T}_\delta$ for $d = 10$, $\delta = 4$. The family of sets $\{\mathcal{J}_p\}_{p\in[4]}$ is depicted in colors with $\mathcal{J}_0 = \{0,1,2,3\}$ in red, $\mathcal{J}_1 = \{3,4,5\}$ in green, $\mathcal{J}_2 = \{5,7,8\}$ in yellow and $\mathcal{J}_3 = \{6,9\}$ in cyan. Note that $\{\mathcal{J}_p\}_{p\in[4]}$ satisfies (3.50) since all colored entries are part of $\mathbb{T}_\delta$. Furthermore, all elements of the main diagonal are covered and, thus, (3.54) is satisfied as well.

Therefore, $|x^p|$ can be recovered by computing the top eigenvector u^p, $\|u^p\|_2 = 1$ of the matrix $|X^p|$ corresponding to the largest magnitude eigenvalue $\||X^p|\|_\infty$ and observing that

$$|x^p| = \sqrt{\||X^p|\|_\infty}|u^p|.$$

If an index $k \in [d]$ belongs to multiple sets $\mathcal{J}_p$, the entry $|x_k|$ is estimated several times. Thus, by setting the counts

$$\mu_k := |\{p \in [P] : k \in \mathcal{J}_p\}| \tag{3.52}$$

as the number of times the entry $|x_k|$ is estimated, averaging the estimates yields

$$|x_k| = \frac{1}{\mu_k} \sum_{p \in [P]} \sqrt{\||X^p|\|_\infty}|u^p_k|.$$

The established estimate in the vector form reads as

$$|x| = \operatorname{diag}(1/\mu) \sum_{p \in [P]} \sqrt{\||X^p|\|_\infty}|u^p|. \tag{3.53}$$

Note that if for some $k \in [d]$, the count μ_k is zero, the corresponding entry $|x_k|$ was not recovered. To avoid such scenarios, we require that all $\mu_k > 0$, so that $\{\mathcal{J}_p\}_{p \in [P]}$ is a covering of the set $[d]$,

$$[d] \subseteq \bigcup_{p \in [P]} \mathcal{J}_p. \tag{3.54}$$

The Block Magnitude Estimation technique is summarized in Algorithm 6.

Algorithm 6: Block Magnitude Estimation, version of [138, 159]

Input : Matrix $Z \in \mathbb{T}_\delta$, a noisy version of X, sets $\{\mathcal{J}_p\}_{p \in [P]}$, satisfying (3.50) and (3.54).

Output: $v \in \mathbb{R}^d$ with $v \approx |x|$.

1. Construct $Z^p = \operatorname{diag}(\mathcal{I}_{\mathcal{J}_p}) Z \operatorname{diag}(\mathcal{I}_{\mathcal{J}_p}) \in \mathbb{H}^d$ for all $p \in [P]$.
2. Compute the largest magnitude eigenvalue $\||Z^p|\|_\infty$ and the corresponding eigenvector $v^p \in \mathbb{C}^d$ with $\|v^p\|_2 = 1$ of the matrix $|Z^p|$ for all $p \in [P]$.
3. Compute $\mu_k = |\{p \in [P] : k \in \mathcal{J}_p\}|$ for $k \in [d]$.
4. Construct $v = \operatorname{diag}(1/\mu) \sum_{p \in [P]} \sqrt{\||Z^p|\|_\infty}|v^p|$.

The computational complexity of Algorithm 6 can be computed by summing up the complexities of the separate steps. In view of the condition (3.50), each $\mathcal{J}_p$ has at most δ entries and, thus, for a single set $\mathcal{J}_p$ Step 1 requires $\mathcal{O}(\delta^2)$ operations and Step 2 requires $\mathcal{O}(\delta^2 \log \delta)$ operations to reach the machine precision. Also, for the same reason Step 3 is performed in $\mathcal{O}(P\delta)$ operations. At last, Step 4 requires $\mathcal{O}(P\delta + d)$ operations which grants the total complexity of $\mathcal{O}(d + P\delta^2 \log \delta)$. It is higher compared to the Main Diagonal method which only requires $\mathcal{O}(d)$ operations, but if $P = \mathcal{O}(d)$ and δ is small, the difference is not extreme.

The recovery guarantees for Algorithm 6 are provided by the next lemma.

Lemma 3.6.14 (Version of [159, Proposition 6]). *Let $x \in \mathbb{C}^d$, $Z \in \mathbb{T}_\delta$ and consider X as in (3.31). Further, let $\{\mathcal{J}_p\}_{p\in[P]}$ be a family of sets satisfying conditions (3.50) and (3.54). Let v be the output of Algorithm 6. Then,*

$$\|v - |x|\|_2 \leq \sqrt{\frac{\max_{k\in[d]} \mu_k}{\min_{j\in[d]} \mu_j}} \cdot \frac{(1+2\sqrt{2})\,\|Z - X\|_F}{\min_{p\in[P]} \|x^p\|_2},$$

where μ_k and x^p are given by equations (3.52) and (3.51), respectively.

Proof of Lemma 3.6.14. Using the sum representations of v and $|x|$ we get

$$\|v - |x|\|_2^2 = \left\| \operatorname{diag}(1/\mu) \left[\sum_{p\in[P]} \sqrt{\||Z^p|\|_\infty}|v^p| - \sqrt{\||X^p|\|_\infty}|u^p| \right] \right\|_2^2.$$

Recall that $\sqrt{\||X^p|\|_\infty}|u^p| = |x^p|$ and expand the square of the norm on the right-hand side as

$$\|v - |x|\|_2^2 = \sum_{k\in[d]} \frac{1}{\mu_k^2} \left[\sum_{p\in[P]} \sqrt{\||Z^p|\|_\infty}|v_k^p| - |x_k^p| \right]^2.$$

For a fixed $k \in [d]$ each of the inner sums contains μ_k non-zero summands. We can apply inequality

$$\left[\sum_{j\in[n]} a_j \right]^2 \leq n \sum_{j\in[n]} a_j^2$$

for all $a \in \mathbb{R}^n$ to obtain

$$\begin{aligned} \|v - |x|\|_2^2 &\leq \sum_{k\in[d]} \frac{1}{\mu_k} \sum_{p\in[P]} \left| \sqrt{\||Z^p|\|_\infty}|v_k^p| - |x_k^p| \right|^2 \\ &\leq \frac{1}{\min_{j\in[d]} \mu_j} \sum_{p\in[P]} \sum_{k\in[d]} \left| \sqrt{\||Z^p|\|_\infty}|v_k^p| - |x_k^p| \right|^2. \end{aligned} \tag{3.55}$$

By the reverse triangle inequality, each of the inner sums is bounded by

$$\sum_{k\in[d]} \left| \sqrt{\||Z^p|\|_\infty}|v_k^p| - |x_k^p| \right|^2 \leq \sum_{k\in[d]} \left| |x_k^p| - \alpha \sqrt{\||Z^p|\|_\infty} v_k^p \right|^2,$$

for any α, $|\alpha| = 1$. Choosing α such that the right-hand side is minimized grants us

$$\begin{aligned} \sum_{k\in[d]} \left| \sqrt{\||Z^p|\|_\infty}|v_k^p| - |x_k^p| \right|^2 &\leq \min_{|\alpha|=1} \sum_{k\in[d]} \left| |x_k^p| - \alpha \sqrt{\||Z^p|\|_\infty} v_k^p \right|^2 \\ &= \operatorname{dist}^2 \left(|x^p|, \sqrt{\||Z^p|\|_\infty} v^p \right). \end{aligned}$$

By Lemma 2.1.1, we bound the resulting distance as

$$\sum_{k\in[d]} \left| \sqrt{\| |Z^p| \|_\infty} |v_k^p| - |x_k^p| \right|^2 \leq \frac{(1+2\sqrt{2})^2 \, \| |Z^p| - |X^p| \|_F^2}{\| |x^p| \|_2^2}.$$

Note that $\| |x^p| \|_2 = \|x^p\|_2$ and by the reverse triangle inequality $\| |Z^p| - |X^p| \|_F \leq \|Z^p - X^p\|_F$. Hence, the final bound on the inner sum reads as

$$\sum_{k\in[d]} \left| \sqrt{\| |Z^p| \|_\infty} |v_k^p| - |x_k^p| \right|^2 \leq \frac{(1+2\sqrt{2})^2 \, \|Z^p - X^p\|_F^2}{\|x^p\|_2^2}.$$

Applying it to (3.55), we obtain

$$\begin{aligned} \| v - |x| \|_2^2 &\leq \frac{(1+2\sqrt{2})^2}{\min_{j\in[d]} \mu_j} \sum_{p\in[P]} \frac{\|Z^p - X^p\|_F^2}{\|x^p\|_2^2} \\ &\leq \frac{1}{\min_{j\in[d]} \mu_j} \cdot \frac{(1+2\sqrt{2})^2}{\min_{p\in[P]} \|x^p\|_2^2} \sum_{p\in[P]} \|Z^p - X^p\|_F^2 . \end{aligned}$$

In the sum $\sum_{p\in[P]} \|Z^p - X^p\|_F^2$, the entries of the matrix $Z - X$ appear multiple times. More precisely, by the definitions of Z^p and X^p we have that

$$\begin{aligned} \sum_{p\in[P]} \|Z^p - X^p\|_F^2 &= \sum_{p\in[P]} \sum_{k,\ell\in[d]} |(Z^p - X^p)_{k,\ell}|^2 = \sum_{p\in[P]} \sum_{k,\ell\in[d]} |(Z-X)_{k,\ell}|^2 \mathcal{I}_{k,\ell\in\mathcal{J}_p} \\ &= \sum_{k,\ell\in[d]} |(Z-X)_{k,\ell}|^2 \sum_{p\in[P]} \mathcal{I}_{k,\ell\in\mathcal{J}_p} \leq \sum_{k,\ell\in[d]} |(Z-X)_{k,\ell}|^2 \sum_{p\in[P]} \mathcal{I}_{k\in\mathcal{J}_p} \\ &= \sum_{k,\ell\in[d]} |(Z-X)_{k,\ell}|^2 \mu_k \leq \max_{k\in[d]} \mu_k \, \|Z - X\|_F^2 . \end{aligned}$$

Thus, we obtain

$$\| v - |x| \|_2^2 \leq \frac{\max_{k\in[d]} \mu_k}{\min_{j\in[d]} \mu_j} \cdot \frac{(1+2\sqrt{2})^2 \, \|Z - X\|_F^2}{\min_{p\in[P]} \|x^p\|_2^2},$$

which concludes the proof. □

The bound of Lemma 3.6.14 depends on three components: the counts μ_k, the norms $\|x^p\|_2$ and the error of the inversion step $\|Z - X\|_F$. It is dimension independent in comparison to Lemma 3.6.13, since for a reasonable choice of family $\{\mathcal{J}_p\}_{p\in[P]}$, the term $\frac{\max_{k\in[d]} \mu_k}{\min_{j\in[d]} \mu_j}$ is either bounded by δ or even an absolute constant. However, the bound depends on $\min \|x^p\|_2$, which may behave poorly. Yet, we will observe a similar dependency for the phase reconstruction in Section 3.6.4.

For the rest of the subsection on Block Magnitude Estimations, we would like to discuss a special construction of sets $\{\mathcal{J}_p\}_{p\in[P]}$, which is convenient to work with. Consider the family $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ of $P = d$ sets

$$\mathcal{J}_p^\gamma := \{p, p+1, \ldots, p+\gamma-1\}, \tag{3.56}$$

with width parameter $1 \leq \gamma \leq \delta$.

Lemma 3.6.15. *Let $1 \leq \gamma \leq \delta$. The family of sets $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ satisfies conditions (3.50) and (3.54) with $\mu_k = \gamma$.*

Proof. Since $\{p\} \subseteq \mathcal{J}_p^\gamma$, we have that

$$[d] = \bigcup_{p\in[d]} \{p\} \subseteq \bigcup_{p\in[d]} \mathcal{J}_p^\gamma.$$

and the condition (3.54) is satisfied. Also, by the definition of $\mathcal{J}_p^\gamma$, for all $k, \ell \in \mathcal{J}_p^\gamma$,

$$|\ell - k|_c < \gamma \leq \delta,$$

and (3.50) holds true. From the last equation, we observed that δ is the maximal possible choice of γ for (3.50) to be satisfied. For $k \in [d]$, we have that $k \in \mathcal{J}_p^\gamma$, $p \in \{k-\gamma+1, \dots, k\}$ and, thus, $\mu_k = \gamma$. □

This family has a few benefits to work with. Firstly, it uses all entries contained in the diagonals $d^{-\gamma+1}(Z), \dots, d^0(Z), \dots, d^{\gamma-1}(Z)$. In particular, if Block Magnitude Estimation is applied to $\{\mathcal{J}_p^\delta\}_{p\in[d]}$, all entries of the matrix Z are used. Secondly, every entry of $|x|$ is estimated precisely γ times, so that $\mu_k = \gamma$ for all $k \in [d]$.
Considering the fact that for larger γ Algorithm 6 would use more entries of Z, it may seem that using family $\{\mathcal{J}_p^\delta\}_{p\in[d]}$ would provide the best possible reconstruction. However, that is not entirely true. On one hand, monotonicity applies.

Example 3.6.16. *Consider two families $\{\mathcal{J}_{p'}\}_{p'\in[P]}$ and $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ such that*

$$\text{every set } \mathcal{J}_p^\gamma \text{ is a superset of } \mathcal{J}_{p'} \text{ for some } p' \in [P]. \tag{3.57}$$

Then, we have

$$\left\|x^{p'}\right\|_2^2 = \left\|\mathcal{I}_{\mathcal{J}_{p'}} x\right\|_2^2 = \sum_{j\in[d]} |x_j|^2 \mathcal{I}_{j\in\mathcal{J}_{p'}} \leq \sum_{j\in[d]} |x_j|^2 \mathcal{I}_{j\in\mathcal{J}_p^\gamma} = \left\|\mathcal{I}_{\mathcal{J}_p^\gamma} x\right\|_2^2 = \left\|x^p\right\|_2^2.$$

Hence, there exists $p' \in [P]$ such that

$$\min_{p\in[d]} \left\|x^p\right\|_2 \geq \left\|x^{p'}\right\|_2 \geq \min_{p'\in[P]} \left\|x^{p'}\right\|_2.$$

Moreover, $\max_{k\in[d]} \mu_k / \min_{j\in[d]} \mu_j \geq 1$ and, thus, the error bound provided by Lemma 3.6.14 for $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ is smaller than for $\{\mathcal{J}_{p'}\}_{p'\in[P]}$. Therefore, if the condition (3.57) holds, the family $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ is a better choice in terms of magnitude estimation error. In particular, since $\mathcal{J}_p^\gamma \subseteq \mathcal{J}_p^\delta$ for all $p \in [d]$, the maximal width family $\{\mathcal{J}_p^\delta\}_{p\in[d]}$ is the best choice.

On the other hand, it is possible to construct a counterexample, in which (3.57) does not hold and the family $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ is suboptimal in term of the error bound.

Example 3.6.17. *Let $2 \leq \gamma \leq \delta$ and $d \geq 2\gamma - 1$. Consider two families $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ and $\{\mathcal{J}_p^\gamma\}_{p\in[d]\setminus\{0\}}$. Note that $\mathcal{J}_p^\gamma$, $p \neq 0$ are in both families and for $\mathcal{J}_0^\gamma$ there is no set in $\{\mathcal{J}_p^\gamma\}_{p\in[d]\setminus\{0\}}$, which is contained in $\mathcal{J}_0^\gamma$. Therefore, (3.57) does not hold.*

For the family $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ *the counts satisfy* $\mu_k = \gamma$ *for all* $k \in [d]$ *and for* $\{\mathcal{J}_p^\gamma\}_{p\in[d]\setminus\{0\}}$ *the counts are either* $\mu_k^{\{0\}} = \gamma$ *or* $\mu_k^{\{0\}} = \gamma - 1 > 0$.
Let $0 < \varepsilon < 1$ *be arbitrary and consider the object*

$$x_j = \begin{cases} \varepsilon, & j \in [\gamma], \\ 1, & \text{otherwise.} \end{cases}$$

The squared norms $\|x^p\|_2^2$ *are given by*

$$\|x^p\|_2^2 = \sum_{j\in\mathcal{J}_p^\gamma} |x_j|^2 = \begin{cases} (\gamma - |p|)\varepsilon^2 + |p|, & p \in \{-\gamma+1, \dots, \gamma+1\}, \\ \gamma, & \text{otherwise.} \end{cases}$$

Consequently, using that $\varepsilon < 1$, *the minimums of norms for both families are given by*

$$\min_{p\in[d]} \|x^p\|_2^2 = \left\|x^0\right\|_2^2 = \gamma\varepsilon^2$$

and

$$\min_{p\in[d]\setminus\{0\}} \|x^p\|_2^2 = \left\|x^1\right\|_2^2 = (\gamma-1)\varepsilon^2 + 1$$

Therefore, the family dependent factors in the error bound of Lemma 3.6.14 are

$$\frac{1}{\sqrt{\gamma}\varepsilon} \text{ for } \{\mathcal{J}_p^\gamma\}_{p\in[d]} \text{ and } \sqrt{\frac{\gamma}{\gamma-1}} \cdot \frac{1}{\sqrt{(\gamma-1)\varepsilon^2+1}} \text{ for } \{\mathcal{J}_p^\gamma\}_{p\in[d]\setminus\{0\}}.$$

When ε *tends to zero from above, the left fraction diverges to infinity, while the right fraction converges to* $\sqrt{\gamma/(\gamma-1)}$. *Hence, for a sufficiently small* ε, *the error bound of Lemma 3.6.14 for the family* $\{\mathcal{J}_p^\gamma\}_{p\in[d]\setminus\{0\}}$ *is lower than for the family* $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$.
Now, consider $x = \mathbb{1}_d$, *i.e.,* $x_k = 1$ *for all* $k \in [d]$. *Then,* $\|x^p\|_2^2 = \gamma$ *for all* $p \in [d]$, *and, thus, the minimums for both families coincide,*

$$\min_{p\in[d]} \|x^p\|_2^2 = \min_{p\in[d]\setminus\{0\}} \|x^p\|_2^2 = \gamma.$$

Hence, the family dependent factors in the error bound of Lemma 3.6.14 are

$$\frac{1}{\sqrt{\gamma}} \text{ for } \{\mathcal{J}_p^\gamma\}_{p\in[d]} \text{ and } \sqrt{\frac{\gamma}{\gamma-1}} \cdot \frac{1}{\sqrt{\gamma}} = \frac{1}{\sqrt{\gamma-1}} \text{ for } \{\mathcal{J}_p^\gamma\}_{p\in[d]\setminus\{0\}}.$$

Clearly, the factor for $\{\mathcal{J}_p^\gamma\}_{p\in[d]}$ *is smaller for this object.*

From these two examples we can conclude that the best choice in terms of the error bound is always a subfamily of $\{\mathcal{J}_p^\delta\}_{p\in[d]}$, depending on x. Since the object is unknown, it is impossible to select the optimal family. In view of this uncertainty, $\{\mathcal{J}_p^\delta\}_{p\in[d]}$ might be the best choice.
At last, let us consider the special case $\gamma = 1$. Following Algorithm 6, the constructed matrices $|Z^p|$ only have a single non-zero element $|Z_{p,p}| \geq 0$ on the main diagonal. Thus,

it is a rank-one matrix with $\||Z^p|\|_\infty = |Z_{p,p}|$ and the corresponding eigenvector being the standard basis vector e_p. Therefore, the k-th entry of the estimate v is equal to

$$v_k = \frac{1}{\mu_k} \sum_{p\in[d]} \sqrt{\||Z^p|\|_\infty}|(e_p)_k| = \sqrt{|Z_{k,k}|},$$

which implies that for $\{\mathcal{J}_p^1\}_{p\in[d]}$ Block Magnitude Estimation coincides with Diagonal Magnitude Estimation and Lemma 3.6.14 provides the dimension-independent error bound.

Corollary 3.6.18 ([159, Proposition 6]). *Let $X, Z \in \mathbb{H}^d$ and set $|x_j| = \sqrt{X_{j,j}}$, $|z_j| = \sqrt{|Z_{j,j}|}$. Then,*

$$\||x| - |z|\|_2 \le \frac{1+2\sqrt{2}}{\min_{k\in[d]} |x_k|^2} \|X - Z\|_F .$$

3.6.3.3 Log Magnitude Estimation

The third method for the magnitude reconstruction is based on the idea of converting the problem to a linear system by applying the logarithm to the entries of $|X|$, which is why we call this method Log Magnitude Estimation. The logarithm transform is used in the linear regression to transform the multiplicative model into an additive model in terms of logarithms and solve the least squares problem. Recall that the non-zero entries in the lower triangular part of the matrix X are recovered by diagonals $d^j(X) = x \circ S_j\overline{x}$, $j \in [\delta]$. Consequently, for all $k \in [d]$, $j \in [\delta]$, we have

$$|X_{k,k-j}| = |d^j(X)_k| = |(x \circ S_j\overline{x})_k| = |x_k| \cdot |x_{k-j}|.$$

Assuming that $|x_k| \neq 0$ for all $k \in [d]$, we apply the logarithm on both sides which leads to the linear equations

$$\log|X_{k,k-j}| = \log|x_k| + \log|x_{k-j}|. \tag{3.58}$$

Combining these equations in a matrix, we obtain

$$b(X) = B \log|x|,$$

with the double indexed vector $b(X) \in \mathbb{R}^{d\delta}$ defined as

$$b_{(k,j)}(X) := \log|X_{k,k-j}|, \quad k \in [d],\ j \in [\delta], \tag{3.59}$$

and the matrix $B \in \mathbb{R}^{d\delta\times d}$ with entries

$$B_{(k,j),\ell} := \begin{cases} 2, & j = 0 \text{ and } k = \ell, \\ 1, & j \neq 0, \text{ and either } k = \ell \text{ or } k - j = \ell, \\ 0, & \text{otherwise,} \end{cases} \tag{3.60}$$

for $k, \ell \in [d]$, $j \in [\delta]$. Alternatively, the entries of B can be compactly written via indicators,

$$B_{(k,0),\ell} = 2\mathcal{I}_{k=\ell}, \quad B_{(k,j),\ell} = \mathcal{I}_{k=\ell} + \mathcal{I}_{k-j=\ell}, \quad k \in [d],\ j \in [d]\backslash\{0\}. \tag{3.61}$$

Therefore, if B is injective, $\log|x|$ can be recovered as a solution of the least squares problem, that is

$$\log|x| = B^{\dagger}b(X) = (B^*B)^{-1}B^*b(X),$$

with $B^{\dagger}$ being the pseudoinverse, and consequently,

$$|x| = e^{\log|x|} = e^{(B^*B)^{-1}B^*b(X)}, \tag{3.62}$$

where all functions are applied entrywise.
The next theorem shows that B is injective and provides a direct computation of the inverse matrix $(B^*B)^{-1}$.

Theorem 3.6.19. *Let $d \geq 2\delta - 1$. Consider the matrix B defined in (3.60). Then, B^*B admits the decomposition*

$$B^*B = \frac{1}{\sqrt{d}}F_d^* \operatorname{diag}(u)\frac{1}{\sqrt{d}}F_d,$$

with $u \in \mathbb{R}^d$ containing the eigenvalues

$$u_k = \begin{cases} 4\delta, & k = 0, \\ 2\delta + 1 + \frac{\sin\left(\frac{\pi(2\delta-1)k}{d}\right)}{\sin(\pi k/d)}, & k \in [d]\backslash\{0\}, \end{cases}$$

and satisfying

$$u_k \geq 4,$$

for all $k \in [d]$. Consequently, its inverse is given by

$$(B^*B)^{-1} = \frac{1}{\sqrt{d}}F_d^* \operatorname{diag}(1/u)\frac{1}{\sqrt{d}}F_d.$$

Proof. Let us compute the entries of $B^*B \in \mathbb{R}^{d\times d}$. For $\ell, s \in [d]$, we write

$$(B^*B)_{\ell,s} = \sum_{k\in[d]}\sum_{j\in[\delta]} B_{(k,j),\ell}B_{(k,j),s} = \sum_{k\in[d]} B_{(k,0),\ell}B_{(k,0),s} + \sum_{j\in[\delta]\backslash\{0\}}\sum_{k\in[d]} B_{(k,j),\ell}B_{(k,j),s}.$$

Substituting the indicator representation (3.61) for the entries of B, we obtain

$$\begin{aligned}(B^*B)_{\ell,s} &= \sum_{k\in[d]} 4\mathcal{I}_{k=\ell}\mathcal{I}_{k=s} + \sum_{j\in[\delta]\backslash\{0\}}\sum_{k\in[d]} [\mathcal{I}_{k=\ell} + \mathcal{I}_{k-j=\ell}]\cdot[\mathcal{I}_{k=s} + \mathcal{I}_{k-j=s}] \\ &= \sum_{k\in[d]} 4\mathcal{I}_{k=\ell=s} + \sum_{j\in[\delta]\backslash\{0\}}\sum_{k\in[d]} [\mathcal{I}_{k=\ell=s} + \mathcal{I}_{s-\ell=j,k=s} + \mathcal{I}_{k=\ell,j=\ell-s} + \mathcal{I}_{k-j=s=\ell}] \\ &= 4\mathcal{I}_{\ell=s} + \sum_{j\in[\delta]\backslash\{0\}} [\mathcal{I}_{\ell=s} + \mathcal{I}_{s-\ell=j} + \mathcal{I}_{\ell-s=j} + \mathcal{I}_{s=\ell}]\end{aligned}$$

If $\ell = s$, the indicators $\mathcal{I}_{s-\ell=j}$ and $\mathcal{I}_{j=s-\ell}$ are zero and the rest are ones, so that

$$(B^*B)_{\ell,\ell} = 4 + \sum_{j\in[\delta]\backslash\{0\}} 2 = 4 + 2(\delta - 1) = 2\delta + 2. \tag{3.63}$$

If $\ell \neq s$, the indicator $\mathcal{I}_{s=\ell}$ is zero. For the second sum, if $\mathcal{I}_{s-\ell=j_0} = 1$ for some $0 < j_0 \leq \delta - 1$, then $s - \ell = j_0$ and

$$\ell - s \bmod d = -j_0 \bmod d = -j_0 + d \geq -\delta + 1 + 2\delta - 1 = \delta > j_0.$$

Hence, the indicator $\mathcal{I}_{\ell-s=j} = 0$ for all $j \in [\delta]\backslash\{0\}$. Analogously, if $\mathcal{I}_{\ell-s=j_0} = 1$ for some $0 < j_0 \leq \delta - 1$, then $\mathcal{I}_{s-\ell=j} = 0$ for all $j \in [\delta]\backslash\{0\}$. Consequently, there exists at most one index j such that either $\mathcal{I}_{s-\ell=j} = 1$ or $\mathcal{I}_{\ell-s=j} = 1$. Moreover, if $|s - \ell|_c \geq \delta$, both indicators are zero due to the inequality $j < \delta$. Consequently,

$$(B^*B)_{\ell,s} = \mathcal{I}_{|s-\ell|_c<\delta}\mathcal{I}_{\ell\neq s}. \tag{3.64}$$

For the decomposition of B^*B, we use equations (3.63) and (3.64) to rewrite the entries of B^*B as

$$\begin{aligned}(B^*B)_{\ell,s} &= (2\delta+2)\mathcal{I}_{\ell=s} + \mathcal{I}_{|\ell-s|_c<\delta}\mathcal{I}_{\ell\neq s} = (2\delta+1)\mathcal{I}_{\ell=s} + \mathcal{I}_{|\ell-s|_c<\delta} \\ &= (2\delta+1)(I_d)_{\ell,s} + (\mathbb{1}_{d\times d})_{\ell,s}\mathcal{I}_{|\ell-s|_c<\delta} = (2\delta+1)(I_d)_{\ell,s} + (T_\delta(\mathbb{1}_{d\times d}))_{\ell,s},\end{aligned}$$

where $\mathbb{1}_{d\times d} \in \mathbb{C}^{d\times d}$ denotes the matrix with all entries set to 1 and $T_\delta(\mathbb{1}_{d\times d})$ denotes its projection onto $\mathbb{T}_\delta$ given by (3.30). Therefore,

$$B^*B = (2\delta+1)I_d + T_\delta(\mathbb{1}_{d\times d}). \tag{3.65}$$

Note that by (3.30), the matrix $T_\delta(\mathbb{1}_{d\times d})$ is Hermitian and circulant, that is for all $k, \ell, s \in [d]$ the equality

$$T_\delta(\mathbb{1}_{d\times d})_{k+s,\ell+s} = (\mathbb{1}_{d\times d})_{k+s,\ell+s}\mathcal{I}_{|k+s-\ell-s|<\delta} = \mathcal{I}_{|k-\ell|<\delta} = (\mathbb{1}_{d\times d})_{k,\ell}\mathcal{I}_{|k-\ell|<\delta} = T_\delta(\mathbb{1}_{d\times d})_{k,\ell}$$

holds. Therefore, by Theorem 2.2.4, $T_\delta(\mathbb{1}_{d\times d})$ admits the decomposition

$$T_\delta(\mathbb{1}_{d\times d}) = F_d^* \operatorname{diag}(F_d^{-1}[T_\delta(\mathbb{1}_{d\times d})_{(0)}])F_d = \frac{1}{\sqrt{d}}F_d^* \operatorname{diag}(dF_d^{-1}[T_\delta(\mathbb{1}_{d\times d})_{(0)}])\frac{1}{\sqrt{d}}F_d. \tag{3.66}$$

The eigenvalues $dF_d^{-1}[T_\delta(\mathbb{1}_{d\times d})_{(0)}]$ are given by

$$\begin{aligned}dF_d^{-1}[T_\delta(\mathbb{1}_{d\times d})_{(0)}]_k &= \sum_{j\in[d]} T_\delta(\mathbb{1}_{d\times d})_{0,j}e^{\frac{2\pi ijk}{d}} = \sum_{j\in[d]} \mathcal{I}_{|j|_c<\delta}e^{\frac{2\pi ijk}{d}} \\ &= \left[1 + \sum_{j=1}^{\delta-1}\left(e^{\frac{2\pi ijk}{d}} + e^{\frac{-2\pi ijk}{d}}\right)\right] = \left[1 + 2\sum_{j=1}^{\delta-1}\cos\left(\frac{2\pi jk}{d}\right)\right].\end{aligned}$$

For $k = 0$, we have $dF_d^{-1}[T_\delta(\mathbb{1}_{d\times d})_{(0)}]_k = 2\delta - 1$. For $k \neq 0$, we apply a trigonometric identity to transform the sum of cosines as

$$\begin{aligned}2\sum_{j=1}^{\delta-1}\cos\left(\frac{2\pi jk}{d}\right) &= \sum_{j=1}^{\delta-1} 2\cos\left(\frac{2\pi jk}{d}\right)\frac{\sin(\pi k/d)}{\sin(\pi k/d)} \\ &= \frac{1}{\sin(\pi k/d)}\sum_{j=1}^{\delta-1}\left[\sin\left(\frac{2\pi(j+1/2)k}{d}\right) - \sin\left(\frac{2\pi(j-1/2)k}{d}\right)\right] \\ &= \frac{\sin\left(\frac{2\pi(\delta-1+1/2)k}{d}\right) - \sin\left(\frac{2\pi(1-1/2)k}{d}\right)}{\sin(\pi k/d)} = \frac{\sin\left(\frac{\pi(2\delta-1)k}{d}\right)}{\sin(\pi k/d)} - 1.\end{aligned} \tag{3.67}$$

Then,

$$dF_d^{-1}[T_\delta(\mathbb{1}_{d\times d})_{(0)}]_k = \begin{cases} 2\delta-1, & k=0, \\ \frac{\sin\left(\frac{\pi(2\delta-1)k}{d}\right)}{\sin(\pi k/d)}, & k\neq 0. \end{cases} \tag{3.68}$$

By Proposition 2.2.1, the identity matrix can be rewritten as

$$(2\delta+1)I_d = \frac{2\delta+1}{d}F_d^*F_d = F_d^*\operatorname{diag}\left(\frac{2\delta+1}{d}\mathbb{1}_d\right)F_d,$$

where $\mathbb{1}_d \in \mathbb{C}^d$ is the vector with all entries equal to 1. Therefore, we combine the last equality with (3.65), (3.66) and (3.68) to obtain

$$B^*B = F_d^*\operatorname{diag}\left(\frac{2\delta+1}{d}\mathbb{1}_d\right)F_d + F_d^*\operatorname{diag}(F_d^{-1}[T_\delta(\mathbb{1}_{d\times d})_{(0)}])F_d = \frac{1}{\sqrt{d}}F_d^*\operatorname{diag}(u)\frac{1}{\sqrt{d}}F_d.$$

Since, $\frac{1}{\sqrt{d}}F_d^*$ and $\frac{1}{\sqrt{d}}F_d$ are unitary matrices, $\operatorname{diag}(u)$ contains the unordered eigenvalues of B^*B. Therefore, $u_0 = 4\delta \geq 4$ and for $k\neq 0$ via (3.67), we have

$$u_k = 2\delta+1+\frac{\sin\left(\frac{\pi(2\delta-1)k}{d}\right)}{\sin(\pi k/d)} = 2\delta+2+2\sum_{j=1}^{\delta-1}\cos\left(\frac{2\pi jk}{d}\right) \geq 2\delta+2-2(\delta-1) = 4.$$

□

Remark 3.6.20. *Note that if $d<2\delta-1$, then by the equation (3.58) we have*

$$\log|X_{\delta-1,0}| = \log|x_{\delta-1}| + \log|x_0|,$$
$$\log|X_{0,0-(d-\delta+1)}| = \log|X_{0,\delta-1}| = \log|x_{\delta-1}| + \log|x_0|,$$

which implies that the $(\delta-1,\delta-1)$ row and the $(d-\delta+1,0)$ row of B are equal. In fact, it is a consequence of the following equality for diagonals

$$\overline{S_j d^{d-j}(xx^*)} = \overline{S_j(x\circ S_{d-j}\overline{x})} = S_j\overline{x}\circ S_d x = S_j\overline{x}\circ x = x\circ S_j\overline{x} = d^j(xx^*), \quad j\in[d].$$

Recall that the inversion step only recovers the first δ diagonals of xx^* and X. Therefore, if $d\geq 2\delta-1$, for $j\in[\delta]$ the index $d-j$ satisfies

$$d-j \geq 2\delta-1-\delta+1 = \delta > \delta-1.$$

Consequently, $d^{d-j}(xx^*)$ would not be used for the construction of B and the duplication of the rows in B is avoided. On the contrary, for $d<2\delta-2$, the index $d-j$ is in $[\delta]$ for some $j\in[\delta]$, which implies that some rows of B will be equal. For $d=2\delta-2$, consider the diagonal with index $d-\delta+1=\delta-1$. Then, $d^{d-\delta+1}(xx^*) = d^{\delta-1}(xx^*)$ and the duplication of rows as above does not appear. Instead, the magnitudes within the diagonal $d^{\delta-1}(xx^*)$ duplicate,

$$d^{\delta-1}(xx^*)_k = (x\circ S_{\delta-1}\overline{x})_k = \overline{(S_{\delta-1}(S_{-\delta+1}\overline{x}\circ x)_k} = \overline{(S_{\delta-1}\overline{x}\circ x)_{k-\delta+1}} = \overline{d^{\delta-1}}_{k-\delta+1},$$

which gives the duplication of the rows in B.
While the duplication of the rows has no impact on the reconstruction with Log Magnitude Estimation, it would slightly change the analysis in Theorem 3.6.19, which is why we excluded the case $d<2\delta-1$ in this section.

Theorem 3.6.19 guarantees that B^*B is invertible and that $|x|$ can be recovered as $e^{(B^*B)^{-1}B^*b(X)}$. Consequently, in the presence of noise, Log Magnitude Estimation is given by

$$v = e^{(B^*B)^{-1}B^*b(Z)}, \tag{3.69}$$

where Z is the outcome of the inversion step. We require that $Z_{k,k-j} \neq 0$, $k \in [d], j \in [\delta]$ to ensure that the entries of $b(Z)$ are finite. If $Z_{k,k-j} = 0$, we can artificially increase its value to $Z_{k,k-j} = \varepsilon$ for some small parameter $\varepsilon > 0$, which would add a little noise to Z. Note that by (3.61), the vector $B^*b(Z)$ can be computed as

$$\begin{aligned}
(B^*b(Z))_\ell &= \sum_{k\in[d]} \sum_{j\in[\delta]} B_{(k,j),\ell} b(Z)_{(k,j)} \\
&= \sum_{k\in[d]} \left[\mathcal{I}_{k=\ell} b(Z)_{(k,0)} + \sum_{j\in[\delta]\setminus\{0\}} [\mathcal{I}_{k=\ell} + \mathcal{I}_{k-j=\ell}]\, b(Z)_{(k,j)} \right] \\
&= 2\log|Z_{\ell,\ell}| + \sum_{j\in[\delta]\setminus\{0\}} [\log|Z_{\ell,\ell-j}| + \log|Z_{\ell+j,\ell}|] \\
&= 2\log|Z_{\ell,\ell}| + \sum_{j\in[\delta]\setminus\{0\}} [\log|Z_{\ell,\ell-j}| + \log|Z_{\ell,\ell+j}|],
\end{aligned}$$

which requires $\mathcal{O}(d\delta)$ operations. Thanks to Theorem 3.6.19, the multiplication with $(B^*B)^{-1}$ can be computed using Fast Fourier transform, which has a computational complexity of $\mathcal{O}(d\log d)$. The entrywise computation of the exponent requires $\mathcal{O}(d)$ operations. Therefore, Log Magnitude Estimation requires $\mathcal{O}(d\log d + d\delta)$ operations in total. If $\delta = \mathcal{O}(\log d)$, then the computational complexity of Log Magnitude Estimation is $\mathcal{O}(d\log d)$, which is faster than $\mathcal{O}(d\log^2 d\log\log d)$ operations required for Block Magnitude Estimation.

Turning to the recovery guarantees for Log Magnitude Estimation, the following error bound holds.

Lemma 3.6.21. *Let $x \in \mathbb{C}^d$ be non-vanishing. Consider $Z \in \mathbb{T}_\delta$ such that $Z_{k,k-j} \neq 0$ for $k \in [d]$, $j \in [\delta]$ and X as in (3.31). For the vector v obtained via (3.69) the inequality*

$$\| |x| - v \|_2 \leq \sqrt{2}\, \|x\|_\infty \left[e^{\frac{1}{2}\|(b(Z)-b(X))\|_2} - 1 \right]$$

holds with $b(X)$ and $b(Z)$ given by (3.59).

Proof. We start by rewriting the vector v as a corrupted version of $|x|$. Using (3.62), we get

$$\begin{aligned}
v &= e^{(B^*B)^{-1}B^*b(Z)} = e^{(B^*B)^{-1}B^*b(X)+(B^*B)^{-1}B^*(b(Z)-b(X))} \\
&= |x| \circ e^{(B^*B)^{-1}B^*(b(Z)-b(X))} =: |x| \circ e^w.
\end{aligned}$$

Then, we substitute the obtained representation into the error $\| |x| - v \|_2$, which gives

$$\| |x| - v \|_2^2 = \| |x| - |x| \circ e^w \|_2^2 = \sum_{j\in[d]} ||x_j| - |x_j| e^{w_j}|^2 \leq \|x\|_\infty^2 \sum_{j\in[d]} |e^{w_j} - 1|^2. \tag{3.70}$$

Next, we work with the individual summands and apply the Taylor expansion of the exponential function,

$$|e^{w_j} - 1| = \left| \sum_{k=0}^{\infty} \frac{w_j^k}{k!} - 1 \right| = \left| \sum_{k=1}^{\infty} \frac{w_j^k}{k!} \right| \leq \sum_{k=1}^{\infty} \frac{|w_j|^k}{k!} = |w_j| + \sum_{k=2}^{\infty} \frac{|w_j|^k}{k!}.$$

The inequality $(\alpha + \beta)^2 \leq 2\alpha^2 + 2\beta^2$ leads to the bound

$$|e^{w_j} - 1|^2 \leq 2|w_j|^2 + 2 \left[\sum_{k=2}^{\infty} \frac{|w_j|^k}{k!} \right]^2,$$

and the sum over $j \in [d]$ is then bounded by

$$\begin{aligned} \sum_{j\in[d]} |e^{w_j} - 1|^2 &= 2 \sum_{j\in[d]} |w_j|^2 + 2 \sum_{j\in[d]} \left[\sum_{k=2}^{\infty} \frac{|w_j|^k}{k!} \right]^2 \leq 2 \|w\|_2^2 + 2 \left[\sum_{j\in[d]} \sum_{k=2}^{\infty} \frac{|w_j|^k}{k!} \right]^2 \\ &= 2 \|w\|_2^2 + 2 \left[\sum_{k=2}^{\infty} \frac{\|w\|_k^k}{k!} \right]^2 \leq 2 \|w\|_2^2 + 2 \left[\sum_{k=2}^{\infty} \frac{\|w\|_2^k}{k!} \right]^2. \end{aligned}$$

Note that in the first line we used the inequality

$$\sum_{j\in[d]} a_j^2 \leq \left[\sum_{j\in[d]} a_j \right]^2, \text{ for all } a \in \mathbb{R}^d, a_j \geq 0, j \in [d],$$

and in the second the monotonicity of ℓ_p-norms, $\|a\|_p \leq \|a\|_q$, $p \geq q$ was applied. The Taylor expansion of the exponential function yields

$$\begin{aligned} \sum_{j\in[d]} |e^{w_j} - 1|^2 &\leq 2 \|w\|_2^2 + 2 \left[\sum_{k=2}^{\infty} \frac{\|w\|_2^k}{k!} \right]^2 = 2 \|w\|_2^2 + 2 \left[\sum_{k=0}^{\infty} \frac{\|w\|_2^k}{k!} - \|w\|_2^2 - 1 \right]^2 \\ &= 2 \|w\|_2^2 + 2 \left[e^{\|w\|_2} - 1 - \|w\|_2 \right]^2 \end{aligned}$$

Then, using the inequality $(\alpha - \beta)^2 \leq \alpha^2 - \beta^2$ for $\alpha \geq \beta \geq 0$ with $\alpha = e^{\|w\|_2} - 1 = \sum_{k=1}^{\infty} \frac{\|w\|_2^k}{k!} + \|w\|_2$ and $\beta = \|w\|_2$ we get

$$\sum_{j\in[d]} |e^{w_j} - 1|^2 \leq 2 \left[e^{\|w\|_2} - 1 \right]^2.$$

Substituting this result into (3.70), we see that

$$\| |x| - v \|_2 \leq \sqrt{2} \|x\|_\infty \left[e^{\|w\|_2} - 1 \right].$$

We continue to bound $e^{\|w\|_2}$ using the definition of the vector w,

$$\|w\|_2 = \left\| (B^*B)^{-1} B^* (b(Z) - b(X)) \right\|_2 \leq \left\| (B^*B)^{-1} B^* \right\|_\infty \| (b(Z) - b(X)) \|_2.$$

Recall that $\|(B^*B)^{-1}B^*\|_\infty = \sqrt{\lambda_d^{-1}(B^*B)}$ and, by Theorem 3.6.19, $\lambda_d(B^*B) \geq 4$. Then, using the monotonicity of the exponential function we conclude that

$$\||x| - v\|_2 \leq \sqrt{2}\,\|x\|_\infty \left[e^{\|w\|_2} - 1\right] \leq \sqrt{2}\,\|x\|_\infty \left[e^{\frac{1}{2}\|(b(Z)-b(X))\|_2} - 1\right].$$

□

The resulting bound is different from those observed for the other two methods for magnitude estimation. The bound depicts noise in the form of $\|b(Z) - b(X)\|_2$, which decreases to 0 as $\|b(Z) - b(X)\|_2 \to 0$ or $Z \to X$. However, it is hard (yet not impossible) to compare $e^{\|b(Z)-b(X)\|_2}$ and $\|X - Z\|_F$ resulting from the inversion step.
In the form above, Log Magnitude Estimation always employs all entries of Z. However, we can customize which entries of Z are used in analogy to the choice of the partition $\{\mathcal{J}_p\}_{p\in[P]}$ for Block Magnitude Estimation. While arbitrary entries of Z can be discarded, we will only consider the scenario where a diagonal of Z is either completely excluded or fully used for the magnitude estimation. Let us assume that only diagonals $d^j(Z)$ with $j \in \mathcal{J} \subset [\delta]$ are used for the reconstruction process. Then, we adjust the construction of B and b as

$$B_{(k,j),\ell} := \begin{cases} 2, & j = 0, j \in \mathcal{J} \text{ and } k = \ell, \\ 1, & j \neq 0, j \in \mathcal{J}, \text{ and } k = \ell \\ 1, & j \neq 0, j \in \mathcal{J}, \text{ and } k - j = \ell \\ 0, & \text{otherwise,} \end{cases} \tag{3.71}$$

and

$$b(X)_{(k,j)} = \begin{cases} \log|X_{k,k-j}|, & j \in \mathcal{J}, \\ 0, & j \notin \mathcal{J}. \end{cases}$$

Then, similarly to Theorem 3.6.19, we show that the matrix B^*B is invertible for the adjusted construction of B.

Theorem 3.6.22. *Let $d \geq 2\delta - 1$. Consider the matrix B defined in (3.71) with $\mathcal{J} \subseteq [\delta]$. Then, the matrix B^*B admits the decomposition*

$$B^*B = \frac{1}{\sqrt{d}} F_d^* \operatorname{diag}(u) \frac{1}{\sqrt{d}} F_d,$$

with $u \in \mathbb{R}^d$ containing the eigenvalues

$$u_k = 2\mathcal{I}_{0\in\mathcal{J}} + 2|\mathcal{J}| + 2 \sum_{j\in\mathcal{J}\setminus\{0\}} \cos\left(\frac{2\pi jk}{d}\right).$$

If $0 \in \mathcal{J}$, then $u_k \geq 4$ for all $k \in [d]$. Otherwise, if $0 \notin \mathcal{J}$ and there are $j_1, j_2 \in \mathcal{J}$, $j_1 < j_2$ such that $j_2 - j_1$ is coprime with d (has the biggest common divisor 1), then u satisfies $u_k > 0$ for all $k \in [d]$. Consequently, the inverse matrix is given by

$$(B^*B)^{-1} = \frac{1}{\sqrt{d}} F_d^* \operatorname{diag}(1/u) \frac{1}{\sqrt{d}} F_d.$$

Proof. We only highlight the main differences from the proof of Theorem 3.6.19. Similarly to (3.63) and (3.64), the entries of B^*B are given by

$$(B^*B)_{\ell,\ell} = 4\mathcal{I}_{0\in\mathcal{J}} + \sum_{j\in\mathcal{J}\setminus\{0\}} 2 = 4\mathcal{I}_{0\in\mathcal{J}} + 2(|\mathcal{J}| - \mathcal{I}_{0\in\mathcal{J}}) = 2\mathcal{I}_{0\in\mathcal{J}} + 2|\mathcal{J}|, \qquad \ell \in [d],$$

$$(B^*B)_{\ell,s} = \mathcal{I}_{|s-\ell|_c\in\mathcal{J}}, \qquad \ell, s \in [d],\ \ell \neq s.$$

This computation requires the assumption $d \geq 2\delta - 1$. Since the entries $(B^*B)_{\ell,s}$ only depend on the difference $|\ell - s|_c$, the matrix B^*B is circulant and, thus, by Theorem 2.2.4 admits the decomposition

$$B^*B = F_d^* \operatorname{diag}(F_d^{-1}[(B^*B)_{(0)}])F_d = \frac{1}{\sqrt{d}}F_d^* \operatorname{diag}(dF_d^{-1}[(B^*B)_{(0)}])\frac{1}{\sqrt{d}}F_d$$

The eigenvalues $u = dF_d^{-1}[(B^*B)_{(0)}]$ are given by

$$\begin{aligned} u_k &= \sum_{j\in[d]} (B^*B)_{0,j} e^{\frac{2\pi ijk}{d}} = 2\mathcal{I}_{0\in\mathcal{J}} + 2|\mathcal{J}| + \sum_{j\in[d]\setminus\{0\}:|j|_c\in\mathcal{J}} e^{\frac{2\pi ijk}{d}} \\ &= 2\mathcal{I}_{0\in\mathcal{J}} + 2|\mathcal{J}| + 2\sum_{j\in\mathcal{J}\setminus\{0\}} \cos\left(\frac{2\pi jk}{d}\right). \end{aligned}$$

If $0 \in \mathcal{J}$, we can lower bound u_k as

$$u_k \geq 2 + 2|\mathcal{J}| + 2\sum_{j\in\mathcal{J}\setminus\{0\}} (-1) = 2 + 2|\mathcal{J}| - 2(|\mathcal{J}| - 1) = 4 > 0.$$

On the other hand, if $0 \notin \mathcal{J}$, u_k transforms to

$$u_k = 2|\mathcal{J}| + 2\sum_{j\in\mathcal{J}} \cos\left(\frac{2\pi jk}{d}\right) \geq 2|\mathcal{J}| - 2|\mathcal{J}| = 0,$$

with the equality in $\geq$ appearing if and only if all cosines are equal to -1. Let us prove by contradiction that $u_k > 0$. If $k = 0$, then all cosines are equal to one. Therefore, we only need to consider $k \in [d]\setminus\{0\}$. Recall the assumption that there are $j_1, j_2 \in \mathcal{J}$, $j_1 < j_2$ such that $j_2 - j_1$ is coprime with d. Since $\cos(2\pi j_p k/d) = -1$, $p = 1, 2$, the arguments satisfies

$$2\pi j_p k/d = \pi + 2\pi\theta_{k,p}, \text{ for some } \theta_{k,p} \in \mathbb{Z},\ p = 1, 2.$$

Taking the difference and simplifying the equation, we obtain

$$(j_2 - j_1)k/d = \theta_{k,2} - \theta_{k,1} \in \mathbb{Z}.$$

Since $j_2 - j_1$ is coprime with d, for all $k \in [d]\setminus\{0\}$ the integer $(j_2 - j_1)k$ is not divisible by d and, thus, the left-hand side is not an integer unless $j_2 = j_1$, which contradicts $j_1 \neq j_2$. Hence, $u_k > 0$. □

The condition on the existence of indices $j_1, j_2 \in \mathcal{J}$ such that $j_1 < j_2$ and $j_2 - j_1$ is coprime with d is satisfied if, for instance, $j_2 = j_1 + 1$ and two consequent diagonals are used for reconstruction.

At last, we note that taking $\mathcal{J} = \{0\}$ leads to Diagonal Magnitude Estimation and, thus, Log Magnitude Estimation is another possible generalization of Diagonal Magnitude Estimation along with Block Magnitude Estimation.

Notes and References. *Diagonal Magnitude Estimation was used in [35] with a slightly weaker version of Lemma 3.6.13. The constant factor in the bound as seen in Lemma 3.6.13 was later improved in [138].*
The original version of Block Magnitude Estimation was proposed in [40] and later analyzed in [138, 159]. There is a minor difference to Algorithm 6: the construction of the final estimate v in Step 4 used v^p instead of $|v^p|$, which may lead to negative entries of the vector v. While the use of $|v|$ instead of v would solve the issue with signs, we instead incorporated the absolute values $|v^p|$ in the summation step, which helps to avoid potential subtractions of the signed values. The second benefit of using $|v^p|$ is the possibility to include minor corrections to the original proof of Lemma 3.6.14, which contained an error in application of Lemma 2.1.1. We also note that our version of Lemma 3.6.14 improves the bound of [138, 159] by a factor of $\sqrt{\max_{k\in[d]}\mu_k/\min_{j\in[d]}\mu_j}$. Furthermore, based on monotonicity (3.57) the authors of [159] claim that $\{\mathcal{J}_p^\delta\}_{p\in[d]}$ is the optimal choice for Block Magnitude Estimation. However, Example 3.6.17 shows this family to be sub-optimal for some objects x.
Our main contribution is the Log Magnitude Estimation technique, which is an alternative to Block Magnitude Estimation. While the theoretical recovery guarantees for Log Magnitude Estimation are weaker, the numerical experiments in Section 6.1.2.3 suggest that the performance of both methods is somewhat similar and Log Magnitude Estimation is faster.
A preliminary version of Theorem 3.6.22 is a part of Master Thesis of Sarah Dörr jointly supervised with Benedikt Diederichs and Felix Krahmer.
Finally, we would like to elaborate on the polynomial dependency on $\|X-Z\|_F$ of the error bound provided by Lemma 3.6.21. Recall that by the definition of $b(Z)$ we have

$$|b(Z)_{(k,j)} - b(X)_{(k,j)}| = \left|\log\frac{|Z_{k,k-j}|}{|X_{k,k-j}|}\right| = \log\left(1+\frac{||Z_{k,k-j}|-|X_{k,k-j}||}{\min\{|X_{k,k-j}|,|Z_{k,k-j}|\}}\right).$$

Hence, the monotonicity of the norms grants $e^{\frac{1}{2}\|(b(Z)-b(X))\|_2} \le e^{\frac{1}{2}\|(b(Z)-b(X))\|_1}$ and the latter exponent can be transformed into a product. The application of the geometric-arithmetic mean inequality shows a polynomial dependency on $\|X-Z\|_F^p$ with p much larger than one.

3.6.4 Phase estimation

3.6.4.1 Phase estimation as phase synchronization problem

In this section, we discuss the estimation of the phases $\operatorname{sgn} x$ from the matrix $X = T_\delta(xx^*)$. As it was seen in Section 3.6.1, the main component for this step is the matrix,

$$\operatorname{sgn}_0(X_{k,\ell}) = \begin{cases} \operatorname{sgn} x_k \overline{\operatorname{sgn} x}_\ell, & X_{k,\ell} \neq 0, \\ 0, & \text{otherwise}, \end{cases}$$

which contains information about the phases. Note the phases $\operatorname{sgn} x$ are characterized by the angles $\operatorname{sgn} x_k = e^{i\varphi_k}$, $\varphi \in [0, 2\pi)^d$ and, thus, the non-zero entries of the matrix $\operatorname{sgn}_0(X_{k,\ell})$ are given by

$$\operatorname{sgn} x_k \overline{\operatorname{sgn} x}_\ell = e^{i\varphi_k} e^{-i\varphi_\ell} = e^{i(\varphi_k - \varphi_\ell)}.$$

It implies that the matrix $\operatorname{sgn}_0(X_{k,\ell})$ only contains information about the pairwise differences $\varphi_k - \varphi_\ell$ modulo 2π and, in practice, they are additionally corrupted by noise. Therefore, we can reformulate the recovery of the phases as the *angular or phase synchronization problem*:

Given a set of noisy pairwise differences $\varphi_k - \varphi_\ell \bmod 2\pi$ find $\varphi \in [0, 2\pi)^d$.

Let us denote by $E \subset [d]^2$ the set of pairs of indices for which the corresponding pairwise noisy differences are known,

$$E := \{(k,\ell) \in [d]^2 : \text{ noisy pairwise difference } \varphi_k - \varphi_\ell \bmod 2\pi \text{ is known and } k \neq \ell\}, \tag{3.72}$$

where the case $k = \ell$ is excluded, since it does not contain information about the phases. Then, the noiseless and noisy pairwise differences are represented by matrices

$$\Phi_{k,\ell} = \begin{cases} e^{i(\varphi_k - \varphi_\ell)}, & (k,\ell) \in E, \\ 0, & \text{otherwise}, \end{cases} \text{ and } \Psi_{k,\ell} = \begin{cases} e^{i(\varphi_k - \varphi_j + \eta_{k,\ell})}, & (k,\ell) \in E, \\ 0, & \text{otherwise}, \end{cases} \tag{3.73}$$

respectively. The entries $\eta_{k,\ell}$ represent noise and satisfy $\eta_{k,\ell} = -\eta_{\ell,k}$, so that both Φ and Ψ are Hermitian.
Note that for all vectors $\varphi \in [0, 2\pi)^d$ the equality

$$\varphi_k - \varphi_\ell = (\varphi_k - \theta) - (\varphi_\ell - \theta)$$

is satisfied for arbitrary $\theta \in [0, 2\pi)$ and, thus, a vector $\tilde{\varphi} \in [0, 2\pi)^d$ with $\tilde{\varphi}_k = \varphi_k - \theta$ will generate the same set of measurements as φ. Hence, the angular synchronization has a global phase ambiguity similarly to the phase retrieval problem and the notion of unique recovery is understood up to a global phase. Moreover, for a vector $u = e^{i\psi}$, $\psi \in [0, 2\pi)^d$ we measure the distance to the true solution by

$$\operatorname{dist}(e^{i\varphi}, e^{i\psi}) = \operatorname{dist}(\operatorname{sgn} x, u) = \min_{|\alpha|=1} \|\operatorname{sgn} x - \alpha u\|_2 .$$

While the angular representation φ is more convenient to introduce the angular synchronization problem, the sign notation $\operatorname{sgn} x$ is more suitable for formulation and analysis of recovery algorithms. Note that the phases $\operatorname{sgn} x$ satisfy $|\operatorname{sgn} x_k| = 1$ and, thus,

$$\operatorname{sgn} x \in \{u \in \mathbb{C}^d : |u_k| = 1 \text{ for all } k \in [d]\} = \mathbb{T}^d.$$

Hence, our goal would be to find $u \in \mathbb{T}^d$, which fits the measurements the best, i.e., it is a minimizer of the weighted Least Squares Problem (LSP)

$$\min_{u \in \mathbb{T}^d} \frac{1}{2} \sum_{(k,\ell) \in E} W_{k,\ell} |u_k - \Psi_{k,\ell} u_\ell|^2. \tag{LSP}$$

The weighs $W_{k,\ell}$ satisfy

$$W_{k,\ell} = 0, (k,\ell) \notin E, \; W_{k,\ell} \neq 0, (k,\ell) \in E, \; W_{k,\ell} = W_{\ell,k}, (k,\ell) \in E. \tag{3.74}$$

Selecting $W_{k,\ell} = 1, (k,\ell) \in E$ leads to an unweighted LSP. In the noiseless case, it corresponds to finding a vector u, which approximates the phase information the best,

$$\min_{u\in\mathbb{T}^d} \frac{1}{2} \sum_{(k,\ell)\in E} |u_k - \Psi_{k,\ell} u_\ell|^2 = \min_{u\in\mathbb{T}^d} \frac{1}{2} \sum_{(k,\ell)\in E} |u_k u_\ell^* - \Psi_{k,\ell}|^2 = \min_{u\in\mathbb{T}^d} \frac{1}{2} \|P_E(uu^*) - \Psi\|_F^2 ,$$

where P_E nullifies the entries, which are not in E. In the context of ptychography, it might be more beneficial to select the weights based on the amplitude information available in Z.

Example 3.6.23. *Consider Z resulting from the first step of Algorithm 3. Set $E = \{(k,\ell) \in [d]^2 : |Z_{k,\ell}| \neq 0, k \neq \ell\}$ and $\Psi = \operatorname{sgn}_0(Z)$. Let $v \in \mathbb{R}^d$ be the result of the magnitude estimation and assume that $|Z_{k,\ell}| = v_k v_\ell$ for $(k,\ell) \in E$. Then, selecting the weights as $W_{k,\ell} = |Z_{k,\ell}|^2 = v_k^2 v_\ell^2$ leads to the problem*

$$\begin{aligned}\min_{u\in\mathbb{T}^d} \frac{1}{2} \sum_{(k,\ell)\in E} W_{k,\ell} |u_k - \Psi_{k,\ell} u_\ell|^2 &= \min_{u\in\mathbb{T}^d} \frac{1}{2} \sum_{(k,\ell)\in E} |Z_{k,\ell}|^2 |u_k u_\ell^* - \operatorname{sgn}_0(Z_{k,\ell})|^2 \\ &= \min_{u\in\mathbb{T}^d} \frac{1}{2} \sum_{(k,\ell)\in E} |v_k u_k v_\ell u_\ell^* - |Z_{k,\ell}| \operatorname{sgn}_0(Z_{k,\ell})|^2 = \min_{u\in\mathbb{T}^d} \frac{1}{2} \|P_E[(v\circ u)(v\circ u)^*] - Z\|_F^2 ,\end{aligned}$$

which searches for the phases u such that the entries of the rank-one matrix $(v\circ u)(v\circ u)^*$ are close to the measured entries of Z.

The choice of the weights based on the amplitudes of the matrix Z is also motivated by the fact that for a small magnitude the phase is easier corrupted by noise than for a large magnitude. Therefore, we would likely "trust" the phase information corresponding to the large entries of Z more. Moreover, the amplitude-based weights indirectly imply that for the entries of $v \circ u$ with small magnitudes v, the incorrectly estimated phases u would have a minor impact on the total error $\operatorname{dist}(x, v\circ u)$ as the next example suggests.

Example 3.6.24. *Let v and u be estimates of magnitudes and phases of x, respectively. If for some $j \in [d]$ the estimated magnitude v_j satisfies $v_j \leq \varepsilon$ for a small ε, then its contribution to the estimation error*

$$\operatorname{dist}^2(x, v\circ u) = \min_{|\alpha|=1} \sum_{k\in[d]} ||x_k| \operatorname{sgn} x_k - \alpha v_k u_k|^2$$

is bounded by

$$\begin{aligned}||x_j| \operatorname{sgn} x_j - \alpha v_j u_j| &\leq ||x_j| \operatorname{sgn} x_j - v_j \operatorname{sgn} x_j| + |v_j \operatorname{sgn} x_j - \alpha v_j u_j| \\ &= ||x_j| - v_j| + v_j |\operatorname{sgn} x_j - \alpha u_j| \\ &\leq ||x_j| - v_j| + v_j(|\operatorname{sgn} x_j| + |\alpha u_j|) \leq ||x_j| - v_j| + 2\varepsilon.\end{aligned}$$

Hence, the contribution of u_j to the estimation error is small. On the other hand, since $|Z_{k,\ell}| \approx v_k v_\ell$ for $(k,\ell) \in E$, the corresponding weights $W_{k,j}$ are small and the error terms $W_{k,j}|u_k - \Psi_{k,j} u_j|^2$ do not contribute much to the sum of squares. Thus, the solution u of LSP tends to minimize the errors $W_{k,\ell}|u_k - \Psi_{k,\ell} u_\ell|^2$, $k,\ell \neq j$ corresponding to the entries with large magnitudes and ignores those errors corresponding to u_j.

The least squared loss in LSP is often explained by the maximum likelyhood estimation for Gaussian noise, which does not apply to the measurement (3.73). Instead we motivate this choice by a graph theoretic perspective. We will first introduce a few definitions, which will be relevant in this section.

We will consider a weighted undirected graph $G = (V, E, W)$ with three components: the *vertex set* V, the *set of edges* E and the *weights* W. In the context of angular synchronization, each node corresponds to a single entry of the phase vector $\operatorname{sgn} x$ and, thus, we set $V = [d]$. The edges are naturally identified with the observed noisy angular differences and E is the set introduced in (3.72). We note that the pairs (k, k) are excluded from E to avoid loops. The weights W are chosen to be the weights for LSP. Recall that by (3.74), the weights $W_{k,\ell}$ are positive if and only if the corresponding edge (k, ℓ) is present and since the graph is undirected, they are symmetric, $W_{k,\ell} = W_{\ell,k}$.

The *adjacency matrix* A_G of G is given by

$$(A_G)_{k,\ell} = \begin{cases} 1, & (k,\ell) \in E, \\ 0, & (k,\ell) \notin E. \end{cases}$$

With this notation, in view of (3.73), we can rewrite Φ as

$$\Phi = A_G \circ \operatorname{sgn} x \operatorname{sgn} x^* = \operatorname{diag}(\operatorname{sgn} x) A_G \operatorname{diag}(\operatorname{sgn} x)^*. \tag{3.75}$$

By construction, W satisfies $W = W \circ A_G$. Furthermore, if $W = A_G$, we speak of G as an *unweighted graph*.

The *degree of vertex* k is defined as

$$\deg(k) := \sum_{\ell:(k,\ell)\in E} W_{k,\ell},$$

and the corresponding *degree matrix* is the diagonal matrix

$$D = \operatorname{diag}\left(\{\deg(k)\}_{k\in[d]}\right).$$

The *Laplacian* of the graph G is given by

$$L_G = D - W.$$

As W is symmetric, the Laplacian is also symmetric. Moreover, using the symmetry of W and $W_{k,\ell} \geq 0$, we have

$$\begin{aligned} u^* L_G u = u^*(D-W)u &= \sum_{k\in[d]} \left[\deg(k)|u_k|^2 - \sum_{\ell:(k,\ell)\in E} u_k^* W_{k,\ell} u_\ell \right] \\ &= \sum_{k\in[d]} \left[|u_k|^2 \sum_{\ell:(k,\ell)\in E} W_{k,\ell} - \frac{1}{2} \sum_{\ell:(k,\ell)\in E} u_k^* W_{k,\ell} u_\ell - \frac{1}{2} \sum_{\ell:(k,\ell)\in E} u_\ell^* W_{\ell,k} u_k \right] \\ &= \frac{1}{2} \sum_{(k,\ell)\in E} W_{k,\ell}(2|u_k|^2 - 2\operatorname{Re}(u_k^* u_\ell)) = \frac{1}{2} \sum_{(k,\ell)\in E} W_{k,\ell}|u_k - u_\ell|^2 \geq 0, \end{aligned} \tag{3.76}$$

for all $u \in \mathbb{C}^d$. Hence, the Laplacian is positive semidefinite and, therefore, has a spectrum consisting of non-negative real numbers,

$$0 = \lambda_d(L_G) \leq \lambda_{d-1}(L_G) \leq \ldots \leq \lambda_1(L_G).$$

Here $\lambda_d(L_G) = 0$ follows from the observation that the vector $\mathbb{1}_d = (1, \ldots, 1)^T \in \mathbb{R}^d$ satisfies $|(\mathbb{1}_d)_\ell - (\mathbb{1}_d)_j| = 0$ and, thus, by (3.76), we have $\mathbb{1}_d^* L_G \mathbb{1}_d = 0$. The *spectral gap* of G is defined as $\tau_G := \lambda_{d-1}(L_G)$ and the graph G is connected if and only if $\tau_G > 0$ [167]. In that case, the kernel of L_G is spanned by $\mathbb{1}_d$.
Besides the Laplacian L_G the *normalized Laplacian* L_N of G is often used. It is defined as

$$L_N = D^{-1/2} L_G D^{-1/2}.$$

Its spectrum also consists of non-negative real numbers and we write τ_N for its second smallest eigenvalue $\lambda_{d-1}(L_N)$.
The *data dependent Laplacians* associated to Φ and Ψ are defined as

$$L_\Phi = D - W \circ \Phi, \quad \text{and} \quad L_\Psi = D - W \circ \Psi,$$

respectively. We note that the eigenvalues of L_Φ and L_G coincide. This is observed by multiplying L_Φ with the unitary diagonal matrix $C = \operatorname{diag}(\operatorname{sgn} x)$ on both sides. More precisely, using the commutativity of the diagonal matrices and the representation (3.75), we have

$$\begin{aligned} C^* L_\Phi C &= C^* D C - C^*(W \circ \Phi) C = D C^* C - C^*(W \circ (C A_G C^*)) C \\ &= D - W \circ (C^* C A_G C^* C) = D - W \circ A_G = D - W = L_G. \end{aligned}$$

In particular, it yields

$$\lambda_d(L_\Phi) = 0 \text{ with } L_\Phi \operatorname{sgn} x = 0 \text{ and } \lambda_{d-1}(L_\Phi) = \tau_G. \tag{3.77}$$

Similarly to (3.76), we can rewrite

$$u^* L_\Phi u = \frac{1}{2} \sum_{(k,\ell) \in E} W_{k,\ell} |u_k - \Phi_{k,\ell} u_\ell|^2 \quad \text{and} \quad u^* L_\Psi u = \frac{1}{2} \sum_{(k,\ell) \in E} W_{k,\ell} |u_k - \Psi_{k,\ell} u_\ell|^2,$$

which implies that both matrices are positive semidefinite.

3.6.4.2 Results for exact solution of phase synchronization

The last equation also allows for a compact representation of LSP,

$$\min_{u \in \mathbb{T}^d} \frac{1}{2} \sum_{(k,\ell) \in E} W_{k,\ell} |u_k - \Psi_{k,\ell} u_\ell|^2 = \min_{u \in \mathbb{T}^d} u^* L_\Psi u. \tag{3.78}$$

Using this form, we derive the first result connecting the graph properties to the existence of unique solution to the angular synchronization.

Theorem 3.6.25. *Consider a noiseless angular synchronization problem, so that $\Phi = \Psi$. The following statements hold true:*

1. *The optimization problem LSP admits unique solution* $\operatorname{sgn} x$ *(up to a global phase) if and only if* $\ker L_\Phi \cap \mathbb{T}^d = \{\alpha \operatorname{sgn} x : \alpha \in \mathbb{T}\}$.

2. *If* $\tau_G > 0$, *then* $\ker L_\Phi \cap \mathbb{T}^d = \{\alpha \operatorname{sgn} x : \alpha \in \mathbb{T}\}$.

Proof. 1. If $\ker L_\Phi \cap \mathbb{T}^d = \{\alpha \operatorname{sgn} x : \alpha \in \mathbb{T}\}$, then

$$\operatorname{sgn} x^* L_\Phi \operatorname{sgn} x = 0 \text{ and } u^* L_\Phi u > 0, \text{ for all } u \in \mathbb{T}^d \backslash \{\alpha \operatorname{sgn} x : \alpha \in \mathbb{T}\}.$$

Therefore, $\operatorname{sgn} x$ is the unique solution of LSP up to a global phase. On the contrary, let $\operatorname{sgn} x \in \mathbb{T}^d$ be the unique solution up to a global phase. By (3.77) we have $\operatorname{sgn} x \in \ker L_\Phi$ and $\{\alpha \operatorname{sgn} x : \alpha \in \mathbb{T}\} \subseteq \ker L_\Phi \cap \mathbb{T}^d$. Assume that there exists $u \in \ker L_\Phi \cap \mathbb{T}^d$ such that $\operatorname{dist}(\operatorname{sgn} x, u) > 0$. Then, $u^* L_\Phi u = 0$ and u is a solution of LSP different from $\operatorname{sgn} x$, which contradicts uniqueness. Thus, $\{\alpha \operatorname{sgn} x : \alpha \in \mathbb{T}\} = \ker L_\Phi \cap \mathbb{T}^d$.
2. By (3.77) we have $\lambda_d(L_\Phi) = 0$ with corresponding eigenvector $\operatorname{sgn} x$. If $\lambda_{d-1}(L_\Phi) = \tau_G > 0$, the kernel is spanned by $\operatorname{sgn} x$, that is $\ker L_\Phi = \operatorname{span}\{\operatorname{sgn} x\}$. Hence, $\ker L_\Phi \cap \mathbb{T}^d = \{\alpha \operatorname{sgn} x : \alpha \in \mathbb{T}\}$. □

As a consequence of Theorem 3.6.25, uniqueness is equivalent to the condition $\ker L_\Phi \cap \mathbb{T}^d = \{\alpha \operatorname{sgn} x : \alpha \in \mathbb{T}\}$, which depends on the noiseless data in the form of L_Φ. On the other hand, the second part of Theorem 3.6.25 tells us that condition $\tau_G > 0$ is sufficient for unique reconstruction. Moreover, $\tau_G > 0$ is equivalent to the graph G being connected, which is independent of both Φ and W.
In the presence of noise, the quality of reconstruction can be measured by $\operatorname{dist}(\operatorname{sgn} x, u)$ and various results are available. The first result derived specifically for unweighted graphs was derived in [138].

Theorem 3.6.26 ([138, Theorem 9])**.** *Let* Φ *and* Ψ *be defined as in* (3.73). *Suppose that* $G = (V, E)$ *is an undirected and unweighted graph with* $\tau_G > 0$. *Let* $u \in \mathbb{T}^d$ *be the minimizer of LSP. Then,*

$$\operatorname{dist}(\operatorname{sgn} x, u) \leq 2\tau_G^{-1/2} \|\Phi - \Psi\|_F .$$

This result is a special case of the next statement for weighted graphs, where we set $W = A_G$.

Theorem 3.6.27. *Let* Φ *and* Ψ *be defined as in* (3.73). *Suppose that* $G = (V, E, W)$ *is a weighted graph with* $\tau_G > 0$. *Let* $u \in \mathbb{T}^d$ *be the minimizer of LSP. Set* $R \in \mathbb{R}^{d \times d}$ *as* $R_{k,\ell} = W_{k,\ell}^{1/2}$. *Then,*

$$\operatorname{dist}(\operatorname{sgn} x, u) \leq 2\tau_G^{-1/2} \|R \circ (\Phi - \Psi)\|_F . \tag{3.79}$$

Proof. We will proceed by establishing the following three inequalities,

$$\operatorname{dist}^2(\operatorname{sgn} x, u) \leq 2\tau_G^{-1} u^* L_\Phi u, \tag{3.80}$$

$$2u^* L_\Phi u \leq 4u^* L_\Psi u + 4 \|u\|_\infty^2 \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x, \tag{3.81}$$

$$u^* L_\Psi u \leq \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x. \tag{3.82}$$

Then, the consecutive application of these inequalities with the observation that $\|u\|_\infty = 1$ and

$$\begin{aligned} 2\operatorname{sgn} x^* L_\Psi \operatorname{sgn} x &= \sum_{(k,\ell)\in E} W_{k,\ell} |\operatorname{sgn} x_k - \Psi_{k,\ell} \operatorname{sgn} x_\ell|^2 = \sum_{(k,\ell)\in E} W_{k,\ell} |\operatorname{sgn} x_k \overline{\operatorname{sgn} x_\ell} - \Psi_{k,\ell}|^2 \\ &= \sum_{(k,\ell)\in E} W_{k,\ell} |\Phi_{k,\ell} - \Psi_{k,\ell}|^2 = \|R \circ (\Phi - \Psi)\|_F^2 , \end{aligned} \tag{3.83}$$

provides the result of Theorem 3.6.27. The inequalities (3.80), (3.81) and (3.82) are the building blocks for other proofs in this section and we will later return to them.
It remains to prove the three inequalities. Since $\|\operatorname{sgn} x\|_2^2 = d$ and $\|u\|_2^2 = d$, we have that

$$\begin{aligned} \operatorname{dist}^2(\operatorname{sgn} x, u) &= \min_{|\alpha|=1} \|\operatorname{sgn} x - \alpha u\|_2^2 = \min_{|\alpha|=1} \|\operatorname{sgn} x\|_2^2 + \|\alpha u\|_2^2 - 2\operatorname{Re}(\alpha \operatorname{sgn} x^* u) \\ &= 2d - 2 \max_{|\alpha|=1} \operatorname{Re}(\alpha \operatorname{sgn} x^* u) . \end{aligned}$$

The term $\operatorname{Re}(\alpha \operatorname{sgn} x^* u)$ is maximal when $\operatorname{sgn} x^* u$ is aligned with the real axis, which leads to the optimal value $\alpha_o = \overline{\operatorname{sgn}(\operatorname{sgn} x^* u)}$. Hence, we obtain

$$\operatorname{Re}(\alpha_o \operatorname{sgn} x^* u) = \operatorname{Re}\left(\overline{\operatorname{sgn}(\operatorname{sgn} x^* u)} \operatorname{sgn}(\operatorname{sgn} x^* u) \cdot |\operatorname{sgn} x^* u|\right) = |\operatorname{sgn} x^* u|, \tag{3.84}$$

and

$$\operatorname{dist}^2(\operatorname{sgn} x, u) = 2d - 2\operatorname{Re}(\alpha_o \operatorname{sgn} x^* u) = 2d - 2|\operatorname{sgn} x^* u|. \tag{3.85}$$

The projection of $\alpha_o u$ onto the orthogonal complement of $\operatorname{sgn} x$ is given by

$$q := \alpha_o u - \left\langle \alpha_o u, \frac{\operatorname{sgn} x}{\|\operatorname{sgn} x\|_2} \right\rangle \frac{\operatorname{sgn} x}{\|\operatorname{sgn} x\|_2} = \alpha_o u - \frac{|\operatorname{sgn} x^* u|}{d} \operatorname{sgn} x,$$

where we used (3.84). Consequently, as $q \perp x$, the pythagorean theorem yields

$$\|q\|_2^2 = \|\alpha_o u\|_2^2 - \left\| \frac{|\operatorname{sgn} x^* u|}{d} \operatorname{sgn} x \right\|_2^2 = \|u\|_2^2 - \frac{|\operatorname{sgn} x^* u|^2}{d^2} \|\operatorname{sgn} x\|_2^2 = d - \frac{|\operatorname{sgn} x^* u|^2}{d}. \tag{3.86}$$

With the Cauchy-Schwarz inequality and (3.85), this leads to

$$\|q\|_2^2 = d - \frac{|\operatorname{sgn} x^* u|^2}{d} \geq d - |\operatorname{sgn} x^* u| = \frac{1}{2} \operatorname{dist}^2(\operatorname{sgn} x, u).$$

Recall that by (3.77) and the condition $\lambda_{d-1}(L_\Phi) = \tau_G > 0$, we have $\ker L_\Phi = \operatorname{span}\{\operatorname{sgn} x\}$. By the definition of q, it is orthogonal to the kernel of L_Φ, which implies that

$$q^* L_\Phi q = \left(\alpha_o u - \frac{|\operatorname{sgn} x^* u|}{d} \operatorname{sgn} x\right)^* L_\Phi \left(\alpha_o u - \frac{|\operatorname{sgn} x^* u|}{d} \operatorname{sgn} x\right) = u^* L_\Phi u,$$

and

$$u^* L_\Phi u = q^* L_\Phi q \geq \lambda_{d-1}(L_\Phi) \|q\|_2^2 \geq \frac{\tau_G}{2} \operatorname{dist}^2(\operatorname{sgn} x, u),$$

and the inequality (3.80) is proved.

Now we will prove the inequality (3.81). For ease of notation, we introduce the following auxiliary variables

$$g_k := \overline{\operatorname{sgn} x_k} u_k, \text{ and } \Lambda_{k,\ell} := \Phi^*_{k,\ell}\Psi_{k,\ell}.$$

Then, we rewrite $2u^* L_\Phi u$ as

$$\begin{aligned} 2u^* L_\Phi u &= \sum_{(k,\ell)\in E} W_{k,\ell}|u_k - \Phi_{k,\ell}u_\ell|^2 \\ &= \sum_{(k,\ell)\in E} W_{k,\ell}|u_k - \operatorname{sgn} x_k \overline{\operatorname{sgn} x_\ell} u_\ell|^2 = \sum_{(k,\ell)\in E} W_{k,\ell}|g_k - g_\ell|^2. \end{aligned}$$

The inequality $(\alpha+\beta)^2 \le 2\alpha^2 + 2\beta^2$ gives us

$$|g_k - g_\ell|^2 = |g_k - \Lambda_{k,\ell}g_\ell + \Lambda_{k,\ell}g_\ell - g_\ell|^2 \le 2|g_k - \Lambda_{k,\ell}g_\ell|^2 + 2|g_\ell|^2|\Lambda_{k,\ell} - 1|^2.$$

Hence, we obtain

$$2u^* L_\Phi u \le 2\sum_{(k,\ell)\in E} W_{k,\ell}|g_k - \Lambda_{k,\ell}g_\ell|^2 + 2\sum_{(k,\ell)\in E} W_{k,\ell}|g_k|^2|\Lambda_{k,\ell} - 1|^2. \tag{3.87}$$

For the first sum we observe that

$$\begin{aligned} \sum_{(k,\ell)\in E} W_{k,\ell}|g_k - \Lambda_{k,\ell}g_\ell|^2 &= \sum_{(k,\ell)\in E} W_{k,\ell}|\overline{\operatorname{sgn} x_k} u_\ell - \overline{\Phi}_{k,\ell}\Psi_{k,\ell}\overline{\operatorname{sgn} x_\ell} u_\ell|^2 \\ &= \sum_{(k,\ell)\in E} W_{k,\ell}|\overline{\operatorname{sgn} x_k} u_\ell - \overline{\operatorname{sgn} x_k}\operatorname{sgn} x_\ell \Psi_{k,\ell}\overline{\operatorname{sgn} x_\ell} u_\ell|^2 \\ &= \sum_{(k,\ell)\in E} W_{k,\ell}|u_k - \Psi_{k,\ell}u_\ell|^2 = 2u^* L_\Psi u. \end{aligned} \tag{3.88}$$

For the second sum, the equalities $|g_k| = |\overline{\operatorname{sgn} x_k} u_k| = |u_k|$ yield

$$\begin{aligned} \sum_{(k,\ell)\in E} W_{k,\ell}|g_k|^2|\Lambda_{k,\ell} - 1|^2 &\le \max_{k\in[d]}|g_k|^2 \sum_{(k,\ell)\in E} W_{k,\ell}|\Lambda_{k,\ell} - 1|^2 \\ &= \|u\|_\infty^2 \sum_{(k,\ell)\in E} W_{k,\ell}|\Lambda_{k,\ell} - 1|^2. \end{aligned}$$

The next step is to notice that

$$\begin{aligned} \sum_{(k,\ell)\in E} W_{k,\ell}|\Lambda_{k,\ell} - 1|^2 &= \sum_{(k,\ell)\in E} W_{k,\ell}|\overline{\Phi}_{k,\ell}\Psi_{k,\ell} - 1|^2 = \sum_{(k,\ell)\in E} W_{k,\ell}|\overline{\operatorname{sgn} x_k}\operatorname{sgn} x_\ell \Psi_{k,\ell} - 1|^2 \\ &= \sum_{(k,\ell)\in E} W_{k,\ell}|\operatorname{sgn} x_k - \Psi_{k,\ell}\operatorname{sgn} x_\ell|^2 = 2\operatorname{sgn} x^* L_\Psi \operatorname{sgn} x, \end{aligned}$$

and, thus,

$$\sum_{(k,\ell)\in E} W_{k,\ell}|g_k|^2|\Lambda_{k,\ell} - 1|^2 \le 2\|u\|_\infty^2 \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x.$$

Applying this bound and (3.88) to (3.87) gives us the desired inequality (3.81).
At last, for the inequality (3.82) we use the fact that u minimizes LSP, so that

$$u^* L_\Psi u \le \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x,$$

which concludes the proof. □

The term $\|R \circ (\Phi - \Psi)\|_F$ is presented in the form of the weighted difference of true and measured pairwise differences. However, it also has an alternative interpretation as a value of the empirical least squares loss evaluated at the vector $\operatorname{sgn} x$, namely

$$\frac{1}{2}\|R \circ (\Phi - \Psi)\|_F^2 = \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x,$$

which represents the gap between the value of the noise-free objective $\operatorname{sgn} x^* L_\Phi \operatorname{sgn} x = 0$ and the noisy objective $\operatorname{sgn} x^* L_\Psi \operatorname{sgn} x$ at the global minimum of the former one. In addition, we note that $\|R \circ (\Phi - \Psi)\|_F$ can be estimated from above as

$$\|R \circ (\Phi - \Psi)\|_F^2 \leq \|W \circ (\Phi - \Psi)\|_F \cdot \|\Phi - \Psi\|_F = \|L_\Phi - L_\Psi\|_F \cdot \|\Phi - \Psi\|_F , \tag{3.89}$$

which is a mixture of the distances between the Laplacians and between the phase differences, two metrics often observed in the error bounds for angular synchronization. For instance, the difference of the Laplacians can be observed in the alternative error bounds for weighted LSP.

Theorem 3.6.28 ([138, Proposition 12 and Theorem 8])**.** *Let Φ and Ψ be defined as in (3.73). Suppose that $G = (V, E, W)$ is a weighted graph with $\tau_G > 0$. Let $u \in \mathbb{T}^d$ be the minimizer of LSP. Then,*

$$\operatorname{dist}(\operatorname{sgn} x, u) \leq 2\sqrt{d\tau_G^{-1}\|W \circ (\Phi - \Psi)\|_\infty}, \tag{3.90}$$

and

$$\operatorname{dist}(\operatorname{sgn} x, u) \leq 4\sqrt{d}\tau_G^{-1}\|W \circ (\Phi - \Psi)\|_\infty . \tag{3.91}$$

Proof. By the inequality (3.80) it is only required to further bound $u^* L_\Phi u$. Using the fact that u minimizes LSP and that $\operatorname{sgn} x^* L_\Phi \operatorname{sgn} x = 0$, we have

$$\begin{aligned} u^* L_\Phi u &= u^*(L_\Phi - L_\Psi)u + u^* L_\Psi u \leq u^*(L_\Phi - L_\Psi)u + \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x \\ &= u^*(L_\Phi - L_\Psi)u - \operatorname{sgn} x^*(L_\Phi - L_\Psi)\operatorname{sgn} x. \end{aligned} \tag{3.92}$$

Then, applying the inequality

$$a^* M b \leq \|M\|_\infty \|a\|_2 \|b\|_2 , \text{ for all } M \in \mathbb{H}^d, a, b \in \mathbb{C}^d, \tag{3.93}$$

and observing that $\|u\|_2 = \|\operatorname{sgn} x\|_2 = \sqrt{d}$ we arrive at

$$u^* L_\Phi u \leq \|L_\Phi - L_\Psi\|_\infty (\|u\|_2^2 + \|\operatorname{sgn} x\|_2^2) = 2d\|L_\Phi - L_\Psi\|_\infty .$$

Hence, the inequality (3.80) yields

$$\operatorname{dist}^2(\operatorname{sgn} x, u) \leq 2\tau_G^{-1} u^* L_\Phi u \leq 4d\tau_G^{-1}\|L_\Phi - L_\Psi\|_\infty ,$$

which concludes the proof of the inequality (3.90). For (3.91), let α_o be the minimizer of

$$\operatorname{dist}(\operatorname{sgn} x, u) = \min_{|\alpha|=1} \|\operatorname{sgn} x - \alpha u\|_2 .$$

Then, we further rewrite (3.92) as

$$\begin{aligned} u^* L_\Phi u &\leq (\alpha_o u)^* (L_\Phi - L_\Psi) \alpha_o u - \operatorname{sgn} x^* (\Phi - L_\Psi) \operatorname{sgn} x \pm (\alpha_o u)^* (L_\Phi - L_\Psi) \operatorname{sgn} x \\ &= (\alpha_o u)^* (L_\Phi - L_\Psi)(\alpha_o u - \operatorname{sgn} x) + (\alpha_o u - \operatorname{sgn} x)^* (L_\Phi - L_\Psi) \operatorname{sgn} x. \end{aligned}$$

Therefore, by (3.93) and the fact that $\|u\|_2 = \|\operatorname{sgn} x\|_2 = \sqrt{d}$ we have

$$u^* L_\Phi u \leq \operatorname{dist}(\operatorname{sgn} x, u) \|L_\Phi - L_\Psi\|_\infty (\|\alpha_o u\|_2 + \|\operatorname{sgn} x\|_2) = 2\sqrt{d} \operatorname{dist}(\operatorname{sgn} x, u) \|L_\Phi - L_\Psi\|_\infty .$$

Combining the obtained inequality with (3.80) gives us

$$\operatorname{dist}^2(\operatorname{sgn} x, u) \leq 2\tau_G^{-1} u^* L_\Phi u \leq 4\sqrt{d}\tau_G^{-1} \operatorname{dist}(\operatorname{sgn} x, u) \|L_\Phi - L_\Psi\|_\infty ,$$

which is equivalent to

$$\operatorname{dist}(\operatorname{sgn} x, u) \leq 4\sqrt{d}\tau_G^{-1} \|L_\Phi - L_\Psi\|_\infty .$$

□

As the square root in (3.90) leads to a slow convergence as noise diminishes, i.e., Ψ approaches Φ, in many cases the bound (3.91) is stronger than (3.90). Our numerical trials in [164] for unweighted graphs suggest that the bound (3.91) behaves similarly to that of Theorem 3.6.26 and for weighted graphs, Theorem 3.6.27 provides empirically stronger bounds.

3.6.4.3 Eigenvector relaxation of phase synchronization

Due to the quadratic constraint $u \in \mathbb{T}^d$, LSP is a non-convex quadratic minimization problem with quadratic constraints and, thus, NP-hard in general. One way to obtain a computationally feasible problem is to relax the constraint $u \in \mathbb{T}^d$ to $\|u\|_2^2 = d$, which leads to

$$\min_{\|u\|_2^2 = d} u^* L_\Psi u. \tag{EIG}$$

This is nothing else but the determination of the smallest eigenvalue of the matrix L_Ψ and can be solved efficiently. Since the minimizer u is not necessarily in $\mathbb{T}^d$, it is further projected on $\mathbb{T}^d$ as $\operatorname{sgn} u$. We will refer to this optimization problem as eigenvector relaxation (EIG).
The first important result regarding the recovery with EIG derives a sufficient condition on the graph for a successful recovery of the phases in the noiseless case.

Theorem 3.6.29. *Consider a noiseless angular synchronization problem, so that $\Phi = \Psi$. The following statements hold true:*

1. *The optimization problem EIG admits the unique solution $\operatorname{sgn} x$ (up to a global phase) if and only if $\tau_G > 0$.*

2. *If $\operatorname{sgn} x$ is the unique solution of EIG, then $\operatorname{sgn} x$ is the unique solution of LSP.*

Proof. 1. Let us recall that by (3.77) we have $\operatorname{sgn} x \in \ker L_\Phi$ and $\lambda_d(L_\Phi) = 0$. If $\lambda_{d-1}(L_\Phi) = \tau_G > 0$, the kernel is spanned by $\operatorname{sgn} x$, i.e., $\ker L_\Phi = \operatorname{span}\{\operatorname{sgn} x\}$. Hence, for any $u \in \mathbb{C}^d$ with $\|u\|_2^2 = d$ and $u \perp \operatorname{sgn} x$ it holds that $u^* L_\Phi u > 0$ and u is not the minimizer of EIG. Consequently, $\operatorname{sgn} x$ is the unique solution of EIG up to a global phase. On the other hand, let $\operatorname{sgn} x$ be the unique solution of EIG and assume that $\tau_G = 0$. Then there exists $u \perp \operatorname{sgn} x$ such that $u^* L_\Phi u = 0$, which contradicts the uniqueness of the solution of EIG.
2. It is a consequence of 1. and Theorem 3.6.25. □

Unlike LSP, EIG requires the slightly stronger condition $\tau_G > 0$ to ensure the uniqueness of reconstruction as a trade-off for the reduced hardness of the problem.
If noise is present, analogues of Theorems 3.6.26 and 3.6.28 are applicable.

Theorem 3.6.30. *Let Φ and Ψ be defined as in (3.73). Suppose that $G = (V, E, W)$ is a weighted graph with $\tau_G > 0$. Set $R \in \mathbb{R}^{d\times d}$ as $R_{k,\ell} = W_{k,\ell}^{1/2}$. Let $u \in \mathbb{T}^d$ be the minimizer of EIG. Then,*

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn} u) \le 2\tau_G^{-1/2}\sqrt{2 + 2\|u\|_\infty^2}\,\|R \circ (\Phi - \Psi)\|_F\,, \tag{3.94}$$

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn} u) \le 4\sqrt{d\tau_G^{-1}\,\|W \circ (\Phi - \Psi)\|_\infty}, \tag{3.95}$$

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn} u) \le 8\sqrt{d}\tau_G^{-1}\,\|W \circ (\Phi - \Psi)\|_\infty\,. \tag{3.96}$$

Proof. The proof of Theorem 3.6.30 slightly expands on the proofs of Theorems 3.6.26 and 3.6.28. The main addition is the inequality

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn} u) \le 2\operatorname{dist}(\operatorname{sgn} x, u), \tag{3.97}$$

which is a consequence of

$$|\operatorname{sgn}\alpha - \beta| \le 2|\alpha - \beta| \text{ for all } \alpha, \beta \in \mathbb{C},\ |\beta| = 1. \tag{3.98}$$

More precisely, we have

$$\begin{aligned}\operatorname{dist}^2(\operatorname{sgn} x, \operatorname{sgn} u) &= \min_{|\alpha|=1} \sum_{k\in[d]} |\operatorname{sgn} x_k - \alpha \operatorname{sgn} u_k|^2 \\ &\le 4 \min_{|\alpha|=1} \sum_{k\in[d]} |\operatorname{sgn} x_k - \alpha u_k|^2 = 4\operatorname{dist}^2(\operatorname{sgn} x, u).\end{aligned}$$

Since proofs of the inequalities (3.80) and (3.81) only use that $\|u\|_2^2 = d$, they hold true for the solution of EIG. The inequality (3.82) follows directly from the fact that u is the minimizer of EIG. Therefore, by combining inequalities (3.97), (3.80), (3.81), (3.82) and (3.83) we obtain

$$\begin{aligned}\operatorname{dist}^2(\operatorname{sgn} x, \operatorname{sgn} u) \le 4\operatorname{dist}^2(\operatorname{sgn} x, u) &\le \frac{8u^* L_\Phi u}{\tau_G} \le \frac{16u^* L_\Psi u + 16\|u\|_\infty^2 \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x}{\tau_G} \\ &\le \frac{8(2 + 2\|u\|_\infty^2)\operatorname{sgn} x^* L_\Psi \operatorname{sgn} x}{\tau_G} = \frac{4(2 + 2\|u\|_\infty^2)\,\|R \circ (\Phi - \Psi)\|_F^2}{\tau_G}.\end{aligned}$$

For the bounds (3.95) and (3.96), we use (3.97) and then repeat the steps of the proof of Theorem 3.6.28 using that $\|u\|_2^2 = d$ and that u is the minimizer of EIG. □

While the bounds (3.95) and (3.96) are slightly weaker counterparts of Theorem 3.6.28, the bound (3.94) also includes the tightness penalty $\sqrt{2+2\|u\|_\infty^2}$, which varies between 2 if the relaxation is tight, that is $u \in \mathbb{T}^d$, and $\sqrt{2+2d}$ in the worst case. In our numerical experiments in [164], however, this difference is not observed, which suggests that this dimensional factor may be a proof artifact.
Another error bound for the eigenvector-based reconstruction was given in [40] for the case of unweighted graphs. Their proof is based on the Cheeger inequality that is only available for the normalized Laplacian [168], which is why the minimization problem in their theorem has a different normalization than EIG. In the special case that $\deg(k)$ is a constant for all $k \in [d]$ (as in [40]), the two normalizations agree up to a constant. Using the terminology introduced above their result reads as follows.

Theorem 3.6.31 ([40, Theorem 3], [138, Theorem 4])**.** *Let Φ and Ψ be defined as in (3.73). Suppose that $G = (V, E)$ is an undirected connected and unweighted graph with $\tau_N > 0$. Let $u \in \mathbb{C}^d$ be the minimizer of*

$$\min_{\|u\|_2^2 = d} u^* D^{-1/2} L_\Psi D^{-1/2} u.$$

Then,

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn} u) \leq 19 \frac{\|\Phi - \Psi\|_F}{\tau_N \sqrt{\min_{k\in[d]} \deg(k)}}.$$

This result has a constant factor instead of the $\|u\|_\infty$-depending penalty, which potentially makes it a stronger bound. We also refer the reader to Section 4.3.2 of [138] for the comparison of τ_G and $\tau_N \min_{k\in[d]} \deg(k)$.

3.6.4.4 Semidefinite relaxation of phase synchronization

An alternative approach to EIG is based on the idea of lifting the problem to the matrix space. It makes use of the relation

$$u^* L_\Psi u = \sum_{k,\ell \in E} u_k^* \, (L_\Psi)_{k,\ell} \, u_\ell = \operatorname{tr}(L_\Psi u u^*).$$

With this, LSP transforms into

$$\begin{aligned} &\min_{U \in \mathbb{H}^d} \operatorname{tr}(L_\Psi U) \\ &s.t.\ U_{k,k} = 1,\ U \succeq 0,\ \operatorname{rank}(U) = 1. \end{aligned} \tag{3.99}$$

The class of minimization problems with explicit rank constraints is known to include many NP-hard instances [169, Chapter 2], so a common strategy is to perform a semidefinite relaxation. For (3.99), the following semidefinite relaxation has been proposed in [170],

$$\begin{aligned} &\min_{U \in \mathbb{H}^d} \operatorname{tr}(L_\Psi U) \\ &s.t.\ U_{k,k} = 1,\ U \succeq 0. \end{aligned} \tag{SDP}$$

The solution U of SDP admits an eigenvalue decomposition

$$U = \sum_{j=1}^{\operatorname{rank}(U)} \lambda_j(U) u_j u_j^*. \tag{3.100}$$

If U meets the rank condition in (3.99) one obtains that $U = du_1u_1^*$, where $\sqrt{d}u_1$ is a solution of LSP. Without the rank condition, however, the solution of SDP may have higher rank. In this case, the phase factors $\operatorname{sgn}(u_1)$ corresponding to the eigenvector u_1 associated with the largest eigenvalue are used as an approximation of the LSP solution [171]. We note that SDP seeks for a $d \times d$ matrix, which constitutes $\mathcal{O}(d^2)$ number of unknowns.
Just like for LSP and EIG, we are able to establish a sufficient condition for a unique reconstruction via SDP in the noiseless case.

Theorem 3.6.32. *Consider a noiseless angular synchronization problem, so that $\Phi = \Psi$. The following statements hold true:*

1. *If $\tau_G > 0$, the optimization problem SDP admits the unique solution $\operatorname{sgn} x \operatorname{sgn} x^*$.*
2. *If $\operatorname{sgn} x \operatorname{sgn} x^*$ is the unique solution of SDP, then $\operatorname{sgn} x$ is the unique solution of LSP (up to a global phase).*

Proof. 1. By (3.77), $\operatorname{sgn} x$ is the minimizer of LSP with the least squares loss being zero. In view of $L_\Phi \succeq 0$ and inequality

$$\operatorname{tr}(L_\Phi U) \geq 0, \quad \text{for all } U \in \mathbb{H}^d,$$

the matrix $\operatorname{sgn} x \operatorname{sgn} x^*$ is the minimizer of SDP with the objective being zero. Hence, every minimizer U of SDP satisfies $\operatorname{tr}(L_\Phi U) = 0$. Let us consider the eigendecomposition of U given by (3.100). Substituting U into the objective of SDP, we obtain

$$0 = \operatorname{tr}\left(L_\Phi \sum_{j=1}^{\operatorname{rank}(U)} \lambda_j(U) u_j u_j^* \right) = \sum_{j=1}^{\operatorname{rank}(U)} \lambda_j(U) \operatorname{tr}(L_\Phi u_j u_j^*) = \sum_{j=1}^{\operatorname{rank}(U)} \lambda_j(U) u_j^* L_\Phi u_j.$$

Since $\lambda_j(U) > 0$ for all $j = 1, \ldots, \operatorname{rank}(U)$, the above equality is only possible if $u_j \in \ker(L_\Phi)$ for all $j = 1, \ldots, \operatorname{rank}(U)$. Recall that u_j are orthogonal and that under the assumption $\tau_G > 0$, the kernel is given by $\operatorname{span}\{\operatorname{sgn} x\}$. This implies that U has rank at most 1. In fact, the zero matrix is not feasible. Thus, $\operatorname{rank}(U) = 1$ and $u_1 = \alpha \operatorname{sgn} x$ for some $\alpha \in \mathbb{C}$. Furthermore, u_1 satisfies $\|u_1\|_2 = 1$, which implies that $u_1 = \alpha \operatorname{sgn} x / \sqrt{d}$ for some $\alpha \in \mathbb{T}$. Turning to the eigenvalues, the constraint $U_{k,k} = 1$ in SDP yields

$$1 = U_{k,k} = \lambda_1(U) |(u_1)_k|^2 = \lambda_1(U)/d,$$

and, therefore, $U = \lambda_1(U) u_1 u_1^* = \operatorname{sgn} x \operatorname{sgn} x^*$.
2. Assume that $\operatorname{sgn} x$ is not the unique minimizer of LSP. Then, there exists $u \in \ker L_\Phi \cap \mathbb{T}^d$ such that $\operatorname{dist}(\operatorname{sgn} x, u) > 0$. For a matrix $uu^* \succeq 0$ we have $\operatorname{tr}(L_\Phi uu^*) = 0$ and $(uu^*)_{k,k} = |u_k|^2 = 1$. Therefore, uu^* is a solution of SDP different from $\operatorname{sgn} x \operatorname{sgn} x^*$, which contradicts the uniqueness of the solution. □

If noise is present, the reconstruction error bounds for SDP are not derived and, instead, the bounds for LSP are used if the relaxation is tight, i.e., U has rank one. A sufficient condition to guarantee the tightness of relaxation is given by the next lemma.

Lemma 3.6.33 ([138, Lemma 16]). *Suppose that $u \in \mathbb{T}^d$ is the minimizer of LSP and let $L_{uu^*} = D - W \circ uu^*$. If*

$$\|L_\Psi - L_{uu^*}\|_F < \frac{\tau_G}{1+\sqrt{d}},$$

then uu^* is the minimizer of SDP.

As the spectral gap τ_G is typically rather small, as compared to the dimension d, the tightness is guaranteed only for very small noise levels. In fact, our numerical simulations in [164] show that the SDP relaxation is indeed not tight in many cases. Hence, we propose error bounds for SDP, which hold true regardless of the tightness of relaxation.

Theorem 3.6.34. *Let Φ and Ψ be defined as in (3.73). Suppose that $G = (V, E, W)$ is a weighted graph with $\tau_G > 0$. Set $R \in \mathbb{R}^{d\times d}$ as $R_{k,\ell} = W_{k,\ell}^{1/2}$. Let $U \in \mathbb{H}^d$ be the minimizer of SDP and set $u_1 \in \mathbb{C}^d$ be the eigenvector corresponding to the largest magnitude eigenvalue of U. Then,*

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn}(u_1)) \le 4\tau_G^{-1/2}\sqrt{\operatorname{erank}(U)}\,\|R \circ (\Phi - \Psi)\|_F\,, \tag{3.101}$$

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn}(u_1)) \le \sqrt{8d(1+\operatorname{erank}(U))\tau_G^{-1}\,\|W \circ (\Phi - \Psi)\|_\infty}, \tag{3.102}$$

with effective rank $\operatorname{erank}(U) := \|U\|_1 / \|U\|_\infty$.

Proof. The proof of Theorem 3.6.34 resembles the proof of Theorem 3.6.30 with few adaptations. Similarly, we use the inequality (3.98) to transit from $\operatorname{sgn}(u_1)$ to $\sqrt{d}u_1$,

$$\operatorname{dist}^2(\operatorname{sgn} x, \operatorname{sgn}(u_1)) = \operatorname{dist}^2\left(\operatorname{sgn} x, \operatorname{sgn}\left(\sqrt{d}u_1\right)\right) \le 4\operatorname{dist}^2\left(\operatorname{sgn} x, \sqrt{d}u_1\right)$$

Since $\left\|\sqrt{d}u_1\right\|_2^2 = d$, we can apply the inequality (3.80), so that

$$\operatorname{dist}^2(\operatorname{sgn} x, \operatorname{sgn}(u_1)) \le 4\operatorname{dist}\left(\operatorname{sgn} x, \sqrt{d}u_1\right) \le 8\tau_G^{-1}(\sqrt{d}u_1)^* L_\Phi(\sqrt{d}u_1). \tag{3.103}$$

For the error bound (3.101), we use (3.81) and obtain

$$\operatorname{dist}^2(\operatorname{sgn} x, \operatorname{sgn} u_1) \le \tau_G^{-1}\left[16(\sqrt{d}u_1)^* L_\Psi(\sqrt{d}u_1) + 16\left\|\sqrt{d}u_1\right\|_\infty^2 \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x\right]. \tag{3.104}$$

Using that L_Ψ is positive semidefinite, the first summand is further bounded by

$$du_1^* L_\Psi u_1 = \frac{d}{\lambda_1(U)}\lambda_1(U)u_1^* L_\Psi u_1 \le \frac{d}{\lambda_1(U)}\sum_{j=1}^{\operatorname{rank}(U)}\lambda_j(U)u_j^* L_\Psi u_j = \frac{d}{\lambda_1(U)}\langle L_\Psi, U\rangle_F.$$

Since U is the minimizer of SDP, we have

$$du_1^* L_\Psi u_1 \le \frac{d}{\lambda_1(U)}\langle L_\Psi, U\rangle_F \le \frac{d}{\lambda_1(U)}\langle L_\Psi, \operatorname{sgn} x \operatorname{sgn} x^*\rangle_F = \frac{d}{\lambda_1(U)}\operatorname{sgn} x^* L_\Psi \operatorname{sgn} x. \tag{3.105}$$

For the second summand, the norm $\left\|\sqrt{d}u_1\right\|_\infty^2$ can be bounded using the condition $U_{k,k} = 1$ as

$$1 = U_{k,k} = \left[\sum_{j=1}^{\operatorname{rank}(U)} \lambda_j(U) u_j u_j^*\right]_{k,k} = \sum_{j=1}^{\operatorname{rank}(U)} \lambda_j(U)|(u_j)_k|^2 \geq \lambda_1(U)|(u_1)_k|^2,$$

and, thus,

$$\left\|\sqrt{d}u_1\right\|_\infty^2 = d \max_{k\in[d]} |(u_1)_k|^2 \leq \frac{d}{\lambda_1(U)}.$$

Substituting this bound and (3.105) into (3.104), we obtain

$$\operatorname{dist}^2(\operatorname{sgn} x, \operatorname{sgn} u_1) \leq 32\tau_G^{-1} \frac{d}{\lambda_1(U)} \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x.$$

The proof of the inequality (3.101) is concluded, by observing that

$$d = \sum_{k\in[d]} U_{k,k} = \operatorname{tr}(U) = \|U\|_1,$$

so that

$$d/\lambda_1(U) = \|U\|_1 / \|U\|_\infty = \operatorname{erank}(U),$$

and applying (3.83),

$$\operatorname{dist}^2(\operatorname{sgn} x, \operatorname{sgn} u_1) \leq 16\tau_G^{-1} \operatorname{erank}(U) \|R \circ (\Phi - \Psi)\|_F^2.$$

Turning to the inequality (3.102), we start with (3.103) and transform the term $u_1^* L_\Phi u_1$ analogously to the inequality (3.92) in the proof of Theorem 3.6.28. In view of (3.105), (3.77) and (3.93), we get

$$\begin{aligned} u_1^* L_\Phi u_1 &= u_1^*(L_\Phi - L_\Psi)u_1 + u_1^* L_\Psi u_1 \leq u_1^*(L_\Phi - L_\Psi)u_1 + \frac{1}{\lambda_1(U)} \operatorname{sgn} x^* L_\Psi \operatorname{sgn} x \\ &\leq u_1^*(L_\Phi - L_\Psi)u_1 - \frac{1}{\lambda_1(U)} \operatorname{sgn} x^*(L_\Phi - L_\Psi) \operatorname{sgn} x \\ &\leq (\|u_1\|_2^2 + \frac{1}{\lambda_1(U)} \|\operatorname{sgn} x\|_2^2) \|L_\Phi - L_\Psi\|_\infty = (1 + \operatorname{erank}(U)) \|L_\Phi - L_\Psi\|_\infty. \end{aligned}$$

An application of the obtained bound to (3.103) concludes the proof. □

We note that both bounds in Theorem 3.6.34 include a tightness penalty in the form of $\operatorname{erank}(U)$. If the relaxation is tight, $U = uu^*$ and $\operatorname{erank}(U) = 1$. On the other hand, in the worst case $\operatorname{erank}(U)$ can be as big as d. Just as the tightness penalty for EIG, it may be a proof artifact, since in numerical trials in [164] the actual error does not show the dependency on the effective rank.

3.6.4.5 Results for phase synchronization in context of Block Phase Retrieval

Returning to ptychography, we would like to employ the angular synchronization for phase estimation.
We construct the edge set E by considering the non-zero entries of the matrix Z reconstructed in the inversion step,

$$E = \{(k,\ell) \in [d]^2 : \ |Z_{k,\ell}| > 0 \text{ and } k \neq \ell\}. \tag{3.106}$$

Recall that $Z \in \mathbb{T}_\delta$ and, thus, E is determined by the non-zero elements of the first δ diagonals $d^j(Z)$, $j \in [\delta]\backslash\{0\}$. If all of them are non-vanishing, we can estimate the spectral gap from below.

Lemma 3.6.35 (Version of [40, Lemma 2] or [138, Lemma 2])**.** *Let $d \geq 4\delta$ and $\delta \geq 3$. Assume that $|d^j(Z)_k| > 0$, $j \in [\delta]\backslash\{0\}$, $k \in [d]$. Consider an unweighted graph with the edge set E as in (3.106). Then, the graph is connected and*

$$\tau_G \geq \frac{\pi^2\delta^3}{3d^2}.$$

Proof. Under the assumptions of the lemma, the adjacency matrix A_G corresponding to the edge set E in (3.106) is given by

$$A_G = T_\delta(\mathbb{1}_{d\times d}) - I_d,$$

where $\mathbb{1}_{d\times d}$ is a matrix with all entries equal to 1 and T_δ is the projection onto the space $\mathbb{T}_\delta$ given by (3.30). Each node in the graph has degree $2\delta - 2$ and, thus, by the definition of the graph Laplacian we have

$$L_G = (2\delta-2)I_d - A_G = (2\delta-1)I_d - T_\delta(\mathbb{1}_{d\times d}).$$

Recall that the eigendecomposition of $T_\delta(\mathbb{1}_{d\times d})$ was derived in (3.66) and its eigenvalues are given by the equation (3.68). Moreover, in view of (3.67), we have that for eigenvalues other than $2\delta - 1$, it holds that

$$\frac{\sin\left(\frac{\pi(2\delta-1)k}{d}\right)}{\sin(\pi k/d)} = 2\sum_{j=1}^{\delta-1}\cos\left(\frac{2\pi jk}{d}\right) + 1 \leq 2(\delta-1)+1 = 2\delta-1, \tag{3.107}$$

so that $2\delta - 1$ is the largest eigenvalue of $T_\delta(\mathbb{1}_{d\times d})$. Hence, for the spectral gap it holds that

$$\begin{aligned}\tau_G &= \lambda_{d-1}(L_G) - \lambda_d(L_G) = \lambda_{d-1}((2\delta-1)I_d - T_\delta(\mathbb{1}_{d\times d})) - 0 = 2\delta - 1 - \lambda_2(T_\delta(\mathbb{1}_{d\times d}))\\ &= \lambda_1(T_\delta(\mathbb{1}_{d\times d})) - \lambda_2(T_\delta(\mathbb{1}_{d\times d})).\end{aligned}$$

In order to obtain the desired estimate from below for τ_G, the difference $\lambda_1(T_\delta(\mathbb{1}_{d\times d})) - \lambda_2(T_\delta(\mathbb{1}_{d\times d}))$ is analyzed by repeating the proof of [40, Lemma 2] or [138, Lemma 2]. □

Remark 3.6.36. *If the assumption of Lemma 3.6.35 does not hold, for the vanishing entries $d^j(Z)_k$ we can artificially introduce noise $d^j(Z)_k = \varepsilon e^{i\theta}$ for a small threshold $\varepsilon > 0$ and a randomly selected angle $\theta \in [0, 2\pi)$. It will slightly deteriorate the results, but for weighted graphs the impact of the noise is minor. After these adjustments, the assumption will be satisfied.*

We construct the phase difference matrices as

$$\Phi = A_G \circ T_\delta(\operatorname{sgn} x \operatorname{sgn} x^*) \text{ and } \Psi = A_G \circ \operatorname{sgn}_0(Z).$$

The definition of Ψ naturally implies that the phase differences are only available if $|Z_{k,\ell}| > 0$, which is in line with the construction of E.
For the weights W we consider three possible cases. The first is the unweighted graph with $W = A_G$. If the object is non-vanishing, we observe that $\Phi = A_G \circ \operatorname{sgn}_0(X)$ and the differences $W \circ (\Psi - \Phi), R \circ (\Psi - \Phi)$ appearing throughout this section further simplify to

$$W \circ (\Psi - \Phi) = R \circ (\Psi - \Phi) = A_G \circ (\operatorname{sgn}_0(X) - \operatorname{sgn}_0(Z)).$$

Moreover, the following upper bound holds.

Lemma 3.6.37 (Version of [40, Lemma 6] or [138, Lemma 6])**.** *Let* $\min_{k\in[d]} |x_j| > 0$. *Then,*

$$\|A_G \circ (\operatorname{sgn}_0(X) - \operatorname{sgn}_0(Z))\|_F \leq \|\operatorname{sgn}_0(X) - \operatorname{sgn}_0(Z)\|_F \leq 2\|X - Z\|_F / \min_{k\in[d]} |x_k|^2.$$

Consequently, one of the error bounds, e.g., (3.94) combined with Lemmas 3.6.35 and 3.6.37 reads as

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn} u) \leq \frac{4\sqrt{3}\sqrt{2 + 2\|u\|_\infty^2}\, d \|X - Z\|_F}{\min_{k\in[d]} |x_k|^2 \pi \delta^{3/2}},$$

where u is the solution of EIG.
In the second case, we consider a weighted graph with the weights $W_{k,\ell} = |Z_{k,\ell}| \mathcal{I}_{k\neq\ell}$. Note that $W_{k,\ell} \neq 0$ if and only if $(A_G)_{k,\ell} \neq 0$ and if the assumptions of Lemma 3.6.35 hold, the graph is connected and we can guarantee that $\tau_G > 0$. However, stronger theoretical lower bounds for the spectral gap are not available.
The second choice of weights is convenient for the error bounds expressed via the norm $\|W \circ (\Phi - \Psi)\|_\infty$, which is bounded by

$$\begin{aligned}\|W \circ (\Phi - \Psi)\|_\infty &\leq \|W \circ (\Phi - \Psi)\|_F = \||Z| \circ (\Phi - \Psi)\|_F \\ &\leq \||Z| \circ T_\delta(\operatorname{sgn} x \operatorname{sgn} x^*) \pm |X| \circ T_\delta(\operatorname{sgn} x \operatorname{sgn} x^*) - |Z| \circ \operatorname{sgn}_0(Z)\|_F \\ &\leq \|(|Z| - |X|) \circ T_\delta(\operatorname{sgn} x \operatorname{sgn} x^*)\|_F + \|X - Z\|_F \leq 2\|X - Z\|_F,\end{aligned}$$

where the last inequality follows from the reverse triangle inequality. More precisely,

$$\begin{aligned}\|(|Z| - |X|) \circ T_\delta(\operatorname{sgn} x \operatorname{sgn} x^*)\|_F^2 &= \sum_{k,\ell\in[d]} ||Z_{k,\ell}| - |X_{k,\ell}||^2 |\operatorname{sgn} x_k|^2 |\operatorname{sgn} x_j|^2 \\ &\leq \sum_{k,\ell\in[d]} |Z_{k,\ell} - X_{k,\ell}|^2 = \|X - Z\|_F^2.\end{aligned}$$

Substituting the obtained inequality into the one of the error bounds, e.g., (3.91) gives us

$$\operatorname{dist}(\operatorname{sgn} x, u) \leq 8\tau_G^{-1}\sqrt{d}\|X - Z\|_F,$$

where u is the solution of LSP.

The third and last example of the weight construction is the matrix with the squared magnitudes $W_{k,\ell} = |Z_{k,\ell}|^2 \mathcal{I}_{k\neq\ell}$. In contrast to the previous case, this construction is more convenient when working with the error bounds for angular synchronization based on the norm $\|R \circ (\Phi - \Psi)\|_F$. Similarly, we have

$$\|R \circ (\Phi - \Psi)\|_F \leq \||Z| \circ (\Phi - \Psi)\|_F \leq 2 \|X - Z\|_F ,$$

and, for instance, the error bound (3.101) reads as

$$\operatorname{dist}(\operatorname{sgn} x, \operatorname{sgn}(u_1)) \leq 8\sqrt{\tau_G^{-1} \operatorname{erank}(U)} \|X - Z\|_F ,$$

with u_1 being the top eigenvector of the solution U of SDP.
At last, we would like to consider the total error resulting from the estimation of x by $z = v \circ u$ with magnitudes v and phases u. It can be bounded from above by splitting the magnitude and the phase estimation errors similarly to (3.32) as

$$\begin{aligned}\operatorname{dist}(x, z) &= \min_{|\alpha|=1} \|x - \alpha z\|_2 = \min_{|\alpha|=1} \||x| \circ \operatorname{sgn} x \pm \alpha v \operatorname{sgn} x - \alpha v \circ u\|_2 \\ &\leq \||x| - |z|\|_2 + \min_{|\alpha|=1} \|v \operatorname{sgn} x - \alpha v \operatorname{sgn} z\| = \operatorname{dist}(v \circ \operatorname{sgn} x, v \circ u) + \||x| - |z|\|_2 .\end{aligned}$$

We observe that the phase error is not the plain $\operatorname{dist}(\operatorname{sgn} x, u)$ considered throughout this section, but its weighted analogy $\operatorname{dist}(v \circ \operatorname{sgn} x, v \circ u)$. In this case, the proofs of the bounds can be adjusted to accommodate the weighted error. As an example, we provide the scaled version of the error bound (3.94) and the rest are adjusted analogously.

Theorem 3.6.38. *Let Φ and Ψ be defined as in (3.73). Suppose that $G = (V, E, W)$ is a weighted graph with the Laplacian matrix L_G and the spectral gap $\tau_G > 0$. Let v be a vector satisfying $v_k > 0, k \in [d]$ and let u be the minimizer of EIG. Set $R \in \mathbb{R}^{d\times d}$ as $R_{k,\ell} = W_{k,\ell}^{1/2}$. Then,*

$$\operatorname{dist}(v \circ \operatorname{sgn} x, v \circ \operatorname{sgn} u) \leq 2 \frac{\|v\|_2}{\|v \circ u\|_2} \frac{\sqrt{2 + 2\|u\|_\infty^2}\, \|R \circ (\Phi - \Psi)\|_F}{\lambda_{d-1}^{1/2}(\operatorname{diag}(v)^{-1} L_G \operatorname{diag}(v)^{-1})}. \tag{3.108}$$

Sketch of the proof. The proofs of the inequalities (3.108) and (3.94) are similar and we will only highlight the main differences. Two inequalities

$$\operatorname{dist}^2(v \circ x, v \circ \operatorname{sgn} u) \leq 4 \operatorname{dist}^2 \left(v \circ \operatorname{sgn} x, \frac{\|v\|_2}{\|v \circ u\|_2} v \circ u \right), \tag{3.109}$$

and

$$\operatorname{dist}^2 \left(v \circ \operatorname{sgn} x, \frac{\|v\|_2}{\|v \circ u\|_2} v \circ u \right) \leq \frac{\|v\|_2^2}{\|v \circ u\|_2^2} \frac{2u^* L_\Phi u}{\lambda_{d-1}(\operatorname{diag}(v)^{-1} L_G \operatorname{diag}(v)^{-1})} \tag{3.110}$$

replace the inequalities (3.97) and (3.80), respectively.
Combining the new inequalities with (3.81), (3.82) and (3.83) will grant us the inequality (3.108). The first inequality is required to transit from $v \circ \operatorname{sgn} u$ to $v \circ u$ using the inequality

(3.98). The scaling factor guarantees that the norm of $\frac{\|v\|_2}{\|v\circ u\|_2}\, v\circ u$ is the same as the norm of $v\circ\operatorname{sgn} x$, analogously to u having the same norm as $\operatorname{sgn} x$ in the proof of Theorem 3.6.30. It is crucial for the proof of the second inequality, which allows us to transit from the weighted reconstruction error to the least squares objective. Another difference to (3.94) is the appearance of the matrix $\operatorname{diag}(v)^{-1}L_G\operatorname{diag}(v)^{-1}$ instead of L_G. This change is important, as it makes the nominator free of the scaling matrix $\operatorname{diag}(v)$. The spectral gap of the matrix $\operatorname{diag}(v)^{-1}L_G\operatorname{diag}(v)^{-1}$ is bounded from below by

$$\lambda_{d-1}(\operatorname{diag}(v)^{-1}L_G\operatorname{diag}(v)^{-1}) \geq \lambda_{d-1}(L)\lambda_d^2(\operatorname{diag}(v)^{-1}) = \tau_G\lambda_1^{-2}(\operatorname{diag}(v)) = \tau_G\,\|v\|_\infty^{-2} > 0,$$

which implies that the nullspace of $\operatorname{diag}(v)^{-1}L_\Phi\operatorname{diag}(v)^{-1}$ is spanned by $v\circ\operatorname{sgn} x$. □

We note that (3.108) generalizes (3.94) and by setting $v = \mathbb{1}_d$ both inequalities coincide. We also observe that (3.108) introduces an additional factor $\|v\|_2 / \|v\circ u\|_2$, which seems to be a proof artifact. In fact, we believe that if the inequality (3.109) is avoided and the proof techniques of Theorem 3.6.31 are used instead, it may be possible to avoid this factor, but it would lead to a larger constant factor similarly to the difference between (3.94) and the bound in Theorem 3.6.31.
If a single entry of v is zero, then $\lambda_{d-1}(\operatorname{diag}(v)^{-1}L_G\operatorname{diag}(v)^{-1}) = 0$ and Theorem 3.6.38 is no longer applicable. Moreover, even if v is non-vanishing with some small entries, it strongly affects the aforementioned eigenvalue. In view of Example 3.6.24, we can ignore the estimation of the phases corresponding to the small entries of v, which leads to the following corollary.

Corollary 3.6.39. *Let $v \in \mathbb{C}^d$ be a vector satisfying $v_k \geq 0, k \in [d]$ and define index set*

$$\mathcal{J}_\varepsilon := \{k : \ |v_k| \leq \varepsilon\},$$

for a threshold parameter $\varepsilon \geq 0$. Let v_ε be a vector constructed by excluding the entries in $\mathcal{J}_\varepsilon$. Consider a weighted graph $G = (V, E, W)$ and its subgraph $G_\varepsilon = (V_\varepsilon, E_\varepsilon, W_\varepsilon)$ obtained by removing the all nodes in $\mathcal{J}_\varepsilon$. Assume that the subgraph G_ε is connected so that the spectral gap of its Laplacian $L_{G,\varepsilon}$ is positive. Let Φ and Ψ be defined as in (3.73) and construct Φ_ε and Ψ_ε by removing rows and columns corresponding to $\mathcal{J}_\varepsilon$. Let u_ε be the eigenvector of norm one corresponding to the smallest eigenvalue of a matrix $D_\varepsilon - W_\varepsilon\circ\Psi_\varepsilon$ and construct an estimate of the phases u as

$$u_k = \begin{cases} \operatorname{sgn}(u_\varepsilon)_k, & k \notin \mathcal{J}_\varepsilon, \\ e^{i\theta}, \ \textit{for random } \theta \in [0, 2\pi), & k \in \mathcal{J}_\varepsilon. \end{cases}$$

Then,

$$\operatorname{dist}(v\circ\operatorname{sgn} x, v\circ u) \leq \left(4\varepsilon^2|\mathcal{J}_\varepsilon| + 4\frac{\|v_\varepsilon\|_2^2}{\|v_\varepsilon\circ u_\varepsilon\|_2^2}\frac{(2+2\,\|u_\varepsilon\|_\infty^2)\,\|R_\varepsilon\circ(\Phi_\varepsilon-\Psi_\varepsilon)\|_F^2}{\lambda_{d-|\mathcal{J}_\varepsilon|-1}(\operatorname{diag}(v_\varepsilon)^{-1}L_{G,\varepsilon}\operatorname{diag}(v_\varepsilon)^{-1})}\right)^{1/2},$$

where the matrix R_ε is given by $(R_\varepsilon)_{k,\ell} = (W_\varepsilon)_{k,\ell}^{1/2}$.

Notes and References. *Often, the choice of the quadratic loss function for the data fidelity is motivated by the maximum likelihood estimation under additive Gaussian noise [172, 173, 174, 175] with the availability of all pairwise differences. The only exception is [176], where for pairs $(k,\ell) \notin E$ the measurements $\Psi_{k,\ell} = e^{2\pi\theta_{k,\ell}}$ with $\theta_{k,\ell}$ uniformly distributed in $[0, 2\pi)$.*
While the analysis in the mentioned works is not applicable for the measurements (3.73), it provides an alternative look at the angular synchronization problem. Among the methods analyzed for Gaussian noise are LSP, EIG, SDP as well as generalized power method [172, 174] and an approach based on the cycle consistency combined with message passing [177], which can be connected to the class of iteratively reweighted least squares algorithms [178]. Sometimes, the angular synchronization is viewed as a special case of a more general group synchronization problem [179, 180].
In applications to phase retrieval, the phase synchronization is either considered as a part of the polarization approach [136] or in a context of the Block Phase Retrieval algorithm. The polarization method proposed in [136] constructs the matrix of the phase differences for the unknown phases of the measurements and uses the angular synchronization to recover them. Phase retrieval via polarization was later generalized for the time-frequency measurements [137] and used in ptychography as an initialization [38] for the Error Reduction algorithm discussed in Section 3.5.2.
Originally, BPR was derived with a greedy synchronization [35] and only later it was replaced by the graph formulation of the problem for unweighted graphs and an eigenvalue-based method of phase recovery [40]. In [138], the authors provided an expanded study of LSP for weighted graphs. However, the recovery guarantees for EIG were not derived and it motivated our work [164], results of which are covered in this section. In addition, we included new error bounds for SDP summarized in Theorem 3.6.34, which are applicable even when the relaxation is not tight.

3.6.5 Extension of Block Phase Retrieval for equidistant shifts

While BPR allows for a fast non-iterative reconstruction, it is unfortunately bound to the setting where all shifts are present, $\mathcal{R} = [d]$. In this section, we discuss two ways how to relax this condition and make BPR suitable for the sets of equidistant shifts

$$\mathcal{R}_s = \{0, s, 2s, \dots, d-s\} = \{rs,\ r \in [d/s]\}\,, \tag{3.111}$$

with the shift length $1 \le s < \delta$ such that s is a divisor of both δ and d. The condition $s < \delta$ is required to ensure that two consequent illuminated regions overlap. The generalization of BPR for equidistant shifts was proposed as an attempt to make it applicable for practical scenarios, where all shifts are not available, while $\mathcal{R}$ is a lattice.
The main challenge for an extension of the BPR algorithm to the case of equidistant shifts is the inversion step. Let us recall its outline and highlight the changes arising from the increased shift size. The ptychographic measurements (PTY) can by written as

$$Y_{j,r} = I^m_{j,rs} + N_{j,r} = \mathcal{A}_s(xx^*)_{j,r} + N_{j,r}, \tag{PTY$_s$}$$

for $j \in [m]$, $r \in [d/s]$, with the linear measurement operator $\mathcal{A}_s : \mathbb{H}^d \to \mathbb{R}^{md/s}$ given by

$$\mathcal{A}_s(Z)_{j,r} := \langle Z, S_{rs} w^j (S_{rs} w^j)^* \rangle_F. \tag{3.112}$$

We note that the space $\operatorname{span}\{S_{rs}w^j(S_{rs}w^j)^*,\ j\in[m],\ r\in[d/s]\}$ is a subspace of

$$\mathbb{T}_{d,\delta,s} := \big\{U\in\mathbb{H}^d : U_{k,\ell}=0,\ \text{whenever } k,\ell\in[d] \text{ such that} \\ \text{both } k,\ell\notin\{rs,\dots,rs+\delta-1\} \text{ for all } r\in[d/s]\big\}. \tag{3.113}$$

We will use the short notation $\mathbb{T}_{\delta,s}$ for $\mathbb{T}_{d,\delta,s}$ unless the dimension parameter is not equal to d. Then, in analogue to $s=1$, we only measure the entries of xx^*, which belong to $\mathbb{T}_{\delta,s}$, so that

$$Y=\mathcal{A}_s(xx^*)=\mathcal{A}_s(T_{\delta,s}(xx^*)),$$

where $T_{\delta,s}$ denotes the projection on $\mathbb{T}_{\delta,s}$ given by

$$T_{\delta,s}(U)_{k,\ell}=\begin{cases}U_{k,\ell}, & (k,\ell)\in\{rs,\dots,rs+\delta-1\}^2 \text{ for some } r\in[d/s],\\ 0, & \text{otherwise,}\end{cases} \tag{3.114}$$

for all $U\in\mathbb{H}^d$. The inversion step recovers $T_{\delta,s}(xx^*)$ from the system of linear equations (PTY_s), which is only possible if this system is overdetermined. Otherwise, $T_{\delta,s}(xx^*)$ cannot be uniquely identified from the measurements. The unique recovery of $T_{\delta,s}(xx^*)$ can also be expressed in the form of an equality between spaces,

$$\operatorname{span}\{S_{rs}w^j(S_{rs}w^j)^*,\ j\in[m],\ r\in[r/s]\}=\mathbb{T}_{\delta,s}.$$

Unfortunately, that is no longer true for $s>1$.

Theorem 3.6.40. *Let $d\geq 2\delta-s$. Then, the dimension of $\mathbb{T}_{\delta,s}$ over the field $\mathbb{R}$ is given by*

$$\dim(\mathbb{T}_{\delta,s})=(2\delta-s)d.$$

Furthermore, for $s>1$

$$\operatorname{span}\{S_{rs}w^j(S_{rs}w^j)^*,\ j\in[m],\ r\in[d/s]\}\neq\mathbb{T}_{\delta,s}.$$

For the proof we need to introduce the space of block constant matrices

$$\begin{aligned}\mathbb{C}_s^{d\times d} &:= \left\{U\in\mathbb{C}^{d\times d} : U=\begin{bmatrix}\tilde{U}_{0,0}\mathbb{1}_{s\times s} & \dots & \tilde{U}_{0,d/s-1}\mathbb{1}_{s\times s}\\ \vdots & \ddots & \vdots\\ \tilde{U}_{d/s-1,0}\mathbb{1}_{s\times s} & \dots & \tilde{U}_{d/s-1,d/s-1}\mathbb{1}_{s\times s}\end{bmatrix},\ \tilde{U}\in\mathbb{C}^{d/s\times d/s}\right\}\\ &= \left\{U\in\mathbb{C}^{d\times d} : U=\tilde{U}\otimes\mathbb{1}_{s\times s},\ \tilde{U}\in\mathbb{C}^{d/s\times d/s}\right\}\\ &= \left\{U\in\mathbb{C}^{d\times d} : U_{k_1s+\ell_1,k_2s+\ell_2}=\tilde{U}_{k_1,k_2},\ k_1,k_2\in[d/s],\ \ell_1,\ell_2\in[s],\ \tilde{U}\in\mathbb{C}^{d/s\times d/s}\right\},\end{aligned}$$

with $\mathbb{1}_{s\times s}\in\mathbb{R}^{s\times s}$ denoting the matrix with all entries equal to 1 and $\otimes$ being the tensor product (2.3). By definition, the space $\mathbb{C}_s^{d\times d}$ is isomorphic to $\mathbb{C}^{d/s\times d/s}$ with isomorphism mapping $\tilde{}:\mathbb{C}_s^{d\times d}\to\mathbb{C}^{d/s\times d/s}$ acting as $U\mapsto\tilde{U}$.
The important properties related to $\mathbb{C}_s^{d\times d}$ and $\mathbb{T}_{\delta,s}$ are summarized in the next lemma.

Lemma 3.6.41. *1. The isomorphism $\tilde{}$ is a linear mapping, that is for $U,V\in\mathbb{C}_s^{d\times d}$ and $\alpha,\beta\in\mathbb{C}$, $[\alpha U+\beta V]\tilde{}=\alpha\tilde{U}+\beta\tilde{V}$.*

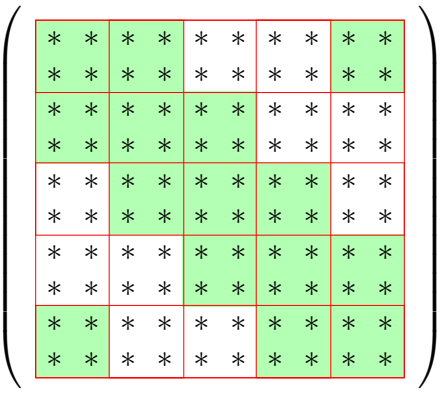


Figure 3.3: Highlighted in green are entries of a matrix, which form the space $\mathbb{T}_{\delta,s}$ for $d = 10$, $\delta = 4$ and $s = 2$. The $s \times s$ blocks correspond to constant values of the matrix $U \in \mathbb{C}_s^{d\times d}$. Consequently, by considering the red blocks as single entries of $\tilde{U}$, we observe that $[T_{\delta,s}(U)]^\sim$ coincides with $T_{d/s,\delta/s,1}$, which is in accordance with Lemma 3.6.41.

2. Let $U \in \mathbb{C}_s^{d\times d}$. Then $T_{\delta,s}(U) \in \mathbb{T}_{\delta,s} \cap \mathbb{C}_s^{d\times d}$ and $[T_{\delta,s}(U)]^\sim = T_{d/s,\delta/s,1}(\tilde{U})$, so that

$$T_{\delta,s}(U) = T_{d/s,\delta/s,1}(\tilde{U}) \otimes \mathbb{1}_{s\times s}.$$

For a visual justification of Lemma 3.6.41, we refer the reader to Figure 3.3.

Proof. 1. Follows directly from Proposition 2.1.2.
2. By (3.114), the matrix $T_{\delta,s}(U)$ is in $\mathbb{T}_{\delta,s}$. Let us show that for all index pairs $(k_1 s + \ell_1, k_2 s + \ell_2)$, $k_1, k_2 \in [d/s]$, $\ell_1, \ell_2 \in [s]$, we have that $T_{\delta,s}(U)_{k_1 s+\ell_1, k_2 s+\ell_2}$ is a constant with respect to ℓ_1 and ℓ_2, i.e., either the whole constant block is preserved or nullified completely by the projection operator. If there exists $r \in [d/s]$ such that $(k_1 s + \ell_1, k_2 s + \ell_2) \in \{rs, \dots, rs + \delta - 1\}^2$, then it is equivalent to

$$rs \le k_p s + \ell_p < rs + \delta, \quad p = 1, 2,$$

and

$$r \le k_p + \ell_p/s < r + \delta/s, \quad p = 1, 2.$$

Since $r, r + \delta/s$, k_1 and k_2 are integers and $0 \le \ell_p \le s - 1$, $p = 1, 2$, we have that

$$r \le \lfloor k_p + \ell_p/s \rfloor = k_p \text{ and } r + \delta/s \ge \lceil k_p + \ell_p/s \rceil = k_p + 1. \quad p = 1, 2. \tag{3.115}$$

Consequently, for any $q_p \in [s]$, $p = 1, 2$, we obtain

$$k_p s + q_p \ge rs + 0 = rs \text{ and } k_p s + q_p \le (r + \delta/s - 1)s + s - 1 = rs + \delta - 1, \quad p = 1, 2,$$

and $(k_1 s + q_1, k_2 s + q_2) \in \{rs, \dots, rs + \delta - 1\}^2$. Therefore, by the definition (3.114) of the projection operator $T_{\delta,s}$, for all $q_1, q_2 \in [s]$, the entries $T_{\delta,s}(U)_{k_1 s+q_1, k_2 s+q_2}$ satisfy

$$T_{\delta,s}(U)_{k_1 s+q_1, k_2 s+q_2} = U_{k_1 s+q_1, k_2 s+q_2} = \tilde{U}_{k_1,k_2}.$$

On the other hand, if $(k_1 s + \ell_1, k_2 s + \ell_2) \notin \{rs, \dots, rs + \delta - 1\}^2$ for all $r \in [d/s]$, then for all $q_1, q_2 \in [s]$ and all $r \in [d/s]$ we have $(k_1 s + q_1, k_2 s + q_2) \notin \{rs, \dots, rs + \delta - 1\}^2$.

For a proof by contradiction, assume that there exists $q_1, q_2 \in [s]$ and $r \in [d/s]$ such that $(k_1 s + q_1, k_2 s + q_2) \in \{rs, \dots, rs + \delta - 1\}^2$. Then, by the considerations above $(k_1 s + \ell_1, k_2 s + \ell_2) \in \{rs, \dots, rs + \delta - 1\}^2$, which is impossible. Consequently, for all $q_1, q_2 \in [s]$ we have

$$T_{\delta,s}(U)_{k_1 s+q_1, k_2 s+q_2} = 0.$$

Hence, $T_{\delta,s}(U) \in \mathbb{C}_s^{d\times d}$ and the inequalities (3.115) together with the definition (3.114) yield

$$\begin{aligned}[T_{\delta,s}(U)]^{\sim}_{k_1,k_2} &= \begin{cases} \tilde{U}_{k_1,k_2}, & (k_1,k_2) \in \{r, \dots, r + \delta/s - 1\} \text{ for some } r \in [d/s], \\ 0, & \text{otherwise}, \end{cases} \\ &= T_{d/s,\delta/s,1}(\tilde{U})_{k_1,k_2}. \end{aligned}$$

□

Now, we are ready to proof Theorem 3.6.40.

Proof of Theorem 3.6.40. We consider a matrix $U \in \mathbb{T}_{\delta,s}$ and count the number of the unknowns which may be non-zero. We note that the entries $U_{k,k}$ on the main diagonal are real-valued, which gives us one real unknown per entry. The off-diagonal entries $U_{k,\ell}$ are complex and give two real unknowns per entry. However, since U is Hermitian, $U_{k,\ell} = \overline{U}_{\ell,k}$ holds and, thus, each non-zero entry contributes precisely one real unknown. Therefore, the dimension of $\mathbb{T}_{\delta,s}$ coincides with the cardinality of the index set

$$\mathcal{K}_s := \{(k,\ell) \in [d]^2 : \text{there exists } r \in [d/s] \text{ such that } k, \ell \in \{rs, \dots, rs + \delta - 1\}\}.$$

Note that a pair (k,ℓ) belongs to $\mathcal{K}_s$ if and only if $T_{\delta,s}(\mathbb{1}_{d\times d})_{k,\ell} = 1$. Hence, the cardinality of $\mathcal{K}_s$ is given by

$$\|T_{\delta,s}(\mathbb{1}_{d\times d})\|_F^2 = \sum_{(k,\ell)\in\mathcal{K}_s} 1 = |\mathcal{K}_s|.$$

The matrix $\mathbb{1}_{d\times d}$ belongs to $\mathbb{C}_s^{d\times d}$ with $[\mathbb{1}_{d\times d}]^{\sim} = \mathbb{1}_{d/s\times d/s}$. Then, by Lemma 3.6.41, we have $T_{\delta,s}(\mathbb{1}_{d\times d}) \in \mathbb{T}_{\delta,s} \cap \mathbb{C}_s^{d\times d}$ with $[T_{\delta,s}(\mathbb{1}_{d\times d})]^{\sim} = T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})$. The definitions of the projection operators (3.114) and (3.30) yield

$$T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})_{k_1,k_2} = \begin{cases} (\mathbb{1}_{d/s\times d/s})_{k_1,k_2}, & |k_1 - k_2|_c < \delta/s, \\ 0, & \text{otherwise}, \end{cases} = \begin{cases} 1, & |k_1 - k_2|_c < \delta/s, \\ 0, & \text{otherwise}, \end{cases}$$

for $k_1, k_2 \in [d/s]$. If $d/s \le (2\delta/s - 1)$, then by Remark 3.6.1, condition $|k_1 - k_2|_c < \delta/s$ is always true and $\mathbb{T}_{d/s,\delta/s,1} = \mathbb{H}^{d/s}$. If $d/s \ge (2\delta/s - 1)$ for a fixed $k_1 \in [d/s]$, there are precisely $2\delta/s - 1$ indices $k_2 \in [d/s]$ satisfying the condition $|k_1 - k_2|_c < \delta/s$.
Consequently, by Proposition 2.1.2 we obtain

$$\begin{aligned}|\mathcal{K}_s| = \|T_{\delta,s}(\mathbb{1}_{d\times d})\|_F^2 &= \left\|T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})\right\|_F^2 \|\mathbb{1}_{s\times s}\|_F^2 = s^2 \sum_{k_1,k_2\in[d/s]} |T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})_{k_1,k_2}|^2 \\ &= s^2 \sum_{k_1\in[d/s]} \sum_{k_2:|k_1-k_2|_c<\delta/s} 1 = s^2 \sum_{k_1\in[d/s]} (2\delta/s - 1) = (2\delta - s)d.\end{aligned}$$

The proof is concluded by comparing dimensions of the spaces. In view of Remark 3.6.5, the dimension of the span over $\mathbb{R}$ is at most $(2\delta-1)|\mathcal{R}_s| = (2\delta-1)d/s$. Let us show that $(2\delta-1)d/s < (2\delta-s)d$ for all $1 < s < \delta$. The desired inequality is equivalent to

$$s^2 - 2\delta s + 2\delta - 1 < 0.$$

This polynomial has a positive leading coefficient 1 and two zeros $s = 1$ and $s = 2\delta - 1$. Therefore, for all $2 \le s \le 2\delta - 2$ it is negative and for $\delta > 1$ we have $2\delta - 2 \ge \delta$. Hence, $s^2 - 2\delta s + 2\delta - 1 < 0$ for $1 < s < \delta$, which concludes the proof. □

Theorem 3.6.40 tells us that the inversion step will fail in case $s > 1$ without further modifications. As a cure for this problem, there are three possible adjustments to ensure the success of the inversion step. The first option is to enhance the recovery algorithm by incorporating the rank-one structure of the matrix xx^*, similarly to the subspace completion in Section 3.6.2.2 or in analogy to low-rank matrix recovery [126, 92] or matrix completion [181] methods. In the second case, additional assumptions on x can be made so that the dimension of the space $\mathbb{T}_{\delta,s}$ intersected with the space introduced by the assumptions will be sufficiently small. The last option is to increase the number of measurements, so that the dimension of the span will match the dimension of $\mathbb{T}_{\delta,s}$. In the rest of the section we will discuss the latter two options.

3.6.5.1 Block Phase Retrieval for piecewise constant objects

In this subsection we will consider the recovery of objects in the subspace of piecewise constant vectors

$$\begin{aligned}\mathbb{C}^d_s &:= \{u \in \mathbb{C}^d : u = (\tilde{u}_0 \mathbb{1}_s^T, \tilde{u}_1 \mathbb{1}_s^T, \dots, \tilde{u}_{d/s-1}\mathbb{1}_s^T)^T,\ \tilde{u} \in \mathbb{C}^{d/s}\} \\ &= \{u \in \mathbb{C}^d : u_{ks+\ell} = \tilde{u}_k,\ k \in [d/s], \ell \in [s]\ \tilde{u} \in \mathbb{C}^{d/s}\},\end{aligned} \tag{3.116}$$

where $\mathbb{1}_s \in \mathbb{R}^s$ is a vector with all entries equal to 1. By its definition, the space $\mathbb{C}^d_s$ is isomorphic to $\mathbb{C}^{d/s}$ and we will abuse the notation for the isomorphism $\tilde{\ } : \mathbb{C}^d_s \to \mathbb{C}^{d/s}$ acting as $u \mapsto \tilde{u}$.
The choice of $\mathbb{C}^d_s$ is motivated by the structure of the set of shifts $\mathcal{R}_s$. For $r \in [d/s]$ we note that $[S_{-rs}x]^{\tilde{}} = S_{-r}\tilde{x}$ and, hence, for $\tilde{x}$ all d/s shifts are observed. However, the measurements with regard to $\tilde{x}$ are not of the form (PTY). More precisely,

$$\begin{aligned}|(F_m P_m[w \circ S_{-rs}x])_j|^2 &= \left|\sum_{t\in[\delta]} w_t x_{t+rs} e^{-\frac{2\pi i t j}{m}}\right|^2 = \left|\sum_{k\in[\delta/s]}\sum_{\ell\in[s]} w_{ks+\ell} x_{(k+r)s+\ell} e^{-\frac{2\pi i (ks+\ell) j}{m}}\right|^2 \\ &= \left|\sum_{k\in[\delta/s]} \tilde{x}_{k+r} e^{-\frac{2\pi i k j}{m/s}} \sum_{\ell\in[s]} w_{ks+\ell} e^{-\frac{2\pi i \ell j}{m}}\right|^2, \quad r \in [d/s],\ j \in [m],\end{aligned}$$

where the second sum is not a constant with respect to j and, moreover, the fraction m/s is not necessarily an integer. Nevertheless, by defining the masks

$$v^j_k := e^{\frac{2\pi i k j}{m/s}} \sum_{\ell\in[s]} \overline{w}_{ks+\ell} e^{\frac{2\pi i \ell j}{m}}, \quad j \in [m],\ k \in [d/r],$$

with $\operatorname{supp}(v^j) \subseteq [\delta/s]$ the measurements with respect to $\tilde{x}$ are of the form

$$Y_{j,r} = |\langle S_{-r}\tilde{x}, v^j\rangle|^2 + N_{j,r} = |\langle \tilde{x}, S_r v^j\rangle|^2 + N_{j,r}, \quad j \in [m],\ r \in [d/s]. \tag{3.117}$$

Originally, the BPR algorithm is designed for the measurements (3.117) and can be applied to recover $\tilde{x}$ [35]. We will not cover the details, however, it is important to mention that the mapping $T_{\delta,s}(xx^*) \mapsto Y$ is based on multiple non-trivial matrix multiplications and, consequently, the inversion step involves the computation of the pseudoinverse matrix for d/s matrices of size $m \times \delta/s$. This may become a computational bottleneck if parameters d, δ or m are large, e.g., when working with two-dimensional images.
One potential treatment of this problem is to discard some frequencies $j \in [m]$ to reduce the size of the matrices. Another approach is to avoid applying BPR to the measurements (3.117) directly and to work with the object $x \in \mathbb{C}^d_s$. In this way, it is possible to reconstruct $T_{d/s,\delta/s,1}(\tilde{x}\tilde{x}^*)$ in the inversion step and to proceed with the rest of the steps of Algorithm 3 for $\tilde{x}$. This allows to avoid the computation of pseudoinverse matrices completely.
For $x \in \mathbb{C}^d_s$, the entries of the matrix xx^* are given by

$$(xx^*)_{k_1 s+\ell_1, k_2 s+\ell_2} = x_{k_1 s+\ell_1}\overline{x}_{k_2 s+\ell_2} = \tilde{x}_{k_1}\overline{\tilde{x}}_{k_2} = (\tilde{x}\tilde{x}^*)_{k_1,k_2}$$

for all $k_1, k_2 \in [d/s]$, $\ell_1, \ell_2 \in [s]$, so that $xx^* \in \mathbb{C}^{d\times d}_s$ with $[xx^*]^\sim = \tilde{x}\tilde{x}^*$. By Lemma 3.6.41 we have

$$T_{\delta,s}(xx^*) \in \mathbb{T}_{\delta,s} \cap \mathbb{C}^{d\times d}_s \text{ and } [T_{\delta,s}(xx^*)]^\sim = T_{d/s,\delta/s,1}(\tilde{x}\tilde{x}^*).$$

Therefore, we can consider the inversion of the measurements on the space $\mathbb{T}_{\delta,s} \cap \mathbb{C}^{d\times d}_s$, which is isomorphic to $\mathbb{T}_{d/s,\delta/s,1}$. By Theorem 3.6.40 the dimension of $\mathbb{T}_{d/s,\delta/s,1}$ over $\mathbb{R}$ is $(2\delta/s - 1)d/s$, which is less than the dimension of the space spanned by the lifted measurement vectors. Thus, it is possible to recover $T_{d/s,\delta/s,1}(\tilde{x}\tilde{x}^*)$ from the measurements. This can be done by an analogue of Theorem 3.6.4 connecting the Fourier coefficients of the diagonals of a matrix $U \in \mathbb{T}_{\delta,s} \cap \mathbb{C}^{d\times d}_s$ to the measurements $\mathcal{A}_s(U)$.

Theorem 3.6.42. *Let $m \geq 2\delta - s$. Consider the measurement operator $\mathcal{A}_s$ defined in (3.112) with the set of shifts $\mathcal{R}_s$ as in (3.111). Then, for all $U \in \mathbb{T}_{\delta,s} \cap \mathbb{C}^{d\times d}_s$ the equality*

$$(F_m^{-1}\mathcal{A}_s(U)F_{d/s})_{(js)} = F_{d/s}[d^j(\tilde{U})] \circ F_{d/s}h^j,$$

holds for all $j \in [\delta/s]$, with vectors $h^j \in \mathbb{C}^{d/s}$, $j \in [\delta/s]$, given by

$$h^j_k := \sum_{\ell\in[s]} (R_d[w \circ S_{js}\overline{w}])_{ks-\ell \bmod d}. \tag{3.118}$$

Proof. We note that the proof of Theorem 3.6.4 up to the equality (3.37) is valid for $\mathcal{R}_s$. We will use (3.37) only for $js \in [m]$ with $j \in [\delta/s]$. Because of that, it is possible to relax condition $m \geq 2\delta - 1$ required for (3.36) to $m \geq 2\delta - s$, since the stronger inequality $\rho(js) = js \leq \delta - s$ is satisfied. Therefore, the equality (3.37) yields

$$F_m^{-1}\mathcal{A}_s(U)_{js,r} = \left(d^{js}(U) *_d (R_d[w \circ S_{js}\overline{w}])\right)_{rs}, \quad j \in [\delta/s],\ r \in [d/s].$$

By the definitions of $d^{js}(U)$ and $\mathbb{C}^{d\times d}_s$, for $j \in [\delta/s]$, $k \in [d/s]$, $\ell \in [s]$, we have

$$d^{js}(U)_{ks+\ell} = U_{ks+\ell,ks+\ell-js} = U_{ks+\ell,(k-j)s+\ell} = \tilde{U}_{k,k-j} = d^j(\tilde{U})_k,$$

so that $d^{js}(U) \in \mathbb{C}_s^d$ and $[d^{js}(U)]^\sim = d^j(\tilde{U})$. Then, we expand the circular convolution as

$$\begin{aligned} F_m^{-1}\mathcal{A}_s(U)_{js,r} &= \sum_{t\in[d]} d^{js}(U)_t (R_d[w \circ S_{js}\overline{w}])_{rs-t \bmod d} \\ &= \sum_{k\in[d/s]} \sum_{\ell\in[s]} d^{js}(U)_{ks+\ell} (R_d[w \circ S_{js}\overline{w}])_{(r-k)s-\ell \bmod d} \\ &= \sum_{k\in[d/s]} d^j(\tilde{U})_k \sum_{\ell\in[s]} (R_d[w \circ S_{js}\overline{w}])_{(r-k)s-\ell \bmod d} \\ &= \sum_{k\in[d/s]} d^j(\tilde{U})_k h^j_{r-k \bmod d/s} = (d^j(\tilde{U}) *_{d/s} h^j)_r. \end{aligned}$$

By the circular convolution theorem (Theorem 2.2.3), the application of the Fourier transform $F_{d/s}$ with respect to the variable r leads to

$$(F_m^{-1}\mathcal{A}_s(U)F_{d/s})_{(js)} = (F_m^{-1}\mathcal{A}_s(U)F_{d/s}^T)_{(js)} = (F_m^{-1}\mathcal{A}_s(U))_{(js)}F_{d/s}^T = F_{d/s}[d^j(\tilde{U})] \circ F_{d/s}h^j.$$

□

In analogy to the inversion step for $\mathcal{R} = [d]$, for $\mathcal{R}_s$ we use Theorem 3.6.42 to construct a matrix $\tilde{Z} \in \mathbb{T}_{d/s,\delta/s,1}$ from the noisy measurements $Y = \mathcal{A}_s(T_{\delta,s}(xx^*)) + N$ by its diagonals $d^j(\tilde{Z})$, $j \in [\delta/s]$ as

$$d^j(\tilde{Z}) = F_{d/s}^{-1}\left[\frac{(F_m^{-1}YF_{d/s})_{(js)}}{F_{d/s}h^j}\right] = d^j(T_{d/s,\delta/s,1}(\tilde{x}\tilde{x}^*)) + F_{d/s}^{-1}\left[\frac{(F_m^{-1}NF_{d/s})_{(js)}}{F_{d/s}h^j}\right]. \quad (3.119)$$

We note that in contrast to the case $s = 1$, for $s > 1$ this procedure does not coincide with $\mathcal{A}_s\big|^\dagger_{\mathbb{T}_{\delta,s}\cap\mathbb{C}_s^{d\times d}}(Y)$ as in the proof of Theorem 3.6.42 we only use entries $F_m^{-1}\mathcal{A}_s(U)_{\ell,r}$, for $\ell = js$, $j \in [\delta/s]$, $r \in [d/s]$, while the rest of the values for $\ell \in [m]$ are discarded. Nevertheless, we can quantify the resulting reconstruction error.

Corollary 3.6.43. *Let $m \geq 2\delta - s$. Consider the ptychographic measurements Y of the form (PTY$_s$) corresponding to the set of shifts $\mathcal{R}_s$ as in (3.111). Let h^j be defined by (3.118) and assume that*

$$\sigma := \sqrt{m} \min_{k,j\in[d/s]} |(F_{d/s}h^j)_k| \quad (3.120)$$

satisfies $\sigma > 0$. Then, for the matrix $\tilde{Z}$ constructed by (3.119) we have

$$\left\|\tilde{Z} - T_{d/s,\delta/s,1}(\tilde{x}\tilde{x}^*)\right\|_F \leq \sigma^{-1}\|N\|_F.$$

Proof. The proof is analogous to the proof of Corollary 3.6.6. □

With the matrix $\tilde{Z}$, the steps of Algorithm 3 can be repeated in order to reconstruct $\tilde{z} \approx \tilde{x}$ and the complete recovery method is summarized below.

Algorithm 7: Block Phase Retrieval for shifts $\mathcal{R}_s$ and objects in $\mathbb{C}_s^d$

Input : Shift size $1 \leq s < \delta$, s divisor of d and δ, noisy measurements $Y \in \mathbb{R}^{m \times d/s}$ of the form (PTY$_s$) with $\mathcal{R}_s$ as in (3.111).

Output: $z \in \mathbb{C}_s^d$ with $z \approx e^{-i\theta}x$ for some $\theta \in [0, 2\pi)$.

1. Construct $\tilde{Z} \in \mathbb{T}_{d/s,\delta/s,1}$ by its diagonals via (3.119) as an estimate of $T_{d/s,\delta/s,1}(\tilde{x}\tilde{x}^*)$.
2. Estimate the magnitudes $\tilde{v} \in \mathbb{R}^{d/s}$ from $\tilde{Z}$ by a method of choice.
3. Estimate the phases $\tilde{u} \in \mathbb{C}^{d/s}$ from $\tilde{Z}$ by a method of choice.
4. Set $\tilde{z} = \tilde{v} \circ \tilde{u} \in \mathbb{C}^{d/s}$ to form $\tilde{z} \approx \tilde{x}$.
5. Construct $z \in \mathbb{C}_s^d$ isomorphic to $\tilde{z}$.

As Algorithm 7 performs the inversion step via (3.119) and then proceeds with recovery of the magnitudes from $\tilde{Z} \approx T_{d/s,\delta/s,1}(\tilde{x}\tilde{x}^*)$. The dimension of the underlying problem is smaller than for $s = 1$. This constitutes the computation complexity of $\mathcal{O}(\frac{\delta^2 d}{s^3} \log \frac{d}{s} + \frac{md}{s} \log m)$ if Diagonal Magnitude Estimation is used.
Furthermore, both the truncation and the subspace completion techniques discussed in Section 3.6.2.2 can be used for the inversion step to treat the cases when σ is either small or equal to zero.
For the reconstruction error of Algorithm 7, we can recycle the results developed for the case $s = 1$. For instance, if Algorithm 7 uses Diagonal Magnitude Estimation and the unweighted phase synchronization, we obtain the following analogue of Theorem 3.6.2.

Corollary 3.6.44. *Consider ptychographic measurements of the form* (PTY$_s$) *with* $\mathcal{R}_s$ *as in* (3.111). *Let the parameters satisfy* $\delta > 2s$, $d \geq 4\delta$, $m \geq 2\delta - s$ *and assume that* σ *defined in* (3.120) *is non-zero. If* $x \in \mathbb{C}_s^d$ *is non-vanishing with* $|x|_{\min} := \min_{k \in [d]} |x_k|$, *then the estimate* $z \in \mathbb{C}_s^d$ *produced in Algorithm 7 satisfies*

$$\operatorname{dist}(x, z) \leq 24 \frac{\|x\|_\infty}{|x|_{\min}^2} \cdot \frac{sd^2}{\delta^{5/2}} \cdot \sigma^{-1} \|N\|_F + (ds)^{1/4} \sqrt{\sigma^{-1} \|N\|_F}.$$

Proof. We observe that

$$\begin{aligned} \operatorname{dist}^2(x, z) &= \min_{|\alpha|=1} \|x - \alpha z\|_2^2 = \min_{|\alpha|=1} \sum_{t \in [d]} |x_t - \alpha z_t|^2 = \min_{|\alpha|=1} \sum_{k \in [d/s]} \sum_{\ell \in [s]} |x_{ks+\ell} - \alpha z_{ks+\ell}|^2 \\ &= s \min_{|\alpha|=1} \sum_{k \in [d/s]} |\tilde{x}_k - \alpha \tilde{z}_k|^2 = s \min_{|\alpha|=1} \|\tilde{x} - \alpha \tilde{z}\|_2^2 = s \operatorname{dist}^2(\tilde{x}, \tilde{z}). \end{aligned}$$

The distance $\operatorname{dist}(\tilde{x}, \tilde{z})$ can be bounded by Theorem 3.6.2. Note that $\tilde{x}, \tilde{z}$ are in $\mathbb{C}^{d/s}$ and in Theorem 3.6.2 parameters d and δ are replaced by d/s and δ/s, respectively. Hence, the requirements on parameters $\delta/s \geq 2$ and $d/s \geq 4\delta/s$ in Theorem 3.6.2 are satisfied. The inversion step for $s \geq 1$ in Corollary 3.6.43 requires $m \geq 2\delta - s$ instead of $m \geq 2\delta - 1$ for the case $s = 1$. Consequently, $\sigma_{\delta d}(\mathcal{A}_s)$ is replaced with σ. Therefore, we have that

$$\begin{aligned} \operatorname{dist}(x, z) &= \sqrt{s} \operatorname{dist}(\tilde{x}, \tilde{z}) \leq \sqrt{s} \left[24 \frac{\|\tilde{x}\|_\infty}{|\tilde{x}|_{\min}^2} \cdot \frac{(d/s)^2}{(\delta/s)^{5/2}} \cdot \sigma^{-1} \|N\|_F + (d/s)^{1/4} \sqrt{\sigma^{-1} \|N\|_F} \right] \\ &= 24 \frac{\|x\|_\infty}{|x|_{\min}^2} \frac{sd^2}{\delta^{5/2}} \sigma^{-1} \|N\|_F + s^{1/4} d^{1/4} \sqrt{\sigma^{-1} \|N\|_F}, \end{aligned}$$

where it was used that

$$\|x\|_\infty = \max_{t\in[d]} |x_t| = \max_{k\in[d/s],\ell\in[s]} |x_{ks+\ell}| = \max_{k\in[d/s],\ell\in[s]} |\tilde{x}_k| = \max_{k\in[d/s]} |\tilde{x}_k| = \|\tilde{x}\|_\infty ,$$

and, analogously, $|x|_{\min} = |\tilde{x}|_{\min}$. □

In view of the inequality (3.32), this corollary can be adapted to any choice of the inversion, the magnitude and the phase reconstruction steps.

3.6.5.2 Block Phase Retrieval for multiple windows

In this subsection, we suppose that the ptychographic experiment was repeated $Q \in \mathbb{N}$ times. For each trial $q \in [Q]$ a different window $w^{(q)}$ with $\operatorname{supp}(w^{(q)}) \subseteq [\delta]$ was used and the resulting measurement operators (3.112), the measurements of the form (PTY_s) and the noise matrices are denoted by $\mathcal{A}_s^q$, Y^q and N^q, respectively.
By considering multiple experiments, the dimension of the span is bigger and, thus, it will be possible to perform the inversion step for matrices in $\mathbb{T}_{\delta,s}$.
The framework for the reconstruction is a generalization of Theorem 3.6.4 and [141, Theorem 4] for the equidistant shifts $\mathcal{R}_s$.

Theorem 3.6.45. *Let $m \geq 2\delta - 1$. Consider the measurement operator $\mathcal{A}_s$ defined in (3.112) with the set of shifts $\mathcal{R}_s$ as in (3.111). For $j \in [m]$ define the transform*

$$\rho(j) := \begin{cases} j, & j \leq \lfloor m/2 \rfloor, \\ j-m, & j > \lfloor m/2 \rfloor. \end{cases}$$

Then, for all $U \in \mathbb{T}_{\delta,s}$ the equality

$$(F_m^{-1}\mathcal{A}_s(U)F_{d/s})_{j,r} = \frac{1}{s}\sum_{k\in[s]} F_d[d^{\rho(j)}(U)]_{r-kd/s}\overline{F_d[\overline{w}\circ S_{\rho(j)}w]}_{r-kd/s}, \quad j\in[m], r\in[d/s]$$

holds. Furthermore, for $\delta \leq j \leq m-\delta$ the coefficients $F_d[\overline{w}\circ S_{\rho(j)}w]_k$ are zero.

Proof. We recall that the proof of Theorem 3.6.4 up to the equality (3.37) is valid for $\mathcal{R}_s$. Therefore, we have

$$F_m^{-1}\mathcal{A}_s(U)_{j,r} = \left(d^{\rho(j)}(U) *_d (R_d[w\circ S_{\rho(j)}\overline{w}])\right)_{rs}, \quad j\in[m],\ r\in[d/s].$$

Consider the subsampling operator $Z_s : \mathbb{C}^d \to \mathbb{C}^{d/s}$, which acts as $(Z_s v)_r = v_{rs}$, $r \in [d/s]$. Then, the right-hand side can be written as $Z_s u^j$ with vectors $u^j := d^{\rho(j)}(U) *_d (R_d[w\circ S_{\rho(j)}\overline{w}])$.
The application of the Fourier transform $F_{d/s}$ with respect to the variable r leads to

$$(F_m^{-1}\mathcal{A}_s(U)F_{d/s})_{(j)} = (F_m^{-1}\mathcal{A}_s(U)F_{d/s}^T)_{(j)} = (F_m^{-1}\mathcal{A}_s(U))_{(j)}F_{d/s}^T = F_{d/s}Z_s u^j$$

By [141, Lemma 6], we obtain

$$(F_m^{-1}\mathcal{A}_s(U)F_{d/s})_{j,r} = \frac{1}{s}\sum_{k\in[s]} F_d u^j_{r-kd/s} = \frac{1}{s}\sum_{k\in[s]} F_d\left[d^{\rho(j)}(U) *_d (R_d[w\circ S_{\rho(j)}\overline{w}])\right]_{r-kd/s}.$$

The last step is to apply Theorem 2.2.3, which gives

$$(F_m^{-1}\mathcal{A}_s(U)F_{d/s})_{j,r} = \frac{1}{s}\sum_{k\in[s]} F_d[d^{\rho(j)}(U)]_{r-kd/s} F_d R_d[w \circ S_{\rho(j)}\overline{w}]_{r-kd/s},$$

and then Proposition 2.2.5, so that the second term is transformed to the desired form,

$$F_d R_d[w \circ S_{\rho(j)}\overline{w}]] = R_d F_d\overline{[w \circ S_{\rho(j)}\overline{w}]} = \overline{F_d[w \circ S_{\rho(j)}\overline{w}]} = \overline{F_d[\overline{w} \circ S_{\rho(j)}w]}.$$

□

The result of Theorem 3.6.45 once again justifies the dimension counting argument of Theorem 3.6.40. For s unknowns on the right-hand side, only a single measurement is available. Thus, repeating the ptychographic experiment $Q \geq s$ times should provide a sufficient number of measurements for the reconstruction of the diagonals.
We recall that $U \in \mathbb{T}_{\delta,s}$ is completely identified by its diagonals $d^j(U)$, $j \in [\delta]$. For a recovery procedure of a single diagonal $d^j(U)$, let us define vectors $v^{j,r} \in \mathbb{C}^s$ with entries

$$v_k^{j,r} := F_d[d^j(U)]_{r-kd/s}, \quad k \in [s],\ r \in [d/s].$$

Note that the vectors $v^{j,r}$ can be viewed as the entries of $F_d[d^j(U)]_\ell$ with $\ell \bmod d/s = r$. Thus, the recovery of $d^j(U)$ is equivalent to the reconstruction of all $v^{j,r}$, $r \in [d/s]$.
For a fixed $r \in [d/s]$ the application of Theorem 3.6.45 for $Y^q = \mathcal{A}_s^q(\mathbb{T}_{\delta,s}(xx^*)) + N^q$, $q \in [Q]$ yields

$$(F_m^{-1}Y^q F_{d/s})_{j,r} = \frac{1}{s}\sum_{k\in[s]} v_k^{j,r}\overline{F_d[\overline{w}^{(q)} \circ S_j w^{(q)}]}_{r-kd/s} + (F_m^{-1}N^q F_{d/s})_{j,r}.$$

Hence, by defining the vectors $b^{j,r}, n^{j,r} \in \mathbb{C}^Q$ as

$$b_q^{j,r} = (F_m^{-1}Y^q F_{d/s})_{j,r} \text{ and } n_q^{j,r} = (F_m^{-1}N^q F_{d/s})_{j,r}, \quad q \in [Q],$$

respectively, and the matrices $M^{j,r} \in \mathbb{C}^{Q\times s}$ with the entries

$$M_{q,k}^{j,r} = \overline{F_d[\overline{w}^{(q)} \circ S_j w^{(q)}]}_{r-kd/s}, \tag{3.121}$$

we end up with the linear system

$$b_q^{j,r} = \frac{1}{s}M^{j,r}v^{j,r} + n^{j,r}. \tag{3.122}$$

Remark 3.6.46. *In the case* $\left\lfloor\frac{\delta-j}{s}\right\rfloor = 0$, *the diagonals* $d^j(U)$ *contain zero entries by construction of the space* $\mathbb{T}_{\delta,s}$ *(see Figure 3.4a below). Consequently, the linear system* (3.122) *should be considered with respect to these non-zero entries of* $d^j(U)$ *instead of the vectors* $v^{j,r}$. *Alternatively, one may discard these diagonals in reconstruction and reduce the requirement on* m *to* $m \geq 2\delta - s$.

If for all $j \in [\delta]$, $r \in [d/s]$, the matrices $M^{j,r}$ are injective, we can obtain the least squares solutions

$$u^{j,r} := s(M^{j,r})^\dagger b_q^{j,r}, \quad j \in [\delta],\ r \in [d/s]. \tag{3.123}$$

Then, we construct a matrix $Z \in \mathbb{T}_{\delta,s}$ by its diagonals $d^j(Z)$, $j \in [\delta]$. Their Fourier coefficients $F_d[d^j(Z)]$ are given by

$$F_d[d^j(Z)]_{r-kd/s} = u_k^{j,r}, \quad k \in [s],\ r \in [d/s],\ j \in [\delta]. \tag{3.124}$$

The resulting matrix Z is an approximation of $T_{\delta,s}(xx^*)$ and the distance between these two matrices is bounded by the next corollary.

Corollary 3.6.47. *Let $m \geq 2\delta - 1$ and $Q \geq s$. For each $q \in [Q]$ consider the ptychographic measurements Y^q of the form (PTY$_s$) with window $w^{(q)}$, $\mathrm{supp}(w^{(q)}) \subseteq [\delta]$ and the set of shifts $\mathcal{R}_s$ as in (3.111). Assume that*

$$\sigma_{min} := \sqrt{\frac{m}{s}} \min_{j\in[\delta], r\in[d/s]} \sigma_s(M^{j,r}) \tag{3.125}$$

satisfies $\sigma_{min} > 0$. Then, for the matrix Z constructed via vectors $u^{j,r}$, $j \in [\delta]$, $r \in [d/s]$, which are defined in (3.123), we have

$$\|Z - T_{d,\delta}(xx^*)\|_F \leq \sigma_{min}^{-1} \left[\sum_{q\in[Q]} \|N^q\|_F^2\right]^{1/2}.$$

Proof. The proof follows the steps of the proof of Corollary 3.6.6. Let us denote $X := T_{d,\delta}(xx^*)$. The equation (3.40) yields

$$\|Z - X\|_F^2 = \left\|d^0(Z) - d^0(X)\right\|_2^2 + \sum_{j=1}^{\delta} 2\left\|d^j(Z) - d^j(X)\right\|_2^2.$$

By construction of Z, the Fourier coefficients of each diagonal is a collection of vectors $u^{j,r}$ and, hence, we have

$$\begin{aligned}
\left\|d^j(Z) - d^j(X)\right\|_2^2 &= \frac{1}{d}\left\|F_d[d^j(Z)] - F_d[d^j(X)]\right\|_2^2 = \frac{1}{d}\sum_{\ell\in[d]} |F_d[d^j(Z)]_\ell - F_d[d^j(X)]_\ell|^2 \\
&= \frac{1}{d}\sum_{r\in[d/s]}\sum_{k\in[s]} |F_d[d^j(Z)]_{r-kd/s} - F_d[d^j(X)]_{r-kd/s}|^2 \\
&= \frac{1}{d}\sum_{r\in[d/s]}\sum_{k\in[s]} |u_k^{j,r} - v_k^{j,r}|^2 = \frac{1}{d}\sum_{r\in[d/s]} \left\|u^{j,r} - v^{j,r}\right\|_2^2
\end{aligned}$$

Since $\sqrt{m/s}\sigma_s(M^{j,r}) \geq \sigma_{min} > 0$, each $M^{j,r}$ is injective and $(M^{j,r})^\dagger M^{j,r} = I_s$. Therefore, the equation (3.123) further simplifies to

$$u^{j,r} := s(M^{j,r})^\dagger b_q^{j,r} = v^{j,r} + s(M^{j,r})^\dagger n^{j,r}.$$

Consequently, we obtain

$$\begin{aligned}\left\|d^j(Z)-d^j(X)\right\|_2^2 &= \frac{s^2}{d}\sum_{r\in[d/s]}\left\|(M^{j,r})^\dagger n^{j,r}\right\|_2^2 \le \frac{s^2}{d}\sum_{r\in[d/s]}\left\|(M^{j,r})^\dagger\right\|_\infty^2\left\|n^{j,r}\right\|_2^2\\ &= \frac{s^2}{d}\sum_{r\in[d/s]}\sigma_s^{-2}(M^{j,r})\sum_{q\in[Q]}|(F_m^{-1}N^qF_{d/s})_{j,r}|^2\\ &\le \frac{ms}{d}\sigma_{min}^{-2}\sum_{q\in[Q]}\left\|(F_m^{-1}N^qF_{d/s})_{(j)}\right\|_2^2\\ &= \sigma_{min}^{-2}\sum_{q\in[Q]}\left\|(\sqrt{m}F_m^{-1}N^q)_{(j)}\right\|_2^2.\end{aligned}$$

Then, similarly to (3.41), the proof is concluded by using the symmetry of the discrete Fourier transform for real vectors,

$$\begin{aligned}\|Z-X\|_F^2 &\le \sigma_{min}^{-2}\sum_{q\in[Q]}\left[\left\|(\sqrt{m}F_m^{-1}N^q)_{(0)}\right\|_2^2+\sum_{j=1}^{\delta}2\left\|(\sqrt{m}F_m^{-1}N^q)_{(j)}\right\|_2^2\right]\\ &= \sigma_{min}^{-2}\sum_{q\in[Q]}\left[\left\|(\sqrt{m}F_m^{-1}N^q)_{(0)}\right\|_2^2+\sum_{j=1}^{\delta}\left\|(\sqrt{m}F_m^{-1}N^q)_{(j)}\right\|_2^2+\left\|\overline{(\sqrt{m}F_m^{-1}N^q)_{(m-j)}}\right\|_2^2\right]\\ &\le \sigma_{min}^{-2}\sum_{q\in[Q]}\left\|\sqrt{m}F_m^{-1}N^q\right\|_F^2=\sigma_{min}^{-2}\sum_{q\in[Q]}\|N^q\|_F^2.\end{aligned}$$

□

We note that the recovery of Z from the measurements $Y=\{Y^q\}_{q\in[Q]}$ corresponds to the application of the pseudoinverse operator $Z=\{\mathcal{A}_s\big|_{\mathbb{T}_{\delta,s}}(Y^q)\}_{q\in[Q]}^\dagger$and σ_{min} is precisely its smallest non-trivial singular value.

Just as in the case $s=1$, the choice of Q and $w^{(q)}$, $q\in[Q]$, is crucial for the stability of the reconstruction. Since $M^{j,r}$ is the $Q\times s$ matrix and has to be injective, the number of windows Q should satisfy $Q\ge s$. For example, let us consider a class of exponential windows in analogy to Proposition 3.6.7, for which we guarantee that U can be recovered from the measurements $\{Y^q\}_{q\in[Q]}$.

Lemma 3.6.48. *Let $Q=s$ and consider windows $w^{(q)}$, $q\in[s]$, of the form*

$$w_k^{(q)}=e^{-k\alpha_q}\mathcal{I}_{k\in[\delta]}=\begin{cases}e^{-k\alpha_q}, & k\in[\delta]\\ 0, & \text{otherwise,}\end{cases}$$

with parameters α_q satisfying

$$\alpha_{q+1}-\alpha_q\ge\frac{\log 2}{2},\quad q\in[s-1], \tag{3.126}$$

$$\alpha_0\ge\frac{s-1}{s}\alpha_{s-1}+\frac{3}{4s}\log s+\frac{(s+1)}{2s}\log 2 \tag{3.127}$$

Then, U can be uniquely reconstructed from the noiseless system of linear equations (3.122).

The proof can be found in Appendix A. For symmetric windows, however, the inversion step fails just as for $s = 1$.

Example 3.6.49. *Let d be even. Consider windows $w^{(q)}$, $q \in [Q]$, satisfying the following symmetry condition*

$$w_k^{(q)} = \overline{w}_{\delta-k-1}^{(q)}, \quad k \in [\delta].$$

In Example 3.6.9 we proved that

$$F_d[\overline{w}^{(q)} \circ S_j w^{(q)}]_{d/2} = 0,$$

for $j \in [\delta]$ such that $\delta + j$ is even. Then, for $r = d/2 \bmod d/s$ the entries of $M^{j,r}$ corresponding to the column $k = s - \lfloor s/2 \rfloor$ and an arbitrary row $q \in [Q]$ is given by

$$\begin{aligned} M^{j,r}_{q,s-\lfloor s/2 \rfloor} &= \overline{F_d[\overline{w}^{(q)} \circ S_j w^{(q)}]}_{d/2 \bmod d/s - (s-\lfloor s/2 \rfloor)d/s} \\ &= \overline{F_d[\overline{w}^{(q)} \circ S_j w^{(q)}]}_{d/2 \bmod d/s + \lfloor (d/2)/(d/s) \rfloor)d/s - d} \\ &= \overline{F_d[\overline{w}^{(q)} \circ S_j w^{(q)}]}_{d/2-d} = \overline{F_d[\overline{w}^{(q)} \circ S_j w^{(q)}]}_{d/2} = 0, \end{aligned}$$

where we used that the indices are understood modulo d. Consequently, $M^{j,r}$ has a zero column and the linear system (3.122) is underdetermined.

Therefore, the truncation and the subspace completion procedures can be adapted for the sets $\mathcal{R}_s$ with $s > 1$ (see Section 3.6.2.2).

For the computational complexity of the inversion step via (3.123) we sum up the complexities of the performed operations. For the each $q \in [Q]$, the matrix $F_m^{-1} Y^q F_{d/s}$ is computed via the fast Fourier transform, which gives a complexity of

$$\mathcal{O}(Q(d/s)m \log m + Qm(d/s)\log(d/s)) = \mathcal{O}(\tfrac{Qdm}{s} \log d).$$

The construction of the matrices $M^{j,r}$ is performed by computing the fast Fourier transform δQ times, once for each diagonal $d^j(Z)$, $j \in [\delta]$, and each window $w^{(q)}$, $q \in [Q]$, which gives $\mathcal{O}(\delta Q d \log d)$ operations. Furthermore, the computation of all pseudoinverses $(M^{j,r})^\dagger$ requires $\mathcal{O}(\delta(d/s)Q^2 s) = \mathcal{O}(\delta d Q^2)$ operations. Finally, the construction of the diagonals $d^j(Z)$ from $u^{j,r}$ is again performed via the fast Fourier transform with a total complexity of $\mathcal{O}(\delta d \log d)$. Therefore, the inversion step requires

$$\mathcal{O}(\tfrac{Qdm}{s} \log d + \delta Q d \log d + \delta d Q^2 + \delta d \log d) = \mathcal{O}(Qd(m/s \log d + \delta \log d + \delta Q))$$

operations. If parameters m, s, Q are of order $\mathcal{O}(\delta)$, then the complexity further simplifies to $\mathcal{O}(\delta^2 d \max\{\log d, \delta\})$, which is slightly worse than $\mathcal{O}(\delta d \log d)$ for $s = 1$.

Finally, we combine the established generalization of the inversion step for multiple windows with the magnitude and the phase estimation steps.

Algorithm 8: Block Phase Retrieval for shifts $\mathcal{R}_s$ and multiple windows

Input : Shift size $1 \leq s < \delta$, s divisor of d and δ. For each window $w^{(q)}, q \in [Q]$, ptychographic measurements $Y^q \in \mathbb{R}^{m \times d/s}$ of the form (PTY_s) with $\mathcal{R}_s$ as in (3.111).

Output: $z \in \mathbb{C}^d$ with $z \approx e^{-i\theta}x$ for some $\theta \in [0, 2\pi)$.

1. Construct $u^{j,r}$ for all $j \in [\delta], r \in [d/s]$ via (3.123).
2. Construct the Fourier coefficients of the diagonals $d^j(Z)$, $j \in [\delta]$ via (3.124).
3. Construct $Z \in \mathbb{T}_{d,\delta,s}$ by its diagonals $d^j(Z)$ as an estimate of $T_{d,\delta,s}(xx^*)$.
4. Estimate the magnitudes $v \in \mathbb{R}^d$ from Z by a method of choice.
5. Estimate the phases $u \in \mathbb{C}^d$ from Z by a method of choice.
6. Set $z = v \circ u \in \mathbb{C}^d$ to form $z \approx x$.

3.6.5.3 Magnitude and phase estimation methods for $\mathbb{T}_{\delta,s}$

In this subsection, we assume that $Z \in \mathbb{T}_{\delta,s}$, the outcome of the inversion step, is given and the goal is to apply suitable magnitude and phase estimation techniques. We recall that the structure of the space $\mathbb{T}_{\delta,s}$ is different from $\mathbb{T}_{\delta,1}$ and, thus, the magnitude and phase estimation methods, which were designed for $\mathbb{T}_{\delta,1}$, have to be adapted for $\mathbb{T}_{\delta,s}$.
We will often refer to the index set

$$\mathcal{K}_s = \{(k,\ell) \in [d]^2 : \text{there exists } r \in [d/s] \text{ such that both } k, \ell \in \{rs, \dots, rs + \delta - 1\}\},$$

describing the non-zero entries of matrices in $\mathbb{T}_{\delta,s}$ (see the proof of Theorem 3.6.40) and $T_{\delta,s}(\mathbb{1}_{d\times d})$, which provides an alternative characterization of the set $\mathcal{K}_s$ as $(k,\ell) \in \mathcal{K}_s$ if and only if $T_{\delta,s}(\mathbb{1}_{d\times d})_{k,\ell} = 1$.
The first approach discards some of the entries in Z and reduces the problem to the case $s = 1$.

Lemma 3.6.50. *Consider the spaces $\mathbb{T}_{\delta,s}$ defined in (3.113). Then,*

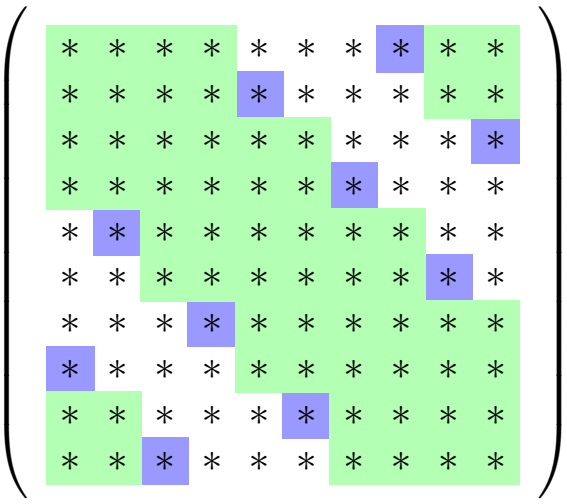
$$\mathbb{T}_{\delta-s+1,1} \subseteq \mathbb{T}_{\delta,s} \subseteq \mathbb{T}_{\delta,1}.$$

Proof. Since all spaces are described by pairs $(k,\ell) \in [d]^2$ corresponding to the non-zero entries of matrices in $\mathbb{T}_{\alpha,\beta}$, we can directly work with these pairs.
Let us start with the inclusion $\mathbb{T}_{\delta-s+1,1} \subseteq \mathbb{T}_{\delta,s}$. By (3.30) and (3.114), it is sufficient to show that for $k, \ell \in [d]$ such that $|k-\ell|_c < \delta - s + 1$ we have $(k,\ell) \in \{rs, \dots, rs+\delta-1\}^2$ for some $r \in [d/s]$. Without loss of generality let $\ell = k + j$ with $0 \leq j \leq \delta - s$. Then, for $r = \lfloor k/s \rfloor$ we have

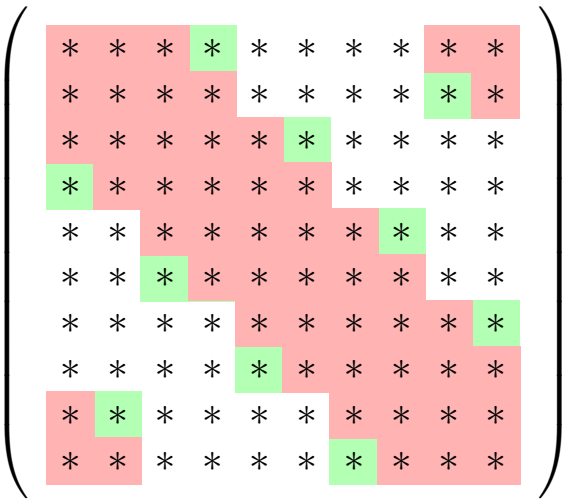
$$rs \leq k \leq \ell = k + j \leq \lfloor k/s \rfloor s + s - 1 + \delta - s = rs + \delta - 1,$$

and, thus, $(k,\ell) \in \mathcal{K}_s$.
For the inclusion $\mathbb{T}_{\delta,s} \subseteq \mathbb{T}_{\delta,1}$ we have that if a pair $(k,\ell) \in \mathcal{K}_s$, then, by (3.113), it satisfies $(k,\ell) \in \{rs, \dots, rs+\delta-1\}^2$ for some $r \in [d/s]$ and, thus, $|k-\ell|_c < \delta$. □

Hence, Z can be projected onto $\mathbb{T}_{\delta-s+1,1}$. Then, the magnitudes and phases are estimated from $T_{\delta-s+1,1}(Z)$ as in Sections 3.6.3 and 3.6.4.



(a) In green are the entries forming the space $\mathbb{T}_{4,2}$ and in blue are the entries, which form $\mathbb{T}_{4,1}$ but not $\mathbb{T}_{4,2}$. Note that the entries in blue are zero entries in corresponding diagonals of $T_{4,2}(xx^*)$.



(b) In red are the entries forming the space $\mathbb{T}_{3,1}$ and in green are the entries, which form in $\mathbb{T}_{4,2}$ but not $\mathbb{T}_{3,1}$.

Figure 3.4: Examples of space inclusions derived in Lemma 3.6.50 for $d = 10$.

As an alternative to the projection approach, the algorithms can be adapted specifically for $\mathbb{T}_{\delta,s}$ to avoid the loss of information.
For Diagonal Magnitude Estimation (Section 3.6.3.1), we observe that the main diagonal is always present in $T_{\delta,s}$. More precisely, by applying Lemma 3.6.50 for $s = 2$, we have $\mathbb{T}_{\delta-1,1} \subseteq \mathbb{T}_{\delta,1}$. Repeating the argument for smaller δ gives us $\mathbb{T}_{\alpha,1} \subseteq \mathbb{T}_{\beta,1}$ for $1 \leq \alpha \leq \beta \leq d$. Therefore, by $\mathbb{T}_{1,1} \subseteq \mathbb{T}_{\delta-s+1,1} \subseteq \mathbb{T}_{\delta,s}$, the main diagonal is present in Z and Diagonal Magnitude Estimation is applicable in Algorithm 8.
For Block Magnitude Estimation (Section 3.6.3.2), the condition (3.50) on the family of index sets $\{\mathcal{J}_j\}_{j\in P}$ has to be replaced by

$$(k, \ell) \in \mathcal{K}_s \text{ for all } k, \ell \in \mathcal{J}_j \text{ and all } j \in [P]. \tag{3.128}$$

Then, the following construction of the index sets is valid.

Lemma 3.6.51. *Consider a family of index sets $\{\mathcal{J}_{rs}^{\delta}\}_{r\in[d/s]}$ with $\mathcal{J}_j^{\gamma}$ defined in (3.56). Then, the conditions (3.128) and (3.54) are satisfied and the counts μ_k given by (3.52) satisfy $\mu_k = \delta/s$, $k \in [d]$. Hence, the Block Magnitude Estimation with $\{\mathcal{J}_{rs}^{\delta}\}_{r\in[d/s]}$ can be used in Algorithm 8.*

Proof. Let $k, \ell \in \mathcal{J}_{rs}^{\delta}$ for some $r \in [d/s]$. Then, $(k, \ell) \in \{rs, \dots, rs + \delta - 1\}^2$ and, thus, $(k, \ell) \in \mathcal{K}_s$. Hence, the condition (3.128) is satisfied. Since all sets $\mathcal{J}_{rs}^{\delta}$ are obtained by shifting $[\delta] = \mathcal{J}_0^{\delta}$ by s, for each $k \in [d]$ there are at most δ/s sets which include k, so that $\mu_k \leq \delta/s$. Let $r_1 = \lfloor k/s \rfloor - \delta/s + 1$ and $r_2 = \lfloor k/s \rfloor$. Then for all $r \in \{r_1, r_1 + 1, \dots, r_2\}$ we have that

$$rs \leq r_2 s = \lfloor k/s \rfloor s \leq k \leq \lfloor k/s \rfloor s + s - 1 = r_1 s + \delta - 1 \leq rs + \delta - 1,$$

and, thus, $\mu_k \geq \delta/s$. Hence, $\mu_k = \delta/s$. We recall that the condition (3.54) is equivalent to showing $\mu_k > 0$, $k \in [d]$. □

Turning to Log Magnitude Estimation (Section 3.6.3.3), we adjust the construction of the matrix $B \in \mathbb{R}^{d\delta \times d}$ and the vector $b(X) \in \mathbb{R}^{d\delta}$ to exclude all pairs $(k,\ell) \notin \mathcal{K}_s$. Therefore, we define

$$B_{(k,j),\ell} := \begin{cases} 2, & j = 0 \text{ and } k = \ell, \\ 1, & j \neq 0,\ k = \ell,\ (k, k-j) \in \mathcal{K}_s \\ 1, & j \neq 0,\ k - j = \ell,\ (k, k-j) \in \mathcal{K}_s \\ 0, & \text{otherwise}, \end{cases} \tag{3.129}$$

and

$$b(X)_{(k,j)} = \begin{cases} \log |X_{k,k-j}|, & (k, k-j) \in \mathcal{K}_s, \\ 0, & (k, k-j) \notin \mathcal{K}_s, \end{cases}$$

for all $k, \ell \in [d], j \in [\delta]$. We note that we artificially introduce zero rows $B_{((k,j))}$ and zero measurements $b_{(k,j)}$ whenever $(k, k-j) \notin \mathcal{K}_s$ to preserve the same indexing. In this way, a part of the proofs for the case $s > 1$ is analogous to $s = 1$. With the new definitions, we can show the invertibility of B^*B similarly to Theorem 3.6.19.

Theorem 3.6.52. *Let $d \geq 2\delta - s$. Consider the matrix B defined in (3.129). Then, the matrix B^*B admits the decomposition*

$$B^*B = U^* \operatorname{diag}(z) U,$$

with the unitary matrix

$$U = \frac{\sqrt{s}}{\sqrt{d}} F_{d/s} \otimes \frac{1}{\sqrt{s}} F_s = \frac{1}{\sqrt{d}} \begin{bmatrix} F_s & F_s & \dots & F_s \\ F_s & e^{-\frac{2\pi i}{d/s}} F_s & \dots & e^{-\frac{2\pi i (d/s-1)}{d/s}} F_s \\ \vdots & \vdots & \ddots & \vdots \\ F_s & e^{-\frac{2\pi i (d/s-1)}{d/s}} F_s & \dots & e^{-\frac{2\pi i (d/s-1)(d/s-1)}{d/s}} F_s \end{bmatrix}, \tag{3.130}$$

and the vector $z \in \mathbb{R}^d$ containing the eigenvalues

$$z_{ks+\ell} = \begin{cases} 2\delta - s + 2, & \ell \neq 0 \\ 4\delta - 2s + 2, & \ell = 0,\ k = 0, \\ 2\delta - s + 2 + s \frac{\sin\left(\frac{\pi(2\delta - s)k}{d}\right)}{\sin(\pi s k / d)}, & \ell = 0,\ k \in [d] \backslash \{0\}, \end{cases}$$

for indices $k \in [d/s]$, $\ell \in [s]$. The entries of z satisfy

$$z_k \geq 2s + 2, \text{ for all } k \in [d],$$

and, consequently, the inverse is given by

$$(B^*B)^{-1} = U^* \operatorname{diag}(1/z) U.$$

We recall that the proof of Theorem 3.6.19 is based on the decomposition of $T_{\delta,1}(\mathbb{1}_{d\times d})$ as a circulant matrix. The next lemma provides a similar decomposition for $T_{\delta,s}(\mathbb{1}_{d\times d})$.

Lemma 3.6.53 ([159, Lemma 1]). *The matrix $T_{\delta,s}(\mathbb{1}_{d\times d}) \in \mathbb{H}^d$ has rank d/s and can be decomposed as*

$$T_{\delta,s}(\mathbb{1}_{d\times d}) = U^*\Lambda U,$$

with the unitary matrix U as in (3.130) and a diagonal matrix $\Lambda \in \mathbb{R}^{d\times d}$. The entries with index $ks+\ell$, $k \in [d/s]$, $\ell \in [s]$ on the main diagonal of Λ are given by

$$\Lambda_{ks+\ell,ks+\ell} = \begin{cases} s\left[1 + 2\sum_{j=1}^{\delta/s-1} \cos\left(\frac{2\pi jk}{d/s}\right)\right], & \ell = 0, \\ 0, & \ell \neq 0, \end{cases}$$

Proof. In the proof of Theorem 3.6.40, we have shown that $T_{\delta,s}(\mathbb{1}_{d\times d}) \in \mathbb{T}_{\delta,s} \cap \mathbb{C}_s^{d\times d}$ with $[T_{\delta,s}(\mathbb{1}_{d\times d})]^\sim = T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})$, so that

$$T_{\delta,s}(\mathbb{1}_{d\times d}) = T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s}) \otimes \mathbb{1}_{s\times s}.$$

For $T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})$ in the proof of Theorem 3.6.52, we obtained the equation (3.66), which reads as

$$T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s}) = \frac{\sqrt{s}}{\sqrt{d}} F_{d/s}^* \operatorname{diag}(c) \frac{\sqrt{s}}{\sqrt{d}} F_{d/s},$$

with

$$c_k = 1 + 2\sum_{j=1}^{\delta/s-1} \cos\left(\frac{2\pi jk}{d/s}\right), \quad k \in [d/s].$$

The matrix $\mathbb{1}_{s\times s}$ satisfies

$$\mathbb{1}_{s\times s} = \mathbb{1}_s \mathbb{1}_s^* = \frac{1}{\sqrt{s}} F_s^* \operatorname{diag}(se_0) \frac{1}{\sqrt{s}} F_s,$$

where e_0 is the first standard basis vector. Then, by Proposition 2.1.2, we have

$$T_{\delta,s}(\mathbb{1}_{d\times d}) = \left(\frac{\sqrt{s}}{\sqrt{d}} F_{d/s} \otimes \frac{1}{\sqrt{s}} F_s\right)^* (\operatorname{diag}(c) \otimes \operatorname{diag}(se_0)) \left(\frac{\sqrt{s}}{\sqrt{d}} F_{d/s} \otimes \frac{1}{\sqrt{s}} F_s\right) = U^*\Lambda U.$$

□

Proof of Theorem 3.6.52. The proof is derived by repeating the proof of Theorem 3.6.19. The entries of the matrix B^*B are computed analogously to (3.63) and (3.64) as

$$(B^*B)_{\ell,\ell} = 4 + \sum_{\substack{j\in[\delta]\setminus\{0\},\\ (\ell,\ell-j)\in\mathcal{K}_s}} 2 = 3 + \sum_{k\in[d]} \mathcal{I}_{(\ell,k)\in\mathcal{K}_s}, \qquad \ell \in [d]$$

$$(B^*B)_{\ell,k} = \mathcal{I}_{(\ell,k)\in\mathcal{K}_s}, \qquad \ell, k \in [d], \ell \neq k.$$

In view of $\mathcal{I}_{(\ell,k)\in\mathcal{K}_s} = T_{\delta,s}(\mathbb{1}_{d\times d})_{\ell,k}$, we obtain

$$(B^*B)_{\ell,\ell} = 3 + \sum_{k\in[d]} T_{\delta,s}(\mathbb{1}_{d\times d})_{\ell,k} \text{ and } (B^*B)_{\ell,k} = T_{\delta,s}(\mathbb{1}_{d\times d})_{\ell,k}.$$

Moreover, if $d \geq 2\delta - s$, in the proof of Theorem 3.6.40 we have shown that for a fixed $\ell \in [d]$ there are precisely $2\delta - s$ values of index $k \in [d]$, such that $T_{\delta,s}(\mathbb{1}_{d\times d})_{\ell,k} = 1$. Therefore, the diagonal entries are given by

$$(B^*B)_{\ell,\ell} = 3 + 2\delta - s = (2\delta - s + 2)(I_d)_{\ell,\ell} + T_{\delta,s}(\mathbb{1}_{d\times d})_{\ell,\ell}.$$

and

$$B^*B = (2\delta - s + 2)I_d + T_{\delta,s}(\mathbb{1}_{d\times d}).$$

Consequently, using Lemma 3.6.53 and $U^*U = I_d$, we rewrite B^*B as

$$B^*B = U^*[(2\delta - s + 2)I_d]U + U^*\Lambda U = U^* \operatorname{diag}(z)U.$$

The definition of Λ and the equation (3.67) yield the formula for the entries of z. For all $k \in [d/s]$, $\ell \in [s]\backslash\{0\}$, we have $z_{ks+\ell} = 2\delta - s + 2 > 0$. If $\ell = 0$, we obtain

$$\begin{aligned} z_{ks} &= 2\delta - s + 2 + s\left[1 + 2\sum_{j=1}^{\delta/s-1} \cos\left(\frac{2\pi jk}{d/s}\right)\right] \\ &\geq 2\delta - s + 2 + s[1 - 2(\delta/s - 1)] = 2\delta - s + 2 - 2\delta + 3s = 2 + 2s. \end{aligned}$$

Hence, B^*B is invertible and its inverse is given by $U^* \operatorname{diag}(1/z)U$. □

Consequently, Log Magnitude Estimation can be used in Algorithm 8.
At last we turn to the phase synchronization (Section 3.6.4). We note, that the phase synchronization was posed as a problem on a weighted graph without a specified construction of the edge set. The necessary condition for the phase synchronization to possess a unique solution is that the graph is connected, which is equivalent to the spectral gap being greater than zero and does not depend the chosen weights. Hence, in the following, we show that for $\mathbb{T}_{\delta,s}$ the corresponding unweighted graph is connected and, in addition, we provide a lower bound for its spectral gap.
For the phase estimation within the BPR algorithm, according to (3.106) the edge set E is given by

$$E = \{(k,\ell) \in [d]^2 : \ |Z_{k,\ell}| > 0 \text{ and } k \neq \ell\}.$$

If $Z_{k,\ell} \neq 0$ for all $(k,\ell) \in \mathcal{K}_s$, $k \neq \ell$, then we have

$$E = \mathcal{K}_s\backslash\{(k,k) \in [d]^2 : k \in [d]\}, \tag{3.131}$$

or equivalently $A_G = T_{\delta,s}(\mathbb{1}_{d\times d}) - I_d$.

Lemma 3.6.54 ([159]). *Let $d \geq 4\delta$ and $\delta \geq 3s$ with $s > 1$. Assume that $Z_{k,\ell} \neq 0$ for all $(k,\ell) \in \mathcal{K}_s$, $k \neq \ell$, and consider an unweighted graph $G = ([d], E)$ with the edge set E as in (3.131). Then, the graph is connected and*

$$\tau_G \geq \min\left\{\frac{\pi^2\delta^3}{3d^2}, 2\delta - s\right\}.$$

Proof. The proof generalizes the proof of Lemma 3.6.35. Retracing the steps, we start with the adjacency matrix $A_G = T_{\delta,s}(\mathbb{1}_{d\times d}) - I_d$.
Since, $d \geq 4\delta \geq 2\delta - s$, we have shown in the proof of Theorem 3.6.40 that for a fixed $\ell \in [d]$ there are precisely $2\delta - s$ indices $k \in [d]$, such that $T_{\delta,s}(\mathbb{1}_{d\times d})_{\ell,k} = 1$. Therefore, for each row of A_G there are $2\delta - s - 1$ non-zero entries and, thus, each node has degree $2\delta - s - 1$. Thus, by the definition of the graph Laplacian we have

$$L_G = (2\delta - s - 1)I_d - A_G = (2\delta - s)I_d - T_{\delta,s}(\mathbb{1}_{d\times d}).$$

Then, similarly to the proof of Lemma 3.6.35 we have,

$$\tau_G = \lambda_{d-1}(L_G) - \lambda_d(L_G) = \lambda_1(T_{\delta,s}(\mathbb{1}_{d\times d})) - \lambda_2(T_{\delta,s}(\mathbb{1}_{d\times d})).$$

By Lemma 3.6.53, the eigenvalues of $T_\delta(\mathbb{1}_{d\times d})$ are either 0 or coincide with the eigenvalues of $T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})$ multiplied with s. Hence, we obtain

$$\tau_G = s\min\{\lambda_1(T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})) - \lambda_2(T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})), \lambda_1(T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})) - 0\}.$$

By (3.107), we have $\lambda_1(T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})) = 2\delta/s - 1$. Moreover, since $d/s \geq 4\delta/s$ and $\delta/s \geq 3$, Lemma 3.6.35 yields

$$\lambda_1(T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})) - \lambda_2(T_{d/s,\delta/s,1}(\mathbb{1}_{d/s\times d/s})) \geq \frac{\pi^2(\delta/s)^3}{3(d/s)^2}.$$

Then, the spectral gap is bounded from below by

$$\tau_G \geq \min\left\{\frac{\pi^2 s(\delta/s)^3}{3(d/s)^2}, s(2\delta/s - 1)\right\} = \min\left\{\frac{\pi^2\delta^3}{3d^2}, 2\delta - s\right\}.$$

□

Consequently, the phase synchronization can be used in Algorithm 8.

Notes and References. *The divisibility of δ by s is only required to simplify the proofs and assumptions in this section and to obtain a regular structure of $\mathbb{T}_{\delta,s}$. In general, it is not necessary.*
The adaptation of BPR to the equidistant shifts was previously done in several works [96, 138, 139, 141, 159]. In [96], the authors consider a version of BPR with Diagonal Magnitude Estimation and the phase synchronization for multiple windows similarly to the setup of Section 3.6.5.2 in the noiseless case. They show that under some assumptions on $w^{(q)}, q \in [Q]$, similar to the condition $\sigma_{min} > 0$ in Corollary 3.6.47, non-vanishing objects can be recovered from the measurements. In [138, 159] the authors consider recovery in the presence of noise. They present the magnitude and phase estimation methods (except of Log Magnitude Estimation) and show that they are applicable for $\mathbb{T}_{\delta,s}$. However, neither the authors of [96] nor [138, 159] provide a construction of windows for which $\sigma_{min} > 0$. In [139], we independently derived the results of [138] covered in Section 3.6.5.3 and, additionally, we provided the first possible construction of the windows (Lemma 3.6.48). We note that our results in [139] were based on the matrix multiplication-based proof similarly to [35] and in this thesis we reworked the proofs in Section 3.6.5.2 based on

results from [141]. We note that working with multiple windows may not be a practical solution as it requires at least s times more data to be obtained, stored and processed. The authors of [141] also considered equidistant shifts combined with additional assumptions on the object. In [141, Theorem 2] the object x is assumed to be bandlimited, so that $\operatorname{supp}(F_d x) \subseteq [\gamma]$ *with parameter γ satisfying $(d/s+1)/2 \leq \gamma \leq 2\delta - 1$. Comparing the lower and upper bounds, we observe that $d \leq s(4\delta - 3)$, which in general requires $s = \mathcal{O}(d/\delta)$. In view of $s < \delta$, this is quite restrictive. In comparison, in Section 3.6.5.1 we reconstruct d/s unknowns without any additional assumptions on s, δ, or d. Furthermore, Algorithm 7 can be used as a heuristics for objects in $\mathbb{C}^d$. Then, we recover x_s, a projection of x onto $\mathbb{C}^d_s$ from the measurements*

$$|Ax|^2 + n = |Ax_s + Ax_\perp|^2 + n = |Ax_s + n_1|^2 + n,$$

where $x_\perp = x - x_s$. However, the noise is not additive and the derived reconstruction guarantees do not apply.

3.6.6 Block Phase Retrieval, uniqueness and stability

In this section, we would like to briefly reflect on the reconstruction guarantees for BPR and their reinterpretation as the uniqueness and stability results discussed in Section 3.3. As the BPR algorithm can be viewed as a constructive proof of the uniqueness of the reconstruction. If we translate the statement of Theorem 3.3.8 for the case $\mathcal{R} = [d]$ into the context of BPR, the condition

$$[F_d(w \circ S_r \overline{w})]_j \neq 0 \text{ for } r = 0, 1,\ j \in [d]$$

implies that the diagonals $d^0(X)$ and $d^1(X)$ are recovered in the inversion step. In the absence of noise, the magnitudes of x are recovered with Diagonal Magnitude Estimation from $d^0(X)$. The phases of x are recovered via angular synchronization and the unique recovery is guaranteed by Theorem 3.6.29 for non-vanishing x as the underlying graph is a connected loop. Consequently, x is uniquely recovered from the noiseless ptychographic measurements, if in BPR for the magnitude and phase estimation the matrix $\mathbb{T}_2(X)$ is used instead of X. This is a weaker version of Theorem 3.6.2. Note that Theorem 3.6.2 requires the stronger assumption

$$[F_d(w \circ S_r \overline{w})]_j \neq 0 \text{ for } r \in [\delta],\ j \in [d].$$

These arguments extend to equidistant shifts $\mathcal{R} = \mathcal{R}_s$ with $s \geq 1$ for non-vanishing piecewise constant objects $x \in \mathbb{C}^d_s$ similarly to Corollary 3.6.44.
Turning to stability, let us consider objects $x, z \in \mathbb{C}^d_s$ with noiseless ptychographic measurements $|Ax|^2$ and $|Az|^2$ with A as in (3.9). By setting

$$y = |Az|^2 = |Ax|^2 + (|Az|^2 - |Ax|^2),$$

and applying BPR, the algorithm will reconstruct z. If a version of Corollary 3.6.44 is used with the magnitude estimation error replaced by Corollary 3.6.18, the reconstruction error is bounded from above by

$$\operatorname{dist}(x, z) \leq \left[24\frac{sd^2}{\delta^{5/2}} \|x\|_\infty + (1 + 2\sqrt{2})\sqrt{s}\right] \frac{1}{|x|^2_{\min}} \sigma^{-1} \left\| |Ax|^2 - |Az|^2 \right\|_2.$$

This can be interpreted as the local stability result on the set $\mathbb{C}_s^d$. However, we must note that the right-hand side depends on $\| |Ax|^2 - |Az|^2 \|_2$ and not $\| |Ax| - |Az| \|_2$. This quadratic dependency on x and z is compensated by the fraction $\|x\|_\infty / |x|_{\min}^2$, where the denominator is quadratic in x, while the nominator is only linear. We also note that these local stability results are similar to [87, Theorem 3.6] for bandlimited objects.

Chapter 4

Blind ptychography

In this section we consider blind ptychography, a more challenging version of ptychography discussed in the previous chapter. In ptychography, recovery of an unknown object $x \in \mathbb{C}^d$ from the measurements (PTY) given by

$$Y_{j,r} = I^m_{j,r} + N_{j,r} = |(F_m P_m[S_{-r}x \circ w])_j|^2 + N_{j,r}, \quad j \in [m], r \in \mathcal{R} \subseteq [d],$$

is considered under the assumption that a window $w \in \mathbb{C}^d$ with $\operatorname{supp}(w) = [\delta]$ is known. In practice, this assumption is not necessarily true and additionally the window has to be estimated as well, which constitutes the following reconstruction problem:

$$\boxed{\text{Reconstruct } x, w \in \mathbb{C}^d \text{ with } \operatorname{supp}(w) = [\delta] \text{ from data (PTY).}}$$

Recall that $\operatorname{supp}(w) = [\delta]$ and, thus, the search space can be reduced by considering the variable

$$\hat{w} := P_\delta w \text{ such that } w = P^*_\delta \hat{w},$$

where P_δ is the projection operator (2.10) on the first δ coordinates and P^*_δ is its adjoint. Let us rewrite the measurements (PTY) as a function of x and $\hat{w}$. First, we observe that

$$(P_m[S_{-r}x \circ w])_k = (P_m[S_{-r}x \circ P^*_\delta \hat{w}])_k = (S_{-r}x)_k (P^*_\delta \hat{w})_k = \begin{cases} (S_{-r}x)_k \hat{w}_k, & k \in [\delta], \\ 0, & k \in [m]\backslash[\delta], \end{cases}$$

and

$$(P^*_\delta[P_\delta S_{-r}x \circ \hat{w}])_k = \begin{cases} (P_\delta S_{-r}x \circ \hat{w})_k, & k \in [\delta], \\ 0, & k \in [m]\backslash[\delta], \end{cases} = \begin{cases} (S_{-r}x)_k \hat{w}_k, & k \in [\delta], \\ 0, & k \in [m]\backslash[\delta]. \end{cases}$$

Note that P^*_δ maps $\mathbb{C}^\delta$ to $\mathbb{C}^d$ in the first case and to $\mathbb{C}^m$ in the second case. Substituting these equalities to the intensity measurements Y yields

$$Y_{j,r} = |(F_m P_m[S_{-r}x \circ P^*_\delta \hat{w}])_j|^2 + N_{j,r} = |(F_m P^*_\delta[P_\delta S_{-r}x \circ \hat{w}])_j|^2 + N_{j,r} \qquad \text{(BPTY)}$$

for all $j \in [m]$, $r \in \mathcal{R}$. Consequently, the reconstruction problem can also be reformulated in terms of x and $\hat{w}$.

$$\boxed{\text{Reconstruct } x \in \mathbb{C}^d, \hat{w} \in \mathbb{C}^\delta \text{ from data (BPTY).}}$$

4.1 Ambiguities and uniqueness of blind ptychography

Similarly to the global phase factor ambiguity in ptychography, the blind ptychography admits unavoidable ambiguities. We summarize the known ambiguities arising in blind ptychography in the next statement.

Theorem 4.1.1 (General ambiguities arising in blind ptychography [182]). *Consider $x \in \mathbb{C}^d$, $\hat{w} \in \mathbb{C}^\delta$ and the corresponding ptychographic measurements* (BPTY). *Then,*

1. *(global phase ambiguity) for all $\alpha, \beta \in \mathbb{T}$ the pair $\alpha x, \beta \hat{w}$ produces the same measurements* (BPTY),
2. *(scaling ambiguity) for all $\gamma \in \mathbb{C}\backslash\{0\}$ the pair $\gamma x, \hat{w}/\gamma$ produces the same measurements* (BPTY),
3. *(linear phase ambiguity) for all $\rho \in \mathbb{R}$ the pair $z \in \mathbb{C}^d, \hat{v} \in \mathbb{C}^\delta$ with $z_k = e^{-i\rho k} x_k$, $k \in [d]$, and $\hat{v}_k = e^{i\rho k}\hat{w}_k$, $k \in [\delta]$, produces the same measurements* (BPTY).

Proof. 1. For all $\alpha, \beta \in \mathbb{T}$ we have

$$|(F_m P_m^*[P_\delta S_{-r} \alpha x \circ \beta \hat{w}])_j|^2 = |\alpha|^2|\beta|^2 \, |(F_m P_m^*[P_\delta S_{-r} x \circ \hat{w}])_j|^2 = |(F_m P_m^*[P_\delta S_{-r} x \circ \hat{w}])_j|^2 .$$

2. For all $\gamma \in \mathbb{C}\backslash\{0\}$ we obtain

$$|(F_m P_m^*[P_\delta S_{-r} \gamma x \circ \hat{w}/\gamma])_j|^2 = |(F_m P_m^*[P_\delta S_{-r} x \circ \hat{w}])_j \gamma/\gamma|^2 = |(F_m P_m^*[P_\delta S_{-r} x \circ \hat{w}])_j|^2 .$$

3. Let z and $\hat{v}$ be defined as in the statement of the theorem. Then, for all $r \in \mathcal{R}$, $k \in [\delta]$, we have

$$(P_\delta S_{-r} z \circ \hat{v})_k = z_{k+r}\hat{v}_k = e^{-i\rho(k+r)} x_{k+r} e^{i\rho k}\hat{w}_k = e^{-i\rho r} x_{k+r}\hat{w}_k = e^{-i\rho r}(P_\delta S_{-r} x \circ \hat{w})_k,$$

and, consequently,

$$|(F_m P_m^*[P_\delta S_{-r} z \circ \hat{v}])_j|^2 = \left|e^{-i\rho r}(F_m P_m^*[P_\delta S_{-r} x \circ \hat{w}])_j\right|^2 = |(F_m P_m^*[P_\delta S_{-r} x \circ \hat{w}])_j|^2 .$$

□

It is not known if the list of ambiguities provided by Theorem 4.1.1 is complete. However, other ambiguities may arise depending on the choice of the set $\mathcal{R}$.

Example 4.1.2 ([183, Proposition III.2]). *Let $\mathcal{R}$ be the set of equidistant shifts $\mathcal{R}_s$ defined by* (3.111) *with s being the divisor of d. Then, for any non-vanishing $\lambda \in \mathbb{C}^s$ the object-window pairs $x \in \mathbb{C}^d, \hat{w} \in \mathbb{C}^\delta$ and $z \in \mathbb{C}^d, \hat{v} \in \mathbb{C}^\delta$ with*

$$z_k := x_k \lambda_{k \bmod s}, \ k \in [d], \quad \text{and} \quad \hat{v}_k := \hat{w}_k \lambda^{-1}_{k \bmod s}, \ k \in [\delta],$$

produce the same measurements (BPTY). *More precisely, for all $r \in [d/s]$ and $k \in [\delta]$ we have*

$$\begin{aligned}(P_\delta S_{-rs} z \circ \hat{v})_k &= z_{k+rs}\hat{v}_k = x_{k+rs}\hat{w}_k \lambda_{k+rs \bmod s}\lambda^{-1}_{k \bmod s} \\ &= x_{k+rs}\hat{w}_k \lambda_{k \bmod s}\lambda^{-1}_{k \bmod s} = (P_\delta S_{-rs} x \circ \hat{w})_k.\end{aligned}$$

Example 4.1.3 (Shift ambiguity). *Let $\mathcal{R}$ be the set of equidistant shifts $\mathcal{R}_s$ defined by (3.111) with s being the divisor of d and let $\delta = d$. Then, for arbitrary $q \in [d]$ the pairs x, w and $S_q x, S_q w$ produce the same measurements, similarly to Example 3.3.3. Then, by Proposition 2.2.2, we have*

$$\begin{aligned}|(F_d[S_{-r}S_q x \circ S_q w])_j|^2 &= |(F_d S_q[S_{-r}x \circ w])_j|^2 = |(M_{-q}F_d[S_{-r}x \circ w])_j|^2 \\ &= |e^{-\frac{2\pi i q j}{d}}(F_d[S_{-r}x \circ w])_j|^2 = |(F_d[S_{-r}x \circ w])_j|^2.\end{aligned}$$

In view of these results, the unique recovery is understood to be the recovery of pairs x, w up to any combination of the listed above ambiguities. The recent study on the uniqueness of reconstruction for blind ptychography provides the following sufficient condition on the number of measurements.

Theorem 4.1.4 ([84, Theorem 2.3]). *Let $m = d$. Consider the ptychographic measurements (BPTY) with the set of equidistant shifts $\mathcal{R}_s$ defined by (3.111) and s being the divisor of d. If*

$$M = d|\mathcal{R}_s| \geq 3(2\delta - 1) + \left\lceil \frac{(4d/s + 1)(d - \delta - 2d/s)}{d/s} \right\rceil,$$

then the unique recovery is possible for all $x \in \mathbb{C}^d$, $\hat{w} \in \mathbb{C}^\delta$ except if $(x, \hat{w}) \in \mathcal{N}$ with $\mathcal{N}$ being a set of zero measure.

Notes and References. *The uniqueness and ambiguities of the blind ptychography are sparsely studied in the literature.*
It was derived that $M \geq 10(d + \delta)$ measurements are sufficient for the unique recovery [183]. Later, the previous results were improved in [84] to $M \geq 4d + 2\delta$ in Theorem 4.1.4. Ambiguities arising in blind ptychography are mostly studied in [183, 182]. In addition to the material covered in this section, in [182] the authors analyze the existence of further ambiguities and study how to choose the set $\mathcal{R}$, so that the additional ambiguities can be avoided, e.g., Example 4.1.2.

4.2 Alternating Amplitude Flow for blind ptychography

Turning to algorithms for blind ptychographic reconstruction, in this section we propose a version of the Amplitude Flow algorithm discussed in Section 3.5.1. Recall that AF is based on the minimization of the loss function $\mathcal{L}_2$. For blind ptychography, we will consider the loss function

$$\begin{aligned}\mathcal{G}(z, \hat{v}) &:= \sum_{r \in \mathcal{R}} \sum_{j \in [m]} \left| \sqrt{|(F_m P_m[S_{-r}z \circ P_\delta^* \hat{v}])_j|^2 + \varepsilon} - \sqrt{Y_{j,r} + \varepsilon} \right|^2 + \alpha_T \|z\|_2^2 + \beta_T \|\hat{v}\|_2^2 \\ &= \sum_{r \in \mathcal{R}} \sum_{j \in [m]} \left| \sqrt{|(F_m P_\delta^*[P_\delta S_{-r}z \circ \hat{v}])_j|^2 + \varepsilon} - \sqrt{Y_{j,r} + \varepsilon} \right|^2 + \alpha_T \|z\|_2^2 + \beta_T \|\hat{v}\|_2^2. \quad (4.1)\end{aligned}$$

with parameters $\varepsilon, \alpha_T, \beta_T \geq 0$. If the parameter $\varepsilon > 0$, the loss function is twice continuously differentiable, which allows us to compute its Wirtinger gradient and potentially apply the convergence theory derived in Section 2.3.

In the absence of noise, the true object-window pair $x, \hat{w}$ minimizes the first term in $\mathcal{G}$. Furthermore, if unique recovery is possible, all global minimizers of the first term in $\mathcal{G}$ are $x, \hat{w}$ up to ambiguities discussed in the previous section. This suggests that the minimization of $\mathcal{G}$ would lead to the recovery of x and $\hat{w}$. An additional inclusion of Tikhonov regularization, i.e., $\alpha_T, \beta_T > 0$, is beneficial in view of the scaling ambiguity and will be crucial later in this section.

We note that $\mathcal{G}$ behaves similarly to a forth order polynomial in its argument $(z, \hat{v})$. This implies that the Hessian matrix behaves as a second order polynomial in $(z, \hat{v})$ and obtaining the inequality (2.17) with a fixed constant L, as required in Theorem 2.3.4, is not possible in the most cases. Therefore, determining the learning rate for the gradient descent in both unknowns,

$$\begin{bmatrix} z^t \\ \hat{v}^t \end{bmatrix} = \begin{bmatrix} z^{t-1} \\ \hat{v}^{t-1} \end{bmatrix} - \mu_t \begin{bmatrix} \nabla_z \mathcal{G}(z^{t-1}, \hat{v}^{t-1}) \\ \nabla_{\hat{v}} \mathcal{G}(z^{t-1}, \hat{v}^{t-1}) \end{bmatrix},$$

becomes a complicated task and has to be performed in a non-trivial way. This, in general, involves multiple extra evaluations of the loss function, which is costly if the dimensions d, δ, m are large.

Instead we consider an alternating minimization approach to minimization of the loss $\mathcal{G}$ with a flavor of [184]. That is, instead of simultaneous optimization of $\mathcal{G}$ with respect both the object z and the window $\hat{v}$, only one of two unknowns is optimized, while the other remains to be fixed. Afterwards, their roles are swapped and the second unknown is optimized, while the first remains to be fixed. The interchanges between the unknowns are continued until the optimization with respect to either of unknowns provides no further improvement on the value of the loss function $\mathcal{G}$.

Algorithm 9: Alternating Amplitude Flow for blind ptychography (informal)

Input : Measurements Y as in (BPTY), number of iterations $T \in \mathbb{N}$, parameters $\varepsilon > 0$, $\alpha_T, \beta_T \geq 0$, initial guesses $z^0 \in \mathbb{C}^d$ and $\hat{v}^0 \in \mathbb{C}^\delta$.

Output: $z \in \mathbb{C}^d$ and $\hat{v} \in \mathbb{C}^\delta$.

for $t = 1, \ldots, T$ **do**

  z^t = Optimize $\mathcal{G}$ with respect to z while $\hat{v} = \hat{v}^{t-1}$ is fixed.

  $\hat{v}^t$ = Optimize $\mathcal{G}$ with respect to $\hat{v}$ while $z = z^t$ is fixed.

end

return $z = z^T, \hat{v} = \hat{v}^T$.

The alternating minimization technique is a popular technique in optimization community [185, 186, 187, 184] and it is sometimes used for blind ptychography. For instance, an alternating minimization version of Douglas-Rachford splitting is considered in [109]. In this section, we choose a gradient descent-based optimization for the optimization subproblems, which in both cases leads to a regularized version of the Amplitude Flow algorithm discussed in Section 3.5.1. Before stating the formal algorithm in Section 4.2.3, we formalize the optimization with respect to each of the variables in the next two sections.

4.2.1 Optimization with respect to the object

Let us fix $\hat{v}$ and consider $\mathcal{G}$ as a function of a single variable z. In this case we are able to make use of the theory developed in Section 3.5.1. Firstly, we recall some concepts from Section 3.5.1.
If $\hat{v}$ is fixed, we can rewrite the phychographic measurements (PTY) in the form of the phase retrieval measurements (PR) with the measurement matrix, the measurements and the noise

$$A_{\hat{v}} = \begin{bmatrix} F_m P_m \operatorname{diag}(P_\delta^* \hat{v}) S_{-r_1} \\ \vdots \\ F_m P_m \operatorname{diag}(P_\delta^* \hat{v}) S_{-r_R} \end{bmatrix}, \quad y = \begin{bmatrix} Y^{(r_1)} \\ \vdots \\ Y^{(r_R)} \end{bmatrix}, \quad n = \begin{bmatrix} N^{(r_1)} \\ \vdots \\ N^{(r_R)} \end{bmatrix}, \tag{4.2}$$

respectively, where $Y^{(r)}$ denotes the r-th column of the matrix Y. Then, the loss function $\mathcal{G}$ can be written as

$$\mathcal{G}(z, \hat{v}) = \mathcal{L}_{2,\varepsilon}(z; A_{\hat{v}}) + \alpha_T \|z\|_2^2 + \beta_T \|\hat{v}\|_2^2 = \mathcal{H}(z; A_{\hat{v}}, \alpha_T) + \beta_T \|\hat{v}\|_2^2, \tag{4.3}$$

with $\mathcal{L}_{2,\varepsilon}$ as in (3.14) and the supplementary loss function $\mathcal{H} : \mathbb{C}^b \to [0, +\infty)$ defined by the family of positive semidefinite matrices $\mathcal{Q} = \{Q_k\}_{k\in[M]} \subset \mathbb{H}^b$ and parameters $\varepsilon, \gamma \geq 0$ as

$$\mathcal{H}(t) = \mathcal{H}(t; \mathcal{Q}, \gamma) = \mathcal{L}_{2,\varepsilon}(t; \mathcal{Q}) + \gamma \|t\|_2^2, \quad \text{for all } t \in \mathbb{C}^b. \tag{4.4}$$

Furthermore, we use the notation $\mathcal{H}(t; A, \gamma)$ instead of $\mathcal{H}(t; \mathcal{Q}, \gamma)$, if $\mathcal{Q}$ corresponds to the phase retrieval measurements (PR) with a measurement matrix A. Note that in (4.3), the function $\mathcal{H}(z; A_{\hat{v}}, \alpha_T)$ includes all terms related to z and, therefore, minimization of $\mathcal{G}$ with respect to z is equivalent to minimization of $\mathcal{H}(z; A_{\hat{v}}, \alpha_T)$ and $\nabla_z \mathcal{G}(z, v) = \nabla_z \mathcal{H}(z; A_{\hat{v}}, \alpha_T)$.

Lemma 4.2.1. *Let $\varepsilon > 0$. The function $\mathcal{H}$ is twice continuously differentiable with the gradient given by*

$$\nabla_t \mathcal{H}(t) = \sum_{k\in[M]} \left(1 - \frac{\sqrt{y_k + \varepsilon}}{\sqrt{t^* Q_k t + \varepsilon}}\right) Q_k t + \gamma t,$$

and its Hessian matrix satisfies

$$\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{H}(t) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \leq \left[\left\| \sum_{k\in[M]} Q_k \right\|_\infty + \gamma \right] \left\| \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \right\|_2^2 \quad \textit{for all } z, u \in \mathbb{C}^b.$$

The proof follows by combining Lemma 3.5.1 with the next result.

Lemma 4.2.2. *Consider a real-valued quadratic function*

$$f_{M_1,M_2}(t) := t^* M_1 t + \operatorname{Re}\left(t^* M_2 \bar{t}\right) = t^* M_1 t + \tfrac{1}{2} t^* M_2 \bar{t} + \tfrac{1}{2} t^T \overline{M_2} t,$$

with Hermitian matrices $M_1, M_2 \in \mathbb{H}^b$. Then, for $t \in \mathbb{C}^b$ we have

$$\nabla_t f_{M_1,M_2}(t) = M_1^* t + \operatorname{Re}(M_2)\bar{t} \text{ and } \nabla^2 f_{M_1,M_2}(t) = \begin{bmatrix} M_1 & \operatorname{Re}(M_2) \\ \operatorname{Re}(M_2) & \overline{M_1} \end{bmatrix}.$$

Proof. By the definition of the Wirtinger derivatives, we have

$$\nabla_t f_{M_1,M_2}(t) = (t^* M_1 + t^T \tfrac{1}{2}(\overline{M}_2 + M_2^*))^* = M_1^* t + \tfrac{1}{2}(M_2 + M_2^T)\bar{t} = M_1^* t + \mathrm{Re}(M_2)\bar{t},$$

and

$$\nabla_{t,t} f_{M_1,M_2}(t) = M_1^* = M_1, \quad \nabla_{\bar{t},t} f_{M_1,M_2}(t) = \mathrm{Re}(M_2)^T = \mathrm{Re}(M_2).$$

Consequently, the Hessian matrix is constant with respect to z and according to (2.14) it is given by

$$\nabla^2 f_{M_1,M_2}(t) = \begin{bmatrix} M_1 & \mathrm{Re}(M_2) \\ \mathrm{Re}(M_2) & \overline{M_1} \end{bmatrix}.$$

□

Proof of Lemma 4.2.1. We note that

$$\|z\|_2^2 = z^* z = z^* I_d z = f_{I_d, O_{d\times d}}(z),$$

with $O_{d\times d}$ denoting the zero matrix. Then, by linearity of the gradient and Lemmas 3.5.1 and 4.2.2, we obtain

$$\nabla_z \mathcal{H}(z) = \nabla_z \mathcal{L}_{2,\varepsilon}(z) + \gamma \nabla_z f_{I_d, O_{d\times d}}(z) = \sum_{k\in[M]} \left(1 - \frac{\sqrt{y_k + \varepsilon}}{\sqrt{z^* Q_k z + \varepsilon}}\right) Q_k z + \gamma z.$$

Furthermore,

$$\begin{aligned} \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{H}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} &= \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{L}_{2,\varepsilon}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} + \gamma \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 f_{I_d, O_{d\times d}}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \\ &\le 2u^* \left[\sum_{k\in[M]} Q_k + \gamma I_d\right] u \le \left[\left\|\sum_{k\in[M]} Q_k\right\|_\infty + \gamma\right] \left\|\begin{bmatrix} u \\ \bar{u} \end{bmatrix}\right\|_2^2. \end{aligned}$$

□

Consequently, we can guarantee that the gradient descent applied to $\mathcal{G}$ for minimization with respect to z will not increase the loss function on each step for a proper choice of the learning rate.

Theorem 4.2.3. *Fix $\varepsilon > 0$. Let $z \in \mathbb{C}^d$, $\hat{v} \in \mathbb{C}^\delta$ be arbitrary. Consider the iteration*

$$z^+ = z - \mu \nabla_z \mathcal{G}(z, \hat{v}),$$

with the gradient

$$\nabla_z \mathcal{G}(z, \hat{v}) = A_{\hat{v}}^* \left[I_M - \mathrm{diag}\left(\frac{\sqrt{y_k + \varepsilon}}{\sqrt{|(A_{\hat{v}} z)_k|^2 + \varepsilon}}\right)\right] A_{\hat{v}} z + \alpha_T z, \tag{4.5}$$

and the learning rate $\mu = \mu(\mathcal{G}, z, \tau, \mu_c, N)$ determined by Algorithm 1 where μ_c satisfies

$$0 < \mu_c \le \left[m \max_{\ell\in[d]} \sum_{r\in\mathcal{R}} |(S_r P_\delta^* \hat{v})_\ell|^2 + \alpha_T\right]^{-1} =: L_{\hat{v}}^{-1}. \tag{4.6}$$

Then,

$$\mathcal{G}(z^+, \hat{v}) - \mathcal{G}(z, \hat{v}) \le -\mu \|\nabla_z \mathcal{G}(z, \hat{v})\|_2^2.$$

Proof. In view of (4.3), we have

$$\mathcal{G}(z,\hat{v}) = \mathcal{H}(z; A_{\hat{v}}, \alpha_T) + \beta_T \|\hat{v}\|_2^2 \quad \text{and} \quad \nabla_z \mathcal{G}(z,\hat{v}) = \nabla_z H(z; A_{\hat{v}}, \alpha_T).$$

By Corollary 3.5.2 and Lemma 4.2.1, the gradient $\nabla_z \mathcal{G}$ is given by (4.5). Moreover, the gradient descent applied to $\mathcal{G}$ is equivalent to the gradient descent applied to $\mathcal{H}(\,\cdot\,; A_{\hat{v}}, \alpha_T)$. Note that the family $\{Q_k^{\hat{v}}\}_{k\in[M]}$ corresponding to $A_{\hat{v}}$ is given by rank-one matrices $Q_k^{\hat{v}} = a_k^{\hat{v}}(a_k^{\hat{v}})^*$, $k \in [M]$. The vectors $a_k^{\hat{v}}$ are the conjugates of the rows of $A_{\hat{v}}$, so that $(A_{\hat{v}} z)_k = (a_k^{\hat{v}})^* z$. Therefore, Lemma 3.3.5 yields

$$\left\| \sum_{k\in[M]} Q_k^{\hat{v}} \right\|_\infty = \left\| \sum_{k\in[M]} a_k^{\hat{v}} (a_k^{\hat{v}})^* \right\|_\infty = \|A_{\hat{v}}^* A_{\hat{v}}\|_\infty = m \max_{\ell\in[d]} \sum_{r\in\mathcal{R}} |(S_r P_\delta^* \hat{v})_\ell|^2.$$

Hence, the choice of a constant learning rate μ_c satisfies the conditions of Theorem 2.3.6. Thus, using Lemma 4.2.1, we obtain

$$\mathcal{H}(z^+; A_{\hat{v}}, \alpha_T) - \mathcal{H}(z; A_{\hat{v}}, \alpha_T) \le -\mu \|\nabla_z \mathcal{H}(z; A_{\hat{v}}, \alpha_T)\|_2^2 .$$

The addition and subtraction of the term $\beta_T \|\hat{v}\|_2^2$ yields

$$\begin{aligned} \mathcal{G}(z^+,\hat{v}) - \mathcal{G}(z,\hat{v}) &= \mathcal{H}(z^+; A_{\hat{v}}, \alpha_T) + \beta_T \|\hat{v}\|_2^2 - \mathcal{H}(z; A_{\hat{v}}, \alpha_T) - \beta_T \|\hat{v}\|_2^2 \\ &\le -\mu \|\nabla_z \mathcal{H}(z; A_{\hat{v}}, \alpha_T)\|_2^2 = -\mu \|\nabla_z \mathcal{G}(z,\hat{v})\|_2^2 . \end{aligned}$$

□

4.2.2 Optimization with respect to the window

Now, let us fix z and consider $\mathcal{G}$ as a function of a single variable $\hat{v}$. Similarly to the previous section, the optimization with respect to $\hat{v}$ can be reduced to the phase retrieval problem of the form (PR). For this, we observe that the intensity measurements can be rewritten as

$$Y_{j,r} = |(F_m P_\delta^* [P_\delta S_{-r} z \circ \tilde{v}])_j|^2 + N_{j,r} = |(F_m P_\delta^* \operatorname{diag}(P_\delta S_{-r} z)\tilde{v}])_j|^2 + N_{j,r},$$

for all $j \in [m]$, $r \in \mathcal{R}$. With the measurement matrix

$$A_z = \begin{bmatrix} F_m P_\delta^* \operatorname{diag}(P_\delta S_{-r_1} z) \\ \vdots \\ F_m P_\delta^* \operatorname{diag}(P_\delta S_{-r_R} z) \end{bmatrix}, \tag{4.7}$$

the measurements y and the noise n as in (4.2), the recovery of $\hat{v}$ is again a phase retrieval problem. Returning to the loss function $\mathcal{G}$, we combine the terms involving $\hat{v}$ as

$$\mathcal{G}(z,\hat{v}) = \mathcal{L}_{2,\varepsilon}(\hat{v}; A_z) + \alpha_T \|z\|_2^2 + \beta_T \|\hat{v}\|_2^2 = \mathcal{H}(\hat{v}; A_z, \beta_T) + \alpha_T \|z\|_2^2 , \tag{4.8}$$

with the functions $\mathcal{L}_{2,\varepsilon}$ and $\mathcal{H}$ defined in (3.14) and (4.4), respectively. Consequently, first order minimization with respect to $\hat{v}$ is analogous to minimization with respect to z, which leads to the following result.

Theorem 4.2.4. *Fix $\varepsilon > 0$. Let $z \in \mathbb{C}^d$, $\hat{v} \in \mathbb{C}^\delta$ be arbitrary. Consider the iteration*

$$\hat{v}^+ = \hat{v} - \nu \nabla_{\hat{v}} \mathcal{G}(z, \hat{v}),$$

with the gradient

$$\nabla_{\hat{v}} \mathcal{G}(z, \hat{v}) = A_z^* \left[I_M - \operatorname{diag}\left(\frac{\sqrt{y_k + \varepsilon}}{\sqrt{|(A_z \hat{v})_k|^2 + \varepsilon}} \right) \right] A_z \hat{v} + \beta_T \hat{v}, \tag{4.9}$$

and the learning rate $\nu = \nu(\mathcal{G}, \hat{v}, \tau, \nu_c, N)$ determined by Algorithm 1, where ν_c satisfies

$$0 < \nu_c \leq \left[m \max_{\ell \in [\delta]} \sum_{r \in \mathcal{R}} |(S_{-r} z)_\ell|^2 + \beta_T \right]^{-1} =: L_z^{-1}. \tag{4.10}$$

Then,

$$\mathcal{G}(z, \hat{v}^+) - \mathcal{G}(z, \hat{v}) \leq -\nu \left\| \nabla_{\hat{v}} \mathcal{G}(z, \hat{v}) \right\|_2^2.$$

Proof. Due to the representation (4.8), the proof is analogous to the proof of Theorem 4.2.3 and is based on Lemma 4.2.1, Corollary 3.5.2 and Theorem 2.3.6. The only difference is the evaluation of the constant learning rate, which depends on A_z. In order to apply Theorem 2.3.6, we first justify that

$$\nu_c \leq \left[\left\| \sum_{k \in [M]} Q_k^z \right\|_\infty + \beta_T \right],$$

where $\{Q_k^z\}_{k \in [M]}$ is the family of rank-one matrices corresponding to the matrix A_z. Recall that by construction $Q_k^z = a_k^z (a_k^z)^*$ with a_k^z being the conjugate of k-th row of A_z. Then, we have

$$\left\| \sum_{k \in [M]} Q_k^z \right\|_\infty = \left\| \sum_{k \in [M]} a_k^z (a_k^z)^* \right\|_\infty = \left\| A_z^* A_z \right\|_\infty,$$

and the product $A_z^* A_z$ is given by

$$\begin{aligned} A_z^* A_z &= \sum_{r \in \mathcal{R}} (F_m P_\delta^* \operatorname{diag}(P_\delta S_{-r} z))^* F_m P_\delta^* \operatorname{diag}(P_\delta S_{-r} z) \\ &= \sum_{r \in \mathcal{R}} \operatorname{diag}(\overline{P_\delta S_{-r} z}) P_\delta F_m^* F_m P_\delta^* \operatorname{diag}(P_\delta S_{-r} z). \end{aligned}$$

Using Proposition 2.2.1 and (2.11), we arrive at

$$\begin{aligned} A_z^* A_z &= \sum_{r \in \mathcal{R}} \operatorname{diag}(\overline{P_\delta S_{-r} z}) P_\delta (m I_m) P_\delta^* \operatorname{diag}(P_\delta S_{-r} z) \\ &= m \sum_{r \in \mathcal{R}} \operatorname{diag}(\overline{P_\delta S_{-r} z}) I_m \operatorname{diag}(P_\delta S_{-r} z) \\ &= m \sum_{r \in \mathcal{R}} \operatorname{diag}(|P_\delta S_{-r} z|^2) = m \operatorname{diag}\left(\sum_{r \in \mathcal{R}} |P_\delta S_{-r} z|^2 \right). \end{aligned}$$

Hence, the spectral norm of $A_z^* A_z$ is given by

$$\left\| \sum_{k\in[M]} Q_k^z \right\|_\infty = \|A_z^* A_z\|_\infty = m \max_{\ell\in[\delta]} \left(\sum_{r\in\mathcal{R}} |P_\delta S_{-r} z|^2 \right)_\ell = m \max_{\ell\in[\delta]} \sum_{r\in\mathcal{R}} |(S_{-r} z)_\ell|^2,$$

which concludes the proof. □

4.2.3 Formal algorithm and convergence guarantees

Now that the optimization with respect to either z or $\hat{v}$ is formalized, we can establish the detailed reconstruction procedure for blind ptychography.

Algorithm 10: Alternating Amplitude Flow for blind ptychography

Input : Measurements Y as in (BPTY), number of iterations $T \in \mathbb{N}$, number of object and window subiterations $T_z \in \mathbb{N}$ and $T_{\hat{v}} \in \mathbb{N}$, parameters $\varepsilon > 0$, $\alpha_T, \beta_T \geq 0$, initial guesses $z^0 \in \mathbb{C}^d$ and $\hat{v}^0 \in \mathbb{C}^\delta$, AG parameters $0 < \tau < 1$, $N \in \mathbb{N} \cup \{0\}$.

Output: $z \in \mathbb{C}^d$ and $\hat{v} \in \mathbb{C}^\delta$.

for $t = 1, \ldots, T$ **do**
- Let $z^{t,0} = z^{t-1}$.
- Set $\mu_{t,c} = L_{\hat{v}^{t-1}}^{-1}$ as in (4.6).
- **for** $i \in [T_z]$ **do**
  - Determine $\mu_{t,i} = \mu_{t,i}(\mathcal{G}, z^{t,i}, \tau, \mu_{t,c}, N)$ via Algorithm 1.
  - Update $z^{t,i+1} = z^{t,i} - \mu_{t,i} \nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})$.
- Let $z^t = z^{t,T_z}$ and $\hat{v}^{t,0} = \hat{v}^{t-1}$
- Set $\nu_{t,c} = L_{z^t}^{-1}$ as in (4.10).
- **for** $j \in [T_{\hat{v}}]$ **do**
  - Determine $\nu_{t,j} = \nu_{t,j}(\mathcal{G}, \hat{v}^{t,j}, \tau, \nu_{t,c}, N)$ via Algorithm 1.
  - Update $\hat{v}^{t,j+1} = \hat{v}^{t,j} - \nu_{t,j} \nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^{t,j})$.
- Let $\hat{v}^t = \hat{v}^{t,T_{\hat{v}}}$.

return $z = z^T, \hat{v} = \hat{v}^T$.

Firstly, let us show that the learning rates are always finite.

Lemma 4.2.5. *Let $\varepsilon, \alpha_T, \beta_T > 0$. Then, for all $z \in \mathbb{C}^d$ and $\hat{v} \in \mathbb{C}^\delta$ we have $L_{\hat{v}} \geq \alpha_T > 0$ and $L_z \geq \beta_T > 0$. Furthermore, the learning rates $\mu_{t,i}$ and $\nu_{t,j}$, $i \in [T_z]$, $j \in [T_{\hat{v}}]$, $t \geq 1$, determined by Algorithm 10 are bounded from above by τ^{-N}/α_T and τ^{-N}/β_T, respectively.*

Proof. By (4.6), $L_{\hat{v}}$ satisfies $L_{\hat{v}} \geq \alpha_T > 0$. For a learning rate $\mu_{t,i}$ determined by Algorithm 1, the inequality (2.21) gives

$$\mu_{t,i} \leq \tau^{-N} \mu_{t,c} = \tau^{-N} L_{\hat{v}^{t-1}}^{-1} \leq \tau^{-N}/\alpha_T < \infty.$$

Analogously, $L_z \geq \beta_T > 0$ and $\nu_{t,j} \leq \tau^{-N}/\beta_T < \infty$. □

Secondly, it is possible to derive the convergence of Algorithm 10 and to show the sublinear convergence rate.

Theorem 4.2.6. *Let $\mathcal{G} : \mathbb{C}^d \times \mathbb{C}^\delta \to [0, \infty)$ be defined as in (4.1) with $\varepsilon, \alpha_T, \beta_T > 0$. Consider two sequences $\{z^t\}_{t\geq 0}, \{\hat{v}^t\}_{t\geq 0}$ defined by Algorithm 10 with arbitrary starting points $z^0 \in \mathbb{C}^d$, $\hat{v}^0 \in \mathbb{C}^\delta$ and let $\mu_{t,i}$ and $\nu_{t,i}$ be the learning rates determined by Algorithm 10. Then, for each subiteration of Algorithm 10 we have*

$$\mathcal{G}(z^{t,i+1}, \hat{v}^{t-1}) - \mathcal{G}(z^{t,i}, \hat{v}^{t-1}) \leq -\mu_{t,i} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 \tag{4.11}$$

$$\mathcal{G}(z^t, \hat{v}^{t,j+1}) - \mathcal{G}(z^t, \hat{v}^{t,j}) \leq -\nu_{t,j} \left\|\nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2. \tag{4.12}$$

for every $t \geq 1$ and $i \in [T_z]$, $j \in [T_{\hat{v}}]$.
Moreover,

$$\lim_{t\to\infty} \left\|\nabla_z \mathcal{G}(z^t, \hat{v}^t)\right\|_2^2 + \left\|\nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^t)\right\|_2^2 = 0,$$

where the rate of convergence is dominated by

$$\frac{\max\{\alpha_T^{-1}, \beta_T^{-1}\} m \mathcal{G}^2(z^0, \hat{v}^0) + \max\{\alpha_T, \beta_T\} \mathcal{G}(z^0, \hat{v}^0)}{T \min\{T_z, T_{\hat{v}}\}}.$$

Proof. Let $t \geq 1$ be fixed. For each iteration of the object, $\hat{v}^{t-1}$ is fixed and the constant learning rate $\mu_{t,c} = L_{\hat{v}^{t-1}}^{-1}$ satisfies the condition (4.6) by construction. Thus, by Theorem 4.2.3 the inequality (4.11) holds. Analogously, Theorem 4.2.4 yields (4.12). These estimates show that every subiteration of Algorithm 10 does not increase the value of the loss function $\mathcal{G}(z, \hat{v})$. Furthermore, summing up the bound for the object and window subiterations, we get

$$\mathcal{G}(z^{t,T_z}, \hat{v}^{t-1}) - \mathcal{G}(z^{t,0}, \hat{v}^{t-1}) \leq -\sum_{i\in[T_z]} \mu_{t,i} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2,$$

$$\mathcal{G}(z^t, \hat{v}^{t,T_{\hat{v}}}) - \mathcal{G}(z^t, \hat{v}^{t,0}) \leq -\sum_{j\in[T_{\hat{v}}]} \nu_{t,j} \left\|\nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2.$$

Note that by construction

$$z^{t,T_z} = z^t, \quad z^{t,0} = z^{t-1}, \quad \hat{v}^{t,T_{\hat{v}}} = \hat{v}^t, \quad \hat{v}^{t,0} = \hat{v}^{t-1}. \tag{4.13}$$

Hence, combining the inequalities leads to

$$\sum_{i\in[T_z]} \mu_{t,i} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 + \sum_{j\in[T_{\hat{v}}]} \nu_{t,j} \left\|\nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2 \leq \mathcal{G}(z^{t-1}, \hat{v}^{t-1}) - \mathcal{G}(z^t, \hat{v}^t).$$

For a fixed $T \geq 1$ the summation over $t = 1, \ldots, T$ gives

$$\begin{aligned}\sum_{t=1}^{T} \left[\sum_{i\in[T_z]} \mu_{t,i} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 + \sum_{j\in[T_{\hat{v}}]} \nu_{t,j} \left\|\nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2\right] \\ \leq \sum_{t=1}^{T} \left[\mathcal{G}(z^{t-1}, \hat{v}^{t-1}) - \mathcal{G}(z^t, \hat{v}^t)\right] = \mathcal{G}(z^0, \hat{v}^0) - \mathcal{G}(z^T, \hat{v}^T) \leq \mathcal{G}(z^0, \hat{v}^0),\end{aligned} \tag{4.14}$$

where we used that $\mathcal{G}(z, \hat{v}) \geq 0$. Hence, if $T \to \infty$, we arrive at

$$\sum_{t=1}^{\infty} \left[\sum_{i\in[T_z]} \mu_{t,i} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 + \sum_{j\in[T_{\hat{v}}]} \nu_{t,j} \left\|\nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2\right] < \infty,$$

which implies that

$$\sum_{i\in[T_z]} \mu_{t,i} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 + \sum_{j\in[T_{\hat{v}}]} \nu_{t,j} \left\|\nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2 \to 0$$

as $t \to \infty$. Since all terms are non-negative we eventually get

$$\mu_{t,i} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 \to 0, \quad \text{and} \quad \nu_{t,j} \left\|\nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2 \to 0$$

for all $i \in [T_z]$, $j \in [T_{\hat{v}}]$ as $t \to \infty$.
In order to show the desired convergence for the norm of the gradients, we have to show that the learning rates $\mu_{t,i}$ and $\nu_{t,j}$ are not converging to zero as $t \to \infty$ for all i, j. Since the learning rate $\mu_{t,i}$ is determined by Algorithm 1, by (2.21) we have

$$\mu_{t,i} \geq \mu_{t,c} = L_{\hat{v}^{t-1}}^{-1}. \tag{4.15}$$

Hence, to prove that $\mu_{t,i}$ does not vanish is equivalent to show that the sequence $\{L_{\hat{v}^t}\}_{t\geq 0}$ is bounded from above. Recall that by (4.6), $L_{\hat{v}^t}$ is given by

$$L_{\hat{v}^t} = m \max_{\ell\in[d]} \sum_{r\in\mathcal{R}} |(S_r P_\delta^* \hat{v}^t)_\ell|^2 + \alpha_T \leq m \max_{\ell\in[d]} \sum_{r\in[d]} |(S_r P_\delta^* \hat{v}^t)_\ell|^2 + \alpha_T.$$

By changing the order of summation, we obtain

$$\sum_{r\in[d]} |(S_r P_\delta^* \hat{v}^t)_\ell|^2 = \sum_{r\in[d]} |(P_\delta^* \hat{v}^t)_{r+\ell}|^2 = \|P_\delta^* \hat{v}\|_2^2.$$

Since P_δ^* only appends zeros, $L_{\hat{v}^t}$ is further bounded from above by

$$L_{\hat{v}^t} \leq m \max_{\ell\in[d]} \left\|P_\delta^* \hat{v}^t\right\|_2^2 + \alpha_T = m \left\|\hat{v}^t\right\|_2^2 + \alpha_T.$$

Consequently, $\{L_{\hat{v}^t}\}_{t\geq 0}$ is bounded if and only if the sequence $\{\|\hat{v}^t\|_2^2\}_{t\geq 0}$ is bounded. Let us show by contradiction that $\{\|\hat{v}^t\|_2^2\}_{t\geq 0}$ is bounded from above by $\mathcal{G}(z^0, \hat{v}^0)/\beta_T$. More precisely, assume that for some $t_0 \geq 0$, the opposite holds, i.e., $\|\hat{v}^{t_0}\|_2^2 > \mathcal{G}(z^0, \hat{v}^0)/\beta_T$. Then, we obtain

$$\mathcal{G}(z^{t_0}, \hat{v}^{t_0}) = \mathcal{H}(z^{t_0}; A_{\hat{v}^{t_0}}, \alpha_T) + \beta_T \left\|\hat{v}^{t_0}\right\|_2^2 \geq \beta_T \left\|\hat{v}^{t_0}\right\|_2^2 > \mathcal{G}(z^0, \hat{v}^0),$$

which is not possible, since we showed in (4.14) that with each iteration the loss function does not increase. Therefore, $\{\|\hat{v}^t\|_2^2\}_{t\geq 0}$ is bounded from above by $\mathcal{G}(z^0, \hat{v}^0)/\beta_T$ and $\{L_{\hat{v}^t}\}_{t\geq 0}$ is also bounded from above by

$$L_{win} := m\beta_T^{-1} \mathcal{G}(z^0, \hat{v}^0) + \alpha_T > 0. \tag{4.16}$$

The strict inequality is due to $L_{win} \geq \alpha_T > 0$. Hence, by (4.15),

$$\mu_{t,i} \geq L_{win}^{-1} > 0, \tag{4.17}$$

and we obtain

$$\left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 \to 0,$$

as $t \to \infty$ for all $i \in [T_z]$. Similarly, $\{L_{z^t}\}_{t\geq 0}$ is bounded from above by

$$L_{obj} := m\alpha_T^{-1}\mathcal{G}(z^0, \hat{v}^0) + \beta_T > 0,$$

so that for all $j \in [T_{\hat{v}}]$, we obtain

$$\nu_{t,j} \geq L_{obj}^{-1} > 0 \quad \text{and} \quad \left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2 \to 0,$$

as $t \to \infty$. In particular,

$$\begin{aligned}\left\|\nabla_z\mathcal{G}(z^t, \hat{v}^t)\right\|_2^2 &= \left\|\nabla_z\mathcal{G}(z^{t+1,0}, \hat{v}^t)\right\|_2^2 \to 0 \text{ as } t \to \infty,\\ \left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t-1})\right\|_2^2 &= \left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t,0})\right\|_2^2 \to 0 \text{ as } t \to \infty.\end{aligned}$$

Recall that our goal is to show that

$$\lim_{t\to\infty} \left\|\nabla_z\mathcal{G}(z^t, \hat{v}^t)\right\|_2^2 + \left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^t)\right\|_2^2 = 0.$$

Thus, it remains to show $\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^t)\|_2^2 \to 0$ as $t \to \infty$. Using the triangle inequality, we obtain

$$0 \leq \left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^t)\right\|_2 \leq \left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^t) - \nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t-1})\right\|_2 + \left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t-1})\right\|_2.$$

It was already shown that the second summand converges to zero as $t \to \infty$. For the first summand we can use the fact that $\nabla_{\hat{v}}\mathcal{G}(z, \hat{v})$ is continuous for $\varepsilon > 0$. Therefore, the first summand converges to zero if $\|\hat{v}^t - \hat{v}^{t-1}\|_2 \to 0$ as $t \to \infty$. In fact, we have,

$$\begin{aligned}0 \leq \left\|\hat{v}^t - \hat{v}^{t-1}\right\|_2 = \left\|\hat{v}^{t,T_{\hat{v}}} - \hat{v}^{t,0}\right\|_2 &= \left\|\sum_{j\in[T_{\hat{v}}]} \nu_{t,j}\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2\\ &\leq \sum_{j\in[T_{\hat{v}}]} \nu_{t,j}\left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2 \leq \tau^{-N}\beta_T^{-1}\sum_{j\in[T_{\hat{v}}]}\left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2,\end{aligned}$$

where in the last line we used Lemma 4.2.5. For $t \to \infty$, we obtain $\|\hat{v}^t - \hat{v}^{t-1}\|_2 \to 0$. Consequently, by continuity $\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^t) - \nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t-1})\|_2 \to 0$, which gives

$$\left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^t)\right\|_2 \to 0,$$

as $t \to 0$. For the convergence speed, we consider the sequence

$$s_t := \max\left\{\min_{i\in[T_z]}\left\|\nabla_z\mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2, \min_{j\in[T_{\hat{v}}]}\left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2\right\}.$$

If s_t is small, it follows that the gradient iterations in either direction are small and the iterates are in a proximity of a fixed point. For a minimum of s_t, the following upper bound applies,

$$\min_{t=1,\dots,T} s_t \leq \tfrac{1}{T}\sum_{t=1}^{T} s_t \leq \tfrac{1}{T}\sum_{t=1}^{T}\left[\min_{i\in[T_z]}\left\|\nabla_z\mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 + \min_{j\in[T_{\hat{v}}]}\left\|\nabla_{\hat{v}}\mathcal{G}(z^t, \hat{v}^{t,j})\right\|_2^2\right].$$

The first minimum is bounded from above as

$$\min_{i\in[T_z]} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 \leq \frac{1}{T_z} \sum_{i\in[T_z]} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2 \\ \leq \frac{1}{T_z \min_{i\in[T_z]} \mu_{t,i}} \sum_{i\in[T_z]} \mu_{t,i} \left\|\nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1})\right\|_2^2.$$

and, similarly, the second minimum is bounded by

$$\min_{j\in[T_{\hat{v}}]} \left\|\nabla_{\hat{v}} \mathcal{G}(z^{t}, \hat{v}^{t,j})\right\|_2^2 \leq \frac{1}{T_{\hat{v}} \min_{j\in[T_{\hat{v}}]} \nu_{t,j}} \sum_{j\in[T_{\hat{v}}]} \nu_{t,j} \left\|\nabla_{\hat{v}} \mathcal{G}(z^{t}, \hat{v}^{t,j})\right\|_2^2.$$

Combined with (4.14), these bounds give

$$\min_{t=1,\dots,T} s_t \leq \frac{1}{T \min\{T_z \min_{i\in[T_z]} \mu_{t,i}, T_{\hat{v}} \min_{j\in[T_{\hat{v}}]} \nu_{t,j}\}} \mathcal{G}(z^0, \hat{v}^0) \\ \leq \frac{1}{T \min\{T_z, T_{\hat{v}}\} \min\{\min_{i\in[T_z]} \mu_{t,i}, \min_{j\in[T_{\hat{v}}]} \nu_{t,j}\}} \mathcal{G}(z^0, \hat{v}^0).$$

In view of (4.17) and (4.16), for the learning rates $\mu_{t,i}$ we obtain

$$\frac{1}{\min_{i\in[T_z]} \mu_{t,i}} \leq \frac{1}{\mu_{t,c}} \leq L_{obj} = m\beta_T^{-1} \mathcal{G}(z^0, \hat{v}^0) + \alpha_T \\ \leq m \max\{\alpha_T^{-1}, \beta_T^{-1}\} \mathcal{G}(z^0, \hat{v}^0) + \max\{\alpha_T, \beta_T\},$$

and for $1/\min_{j\in[T_{\hat{v}}]} \nu_{t,j}$ the upper bound is precisely the same. Then, we arrive at

$$\min_{t=1,\dots,T} s_t \leq \frac{m \max\{\alpha_T^{-1}, \beta_T^{-1}\} \mathcal{G}^2(z^0, \hat{v}^0) + \max\{\alpha_T, \beta_T\} \mathcal{G}(z^0, \hat{v}^0)}{T \min\{T_z, T_{\hat{v}}\}}.$$

□

The first main consequence of Theorem 4.2.6 is that Algorithm 10 will always stop, which is not known for some of the methods applied in practice, for example the extended Ptychographic Iterative Engine algorithm [33]. The second is that the gradient decay is inversely proportional to the number of iterations and quadratically depends on the value of the loss function $\mathcal{G}$ evaluated at the initial guess $z^0, \hat{v}^0$. Thus, the starting point has a major influence on the performance of the algorithm.

We note that the Tikhonov regularization is crucial in Theorem 4.2.6. While the quadratic loss $\mathcal{L}_{2,\varepsilon}(z; A_{\hat{v}})$ is constant under rescaling $z, \hat{v} \mapsto \gamma z, \hat{v}/\gamma$, $\gamma \neq 0$ in accordance with the scaling ambiguity, its gradient $\nabla_z \mathcal{L}_{2,\varepsilon}(z; A_{\hat{v}})$ is not. Thus, it is harder to control the norm $\|\nabla_z \mathcal{L}_{2,\varepsilon}(z; A_{\hat{v}})\|_2$. The inclusion of the Tikhonov regularization resolves the ambiguity and allows to establish the convergence guarantees. Furthermore, we observe that the convergence rate depends on $\max\{\alpha_T^{-1}, \beta_T^{-1}\}$, which grows to infinity as the regularization parameters vanish.

Finally, we point out that similar to Theorem 3.5.5, Theorem 4.2.6 only grants convergence to a fixed point of the loss function $\mathcal{G}$, which is non necessarily is a global minimizer.

4.2.4 Exclusion of Tikhonov regularization via reweighting

One of the issues arising from Algorithm 10 is the selection of the regularization parameters α_T and β_T, which is done empirically in practice. This requires several trial-and-error attempts to find good choices of α_T, β_T. For this reason we propose an alternative version of Algorithm 10, which completely excludes the regularization terms. Since the Tikhonov regularization was a key in controlling the scaling ambiguity and achieving the convergence results in Theorem 4.2.6, an alternative way to resolve the ambiguity is required. This is done by normalizing the fixed variable and scaling the other appropriately to preserve the same value of the loss function $\mathcal{G}$. However, the normalization procedure may cause a division by zero, which has to be treated separately.

Algorithm 11: Alternating Amplitude Flow with reweighting

Input : Measurements Y as in (BPTY), number of iterations $T \in \mathbb{N}$, number of object and window subiterations $T_z \in \mathbb{N}$ and $T_{\hat{v}} \in \mathbb{N}$, parameter $\varepsilon > 0$, initial guesses $z^0 \in \mathbb{C}^d$ and $\hat{v}^0 \in \mathbb{C}^\delta \backslash \{\mathbb{0}_\delta\}$, AG parameters $0 < \tau < 1$, $N \in \mathbb{N} \cup \{0\}$.

Output: $z \in \mathbb{C}^d$ and $\hat{v} \in \mathbb{C}^\delta$.

Set $\alpha_T = 0, \beta_T = 0$ in $\mathcal{G}$.

Set $z^0 = \|\hat{v}^0\|_2 z^0$ and $\hat{v}^0_{nor} = \hat{v}^0 / \|\hat{v}^0\|_2$.

for $t = 1, \dots, T$ **do**

- Let $z^{t,0} = z^{t-1}$.
- Set $\mu_{t,c} = L^{-1}_{\hat{v}^{t-1}_{nor}}$ as in (4.6).
- **for** $i \in [T_z]$ **do**
  - Determine $\mu_{t,i} = \mu_{t,i}(\mathcal{G}, z^{t,i}, \tau, \mu_{t,c}, N)$ via Algorithm 1.
  - Update $z^{t,i+1} = z^{t,i} - \mu_{t,i} \nabla_z \mathcal{G}(z^{t,i}, \hat{v}^{t-1}_{nor})$.
- **if** $\left\|z^{t,T_z}\right\|_2 = 0$ **then**
  - **return** $z = z^{t,T_z}$, $\hat{v} = \hat{v}^{t-1}_{nor}$.
- Let $\hat{v}^{t,0} = \left\|z^{t,T_z}\right\|_2 \hat{v}^{t-1}_{nor}$ and $z^t_{nor} = z^{t,T_z} / \left\|z^{t,T_z}\right\|_2$
- Set $\nu_{t,c} = L^{-1}_{z^t_{nor}}$ as in (4.10).
- **for** $j \in [T_{\hat{v}}]$ **do**
  - Determine $\nu_{t,j} = \nu_{t,j}(\mathcal{G}, \hat{v}^{t,j}, \tau, \nu_{t,c}, N)$ via Algorithm 1.
  - Update $\hat{v}^{t,j+1} = \hat{v}^{t,j} - \nu_{t,j} \nabla_{\hat{v}} \mathcal{G}(z^t_{nor}, \hat{v}^{t,j})$.
- **if** $\left\|\hat{v}^{t,T_{\hat{v}}}\right\|_2 = 0$ **then**
  - **return** $z = z^t_{nor}$, $\hat{v} = \hat{v}^{t,T_{\hat{v}}}$.
- Let $z^t = \left\|\hat{v}^{t,T_{\hat{v}}}\right\|_2 z^t_{nor}$ and $\hat{v}^t_{nor} = \hat{v}^{t,T_{\hat{v}}} / \left\|\hat{v}^{t,T_{\hat{v}}}\right\|_2$.

return $z = z^T, \hat{v} = \hat{v}^T_{nor}$.

Similarly to Lemma 4.2.5, we show that the learning rates are always well-defined in Algorithm 11.

Lemma 4.2.7. *Let $\varepsilon > 0$, $\alpha_T = 0$, $\beta_T = 0$. Assume that*

$$\cup_{r \in \mathcal{R}} \{r, r+1, \dots, r+\delta-1\} = [d]. \tag{4.18}$$

Then, for all $z \in \mathbb{C}^d$ and $\hat{v} \in \mathbb{C}^\delta$ such that $\|z\|_2 = 1$ and $\|\hat{v}\|_2 = 1$ we have $L_{\hat{v}} \geq m/\delta > 0$ and $L_z \geq m/d > 0$. Furthermore, the learning rates $\mu_{t,i}$ and $\nu_{t,j}$, $i \in [T_z]$, $j \in [T_{\hat{v}}]$, $t \geq 1$, determined by Algorithm 11 are bounded by $\tau^{-N}\delta/m$ and $\tau^{-N}d/m$, respectively.

Proof. In view of (4.6), $L_{\hat{v}}$ is given by

$$L_{\hat{v}} = m \max_{\ell \in [d]} \sum_{r \in \mathcal{R}} |(S_r P_\delta^* \hat{v})_\ell|^2 + \alpha_T = m \max_{\ell \in [d]} \sum_{r \in \mathcal{R}} |(S_r P_\delta^* \hat{v})_\ell|^2.$$

Let $r_0 \in \mathcal{R}$ and let $\ell_0 \in [\delta]$ be an index such that $|\hat{v}_{\ell_0}| = \|\hat{v}\|_\infty$. Then,

$$L_{\hat{v}} \geq m \sum_{r \in \mathcal{R}} |(S_r P_\delta^* \hat{v})_{\ell_0 + r_0}|^2 \quad \geq m |(S_{r_0} P_\delta^* \hat{v})_{\ell_0 + r_0}|^2 = m |(P_\delta^* \hat{v})_{\ell_0}|^2 = m |\hat{v}_{\ell_0}|^2 = m \|\hat{v}\|_\infty^2 .$$

Since $\|\hat{v}\|_2 = 1$, the infinity norm is bounded from below as

$$1 = \|\hat{v}\|_2^2 \leq \delta \|\hat{v}\|_\infty^2 ,$$

which gives

$$L_{\hat{v}} \geq m/\delta.$$

Similarly, let $\ell_1 \in [d]$ be such that $|z_{\ell_1}| = \|z\|_\infty$. By (4.18), there exists $r_1 \in \mathcal{R}$ such that $\ell_1 - r_1 \in [\delta]$. Thus,

$$\begin{aligned} L_z &= m \max_{\ell \in [\delta]} \sum_{r \in \mathcal{R}} |(S_{-r} z)_\ell|^2 + \beta_T \geq m \sum_{r \in \mathcal{R}} |(S_{-r} z)_{\ell_1 - r_1}|^2 \\ &\geq m |(S_{-r_1} z)_{\ell_1 - r_1}|^2 = m |z_{\ell_1}|^2 = m \|z\|_\infty^2 \geq m/d. \end{aligned}$$

Therefore, for the learning rates $\mu_{t,i}$ selected via Algorithm 1, the inequality (2.21) yields

$$\mu_{t,i} \leq \tau^{-N} \mu_{t,c} = \tau^{-N} L_{\hat{v}^{t-1}}^{-1} \leq \tau^{-N} \delta / m < \infty,$$

and analogously $\nu_{t,j} \leq \tau^{-N} d/m < \infty$. □

The set $\cup_{r \in \mathcal{R}} \{r, r+1, \ldots, r+\delta-1\}$ is the set of all observed entries of the object. Then, (4.18) is similar to the condition on the injectivity of the matrix A provided by Lemma 3.3.5, which requires that the whole object is illuminated during the experiments. If some of the entries are not observed, i.e., the assumption (4.18) is violated, it may cause a division by zero during the computation of the learning rates.

Example 4.2.8. *Assume that (4.18) does not hold and there exists ℓ_0 such that for all $r \in \mathcal{R}$ we have $\ell_0 \notin \{r, r+1, \ldots, r+\delta-1\}$ or equivalently $\ell_0 \neq r + \ell$ for all $\ell \in [\delta]$. Let z be the standard basis vector e_{ℓ_0} so that $z_\ell = \mathcal{I}_{\ell=\ell_0}$, $\ell \in [d]$, and $\|z\|_2 = 1$. Then, we obtain*

$$L_z = m \max_{\ell \in [\delta]} \sum_{r \in \mathcal{R}} |(S_{-r} z)_\ell|^2 = m \max_{\ell \in [\delta]} \sum_{r \in \mathcal{R}} |z_{r+\ell}|^2 = m \max_{\ell \in [\delta]} \sum_{r \in \mathcal{R}} \mathcal{I}_{r+\ell=\ell_0} = 0,$$

and the corresponding learning rate $\mu_c = L_z^{-1}$ is undefined.

The convergence of Algorithm 10 is summarized in the next theorem.

Theorem 4.2.9. *Let $\mathcal{G} : \mathbb{C}^d \times \mathbb{C}^\delta \to [0, \infty)$ be defined as in (4.1) with $\varepsilon > 0$, $\alpha_T = 0$, $\beta_T = 0$. Assume that (4.18) holds. Consider the sequences and the learning rates determined*

by Algorithm 11 with arbitrary starting points $z^0 \in \mathbb{C}^d$, $\hat{v}^0 \in \mathbb{C}^\delta \backslash \{\mathbb{0}_d\}$. *Then, for each subiteration of Algorithm 11 we have*

$$\mathcal{G}(z^{t,i+1}, \hat{v}_{nor}^{t-1}) - \mathcal{G}(z^{t,i}, \hat{v}_{nor}^{t-1}) \leq -\mu_{t,i} \left\| \nabla_z \mathcal{G}(z^{t,i}, \hat{v}_{nor}^{t-1}) \right\|_2^2 \tag{4.19}$$

$$\mathcal{G}(z_{nor}^{t}, \hat{v}^{t,j+1}) - \mathcal{G}(z_{nor}^{t}, \hat{v}^{t,j}) \leq -\nu_{t,j} \left\| \nabla_{\hat{v}} \mathcal{G}(z_{nor}^{t}, \hat{v}^{t,j}) \right\|_2^2. \tag{4.20}$$

for every $t \geq 1$ *and* $i \in [T_z]$, $j \in [T_{\hat{v}}]$.
If there exists a $T_0 \geq 1$ *such that either* $\left\| z^{T_0,T_z} \right\|_2 = 0$ *or* $\left\| \hat{v}^{T_0,T_{\hat{v}}} \right\|_2 = 0$, *then*

$$\left\| \nabla_z \mathcal{G}(z^{T_0}, \hat{v}^{T_0}) \right\|_2^2 + \left\| \nabla_{\hat{v}} \mathcal{G}(z^{T_0}, \hat{v}^{T_0}) \right\|_2^2 = 0.$$

Otherwise,

$$\lim_{t \to \infty} \left\| \nabla_z \mathcal{G}(z^t, \hat{v}_{nor}^t) \right\|_2^2 + \left\| \nabla_{\hat{v}} \mathcal{G}(z^t, \hat{v}_{nor}^t) \right\|_2^2 = 0.$$

In any case, the rate of convergence is dominated by

$$\frac{m \mathcal{G}(z^0, \hat{v}^0)}{T_1 \min\{T_z, T_{\hat{v}}\}},$$

where $T_1 := \min\{T_0 - 1, T\}$ *is the total number of iterations Algorithm 11.*

Proof. The structure of the proof resembles the proof of Theorem 4.2.6 and we will only highlight the differences when necessary. The inequalities (4.19) and (4.20) are obtained by applying Theorem 4.2.3 and Theorem 4.2.4, respectively.
If $T_0 \geq 1$ exists and $\left\| z^{T_0,T_z} \right\|_2 = 0$, then by (4.5) we obtain $\left\| \nabla_z \mathcal{G}(z^{T_0,T_z}, \hat{v}_{nor}^{T_0-1}) \right\|_2 = 0$. Furthermore, the matrix A^z is the zero matrix, which in combination with (4.9) yields $\left\| \nabla_{\hat{v}} \mathcal{G}(z^{T_0,T_z}, \hat{v}_{nor}^{T_0-1}) \right\|_2 = 0$. The case $\left\| \hat{v}^{T_0,T_{\hat{v}}} \right\|_2 = 0$ is analogous.
In the case $T_0 \geq 1$ does not exist, we repeat the series argument of Theorem 4.2.6, with (4.13) is replaced by

$$\mathcal{G}(z^{t,T_z}, \hat{v}_{nor}^{t-1}) = \mathcal{G}(z_{nor}^{t}, \hat{v}^{t,0}) \quad \text{and} \quad \mathcal{G}(z_{nor}^{t}, \hat{v}^{t,T_{\hat{v}}}) = \mathcal{G}(z^{t+1,0}, \hat{v}_{nor}^{t}). \tag{4.21}$$

Consequently, we arrive at

$$\mu_{t,i} \left\| \nabla_z \mathcal{G}(z^{t,i}, \hat{v}_{nor}^{t-1}) \right\|_2^2 \to 0, \quad \text{and} \quad \nu_{t,j} \left\| \nabla_{\hat{v}} \mathcal{G}(z_{nor}^{t}, \hat{v}^{t,j}) \right\|_2^2 \to 0$$

for all $i \in [T_z]$, $j \in [T_{\hat{v}}]$ as $t \to \infty$. Again, the learning rates are bounded from below by

$$\mu_{t,i} \geq \mu_{t,c} = L_{\hat{v}_{nor}^{t-1}}^{-1} \geq \left[m \left\| \hat{v}_{nor}^{t-1} \right\|_2^2 + \alpha_T \right]^{-1} = m^{-1}, \tag{4.22}$$

and, analogously, $\nu_{t,j} \geq m^{-1}$. Hence, we obtain

$$\left\| \nabla_z \mathcal{G}(z^{t,i}, \hat{v}_{nor}^{t-1}) \right\|_2^2 \to 0, \quad i \in [T_z], \text{ and } \left\| \nabla_{\hat{v}} \mathcal{G}(z_{nor}^{t}, \hat{v}^{t,j}) \right\|_2^2 \to 0, \quad j \in [T_{\hat{v}}],$$

as $t \to \infty$. In particular,

$$\left\| \nabla_z \mathcal{G}(z^t, \hat{v}_{nor}^t) \right\|_2^2 = \left\| \nabla_z \mathcal{G}(z^{t+1,0}, \hat{v}_{nor}^t) \right\|_2^2 \to 0 \text{ as } t \to \infty, \tag{4.23}$$

$$\left\| \nabla_{\hat{v}} \mathcal{G}(z_{nor}^t, \hat{v}^{t,0}) \right\|_2^2 \to 0 \text{ as } t \to \infty. \tag{4.24}$$

To prove that $\|\nabla_{\hat{v}}\mathcal{G}(z^t,\hat{v}^t_{nor})\|_2 \to 0$ as $t\to\infty$, we note that by (4.7) we have $A^{z^t} = \|\hat{v}^{t,T_{\hat{v}}}\|_2 A^{z^t_{nor}}$. Therefore, the gradient satisfies

$$\nabla_{\hat{v}}\mathcal{G}(z^t,\hat{v}^t_{nor}) = \nabla_{\hat{v}}\mathcal{G}(\|\hat{v}^{t,T_{\hat{v}}}\|_2 z^t_{nor}, \hat{v}^{t,T_{\hat{v}}}/\|\hat{v}^{t,T_{\hat{v}}}\|_2) = \|\hat{v}^{t,T_{\hat{v}}}\|_2 \nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,T_{\hat{v}}}). \tag{4.25}$$

In view of the construction of the iterates in Algorithm 11, the sequence $\{\|\hat{v}^{t,T_{\hat{v}}}\|_2\}_{t\geq 1}$ satisfies

$$\begin{aligned}\|\hat{v}^{t,T_{\hat{v}}}\|_2 &= \|\hat{v}^{t,T_{\hat{v}}}\|_2 - \|\hat{v}^{t,0}\|_2 + \|\hat{v}^{t,0}\|_2 = \|\hat{v}^{t,T_{\hat{v}}}\|_2 - \|\hat{v}^{t,0}\|_2 + \|z^{t,T_z}\|_2\\ &= \|\hat{v}^{t,T_{\hat{v}}}\|_2 - \|\hat{v}^{t,0}\|_2 + \|z^{t,T_z}\|_2 - \|z^{t,0}\|_2 + \|z^{t,0}\|_2\\ &= \|\hat{v}^{t,T_{\hat{v}}}\|_2 - \|\hat{v}^{t,0}\|_2 + \|z^{t,T_z}\|_2 - \|z^{t,0}\|_2 + \|\hat{v}^{t-1,T_{\hat{v}}}\|_2,\end{aligned}$$

and, consequently,

$$\begin{aligned}0 \leq \big|\|\hat{v}^{t,T_{\hat{v}}}\|_2 - \|\hat{v}^{t-1,T_{\hat{v}}}\|_2\big| &\leq \big|\|\hat{v}^{t,T_{\hat{v}}}\|_2 - \|\hat{v}^{t,0}\|_2\big| + \big|\|z^{t,T_z}\|_2 - \|z^{t,0}\|_2\big|\\ &\leq \|\hat{v}^{t,T_{\hat{v}}} - \hat{v}^{t,0}\|_2 + \|z^{t,T_z} - z^{t,0}\|_2.\end{aligned}$$

Recalling that

$$\hat{v}^{t,T_{\hat{v}}} - \hat{v}^{t,0} = \sum_{j\in[T_{\hat{v}}]} \nu_{t,j}\nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,j}) \quad \text{and} \quad z^{t,T_z} - z^{t,0} = \sum_{i\in[T_z]} \mu_{t,i}\nabla_z\mathcal{G}(z^{t,i},\hat{v}^{t-1}_{nor}),$$

by the triangle inequality and Lemma 4.2.7 we obtain

$$\|\hat{v}^{t,T_{\hat{v}}} - \hat{v}^{t,0}\|_2 \leq \sum_{j\in[T_{\hat{v}}]} \nu_{t,j}\|\nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,j})\|_2 \leq \frac{\tau^{-N}d}{m}\sum_{j\in[T_{\hat{v}}]}\|\nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,j})\|_2 \to 0, \tag{4.26}$$

as $t\to\infty$ and, analogously, $\|z^{t,T_z} - z^{t,0}\|_2 \to 0$ as $t\to\infty$. Consequently, $\{\|\hat{v}^{t,T_{\hat{v}}}\|_2\}_{t\geq 1}$ is the Cauchy sequence and, thus, it converges and is also bounded from above by some constant $c_{norm}\geq 0$. Returning to (4.25), we use the obtained bound to get

$$\begin{aligned}\|\nabla_{\hat{v}}\mathcal{G}(z^t,\hat{v}^t_{nor})\|_2 &= \|\hat{v}^{t,T_{\hat{v}}}\|_2\|\nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,T_{\hat{v}}})\|_2\\ &\leq c_{norm}\left[\|\nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,T_{\hat{v}}}) - \nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,0})\|_2 + \|\nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,0})\|_2\right].\end{aligned}$$

Since $\varepsilon > 0$, the gradient $\nabla_{\hat{v}}\mathcal{G}$ is continuous and the first term converges to zero as $t\to\infty$ due to (4.26). The second term also converges to 0 as $t\to\infty$ by (4.24) and, hence, $\|\nabla_{\hat{v}}\mathcal{G}(z^t,\hat{v}^t_{nor})\|_2$ vanishes as $t\to\infty$. Combining this with (4.23) gives

$$\lim_{t\to\infty}\|\nabla_z\mathcal{G}(z^t,\hat{v}^t_{nor})\|_2^2 + \|\nabla_{\hat{v}}\mathcal{G}(z^t,\hat{v}^t_{nor})\|_2^2 = 0.$$

The derivation of the convergence speed is analogous to the proof of Theorem 4.2.6 and considers

$$s_t := \max\left\{\min_{i\in[T_z]}\|\nabla_z\mathcal{G}(z^{t,i},\hat{v}^{t-1}_{nor})\|_2^2, \min_{j\in[T_{\hat{v}}]}\|\nabla_{\hat{v}}\mathcal{G}(z^t_{nor},\hat{v}^{t,j})\|_2^2\right\},$$

for $1\leq t\leq T_1$. Repeating the steps with an incorporation of the lower bound (4.22) for the learning rates yields

$$\min_{t=1,\dots,T_1} s_t \leq \frac{m\mathcal{G}(z^0,\hat{v}^0)}{T_1\min\{T_z,T_{\hat{v}}\}}.$$

□

Theorem 4.2.9 is similar to Theorem 4.2.6 in all aspects and the only difference is the convergence rate, which in Theorem 4.2.9 is better and depends on $\mathcal{G}(z^0, \hat{v}^0)$ linearly.
If Algorithm 11 terminates at some point $\left\|z^{T_0,T_z}\right\|_2 = 0$, it means that the gradient optimization with respect to object stopped at a local maximum. This is unlikely, but not impossible. Similarly, points satisfying $\left\|\hat{v}^{T_0,T_{\hat{v}}}\right\|_2 = 0$ are local maxima for the optimization with respect to the window. Furthermore, these points could also be an output of Algorithm 10, but they require no special attention in Algorithm 10 as it does not perform the normalization steps.
Finally, note that even if the condition (4.18) is not fulfilled, Algorithm 11 and its analysis can be adjusted by only considering entries in $\cup_{r\in\mathcal{R}}\{r, r+1, \ldots, r+\delta-1\}$ for normalization and ignoring the rest.

4.3 Extended Ptychographic Iterative Engine

Among practitioners, the popular choice for recovery from the measurements (BPTY) is extended Ptychographic Iterative Engine (ePIE). As the name suggests, it is an extension of the PIE algorithm for the joint recovery of the object and the window. The algorithm ePIE is initialized with a pair $z^0 \in \mathbb{C}^d$, $\hat{v}^0 \in \mathbb{C}^\delta$ and constructs the t-th iterate by performing the following steps.

Algorithm 12: ePIE iteration, version of [33]

Input : Shift position $r^t \in \mathcal{R}$ and corresponding measurements $Y^{(r)}$, previous iterates iterate $z^t \in \mathbb{C}^d$, $\hat{v}^t \in \mathbb{C}^\delta$, parameters $\alpha, \beta > 0$.
Output: $z^{t+1} \in \mathbb{C}^d$, $\hat{v}^{t+1} \in \mathbb{C}^\delta$.
1. Select a shift position $r^t \in \mathcal{R}$.
2. Construct an exit wave $\psi = S_{-r^t} z^t \circ P_\delta^* \hat{v}^t$.
3. Compute its Fourier transform $\Psi = F_m P_m \psi$.
4. Correct the magnitudes of Ψ as $\Psi' = \sqrt{Y^{(r^t)}} \circ \operatorname{sgn}_0 \Psi$.
5. Find an exit wave ψ' corresponding to Ψ' via $\psi' = P_m^* F_m^{-1} \Psi'$
6. Return

$$z^{t+1} = z^t + \frac{\alpha S_{r^t} \operatorname{diag}(\overline{\hat{v}}^t)}{\|\hat{v}\|_\infty^2}[\psi' - \psi], \quad \hat{v}^{t+1} = \hat{v}^t + \frac{\beta \operatorname{diag}(P_\delta S_{-r^t} \overline{z}^t)}{\|P_\delta S_{-r^t} z^t\|_\infty^2}[\psi' - \psi].$$

Several interpretations of the ePIE iterations are available in the literature, similarly to PIE. The ePIE algorithm can also be understood as stochastic gradient descent analogously to our interpretation of PIE in Section 3.5.3.We note that by (4.1), the function $\mathcal{G}$ is already the sum of functions corresponding to the shift positions $r \in \mathcal{R}$.

Theorem 4.3.1. *Let $\mathcal{G} : \mathbb{C}^d \times \mathbb{C}^\delta \to [0, \infty)$ be defined as in (4.1) with $\varepsilon = 0$, $\alpha_T = 0$, $\beta_T = 0$. If for each iteration $t \geq 1$, the shift position r^t is sampled uniformly at random from the set $\mathcal{R}$, then, the iteration of ePIE is equal to*

$$\begin{bmatrix} z^{t+1} \\ \hat{v}^{t+1} \end{bmatrix} = \begin{bmatrix} z^t \\ \hat{v}^t \end{bmatrix} - \begin{bmatrix} \mu_t I_d & O_{d\times\delta} \\ O_{\delta\times d} & \nu_t I_\delta \end{bmatrix} g_{\mathcal{G}}(z^t, \hat{v}^t)$$

where $O_{a\times b} \in \mathbb{C}^{a\times b}$ is the matrix with all entries equal to zero. The learning rates for the object and the window are given by

$$\mu_t = \frac{\alpha}{m|\mathcal{R}|\,\|\hat{v}^t\|_\infty^2} \quad \text{and} \quad \nu_t = \frac{\beta}{m|\mathcal{R}|\,\|P_\delta S_{-r^t} z^t\|_\infty^2},$$

respectively, and $g_{\mathcal{G}}$ is the stochastic gradient of $\mathcal{G}$ given by (2.23). The sampling variables v_r in $g_{\mathcal{G}}$ correspond to the sampling with replacement (2.27) for $K = 1$ and probabilities $1/|\mathcal{R}|$ as in (2.31).

Proof. Repeating the proof of Theorem 3.5.12, we obtain the analogue of (3.22), that is

$$z^{t+1} = z^t - \frac{\alpha}{m\,\|\hat{v}^t\|_\infty^2} \nabla_z \mathcal{L}_2(z^t; A_{\hat{v}^t, r^t}),$$

where $A_{\hat{v},r}$ denotes the row-block of matrix $A_{\hat{v}}$ corresponding to a shift position $r \in \mathcal{R}$. Similarly, we have

$$\hat{v}^{t+1} = \hat{v}^t - \frac{\beta}{m\,\|P_\delta S_{-r^t} z^t\|_\infty^2} \nabla_{\hat{v}} \mathcal{L}_2(\hat{v}^t; A_{z^t, r^t}),$$

with $A_{z,r}$ being the row-block of matrix A_z corresponding to a shift position $r \in \mathcal{R}$. Let us denote by $\mathcal{G}_r$ the summands of the loss function $\mathcal{G}$, so that

$$\mathcal{G}(z, \hat{v}) = \sum_{r\in\mathcal{R}} \mathcal{G}_r(z, \hat{v})$$

and analogously to (4.3) and (4.8), we obtain

$$\nabla_z \mathcal{G}_r(z, \hat{v}) = \nabla_z \mathcal{L}_2(z; A_{\hat{v},r}) \quad \text{and} \quad \nabla_{\hat{v}} \mathcal{G}_r(z, \hat{v}) = \nabla_{\hat{v}} \mathcal{L}_2(\hat{v}; A_{z,r}).$$

Furthermore, (3.23) yields

$$|\mathcal{R}| g_{\mathcal{G}}(z^t, \hat{v}^t) = \begin{bmatrix} \nabla_z \mathcal{G}_{r^t}(z^t, \hat{v}^t) \\ \nabla_{\hat{v}} \mathcal{G}_{r^t}(z, \hat{v}^t) \end{bmatrix} = \begin{bmatrix} \nabla_z \mathcal{L}_2(z^t; A_{\hat{v}^t, r^t}) \\ \nabla_{\hat{v}} \mathcal{L}_2(\hat{v}^t; A_{z^t, r^t}) \end{bmatrix},$$

and, consequently, we obtain

$$\begin{bmatrix} z^{t+1} \\ \hat{v}^{t+1} \end{bmatrix} = \begin{bmatrix} z^t \\ \hat{v}^t \end{bmatrix} - \begin{bmatrix} \frac{\alpha}{m\|\hat{v}^t\|_\infty^2} \nabla_z \mathcal{G}_{r^t}(z^t, \hat{v}^t) \\ \frac{\beta}{m\|P_\delta S_{-r^t} z^t\|_\infty^2} \nabla_{\hat{v}} \mathcal{G}_{r^t}(z^t, \hat{v}^t) \end{bmatrix} = \begin{bmatrix} z^t \\ \hat{v}^t \end{bmatrix} - \begin{bmatrix} \mu_t I_d & O_{d\times\delta} \\ O_{\delta\times d} & \nu_t I_\delta \end{bmatrix} g_{\mathcal{G}}(z^t, \hat{v}^t).$$

□

Recall that at the beginning of Section 4.2, we argued that the function $\mathcal{G}$ (with $\varepsilon > 0$) does not satisfy the inequality (2.17), while the most recent mathematical analysis of stochastic gradient descent methods is derived under the assumption that (2.17) holds. Therefore, despite of the stochastic gradient descent representation of ePIE, we are not able to establish the convergence guarantees for ePIE, in contrast to Theorem 3.5.13 for PIE.

Finally, we note that the learning rates are chosen as $\mu_t = \alpha/|\mathcal{R}|\,\|A_{\hat{v}^t, r^t}\|_\infty^2$ and $\nu_t = \beta/|\mathcal{R}|\,\|A_{z^t, r^t}\|_\infty^2$, which corresponds to the constants in (2.17) for the functions $\mathcal{G}_{r^t}(\cdot, \hat{v})$ and $\mathcal{G}_{r^t}(z, \cdot)$, respectively.

Notes and References. *Several approaches towards blind ptychographic recovery are present in the literature. In [188] an algorithm with guaranteed convergence based on alternating direction method of multiplier is proposed. The previously mentioned alternating Douglas-Rachford splitting is considered in [109], however, its convergence for blind ptychography is not analyzed. Another popular choice among practitioners is the extended Ptychographic Iterative Engine algorithm or its variants [33], which performs simultaneous optimization with respect to both the object and the window. Despite of its popularity, the convergence of ePIE is not studied and strongly depends on the choice of the parameters. The lack of convergence guarantees for extended Ptychographic Iterative Engine was the starting point for our research, which resulted in the development of the two methods, Algorithm 10 and Algorithm 11, with the guaranteed convergence. The only other method in the literature with quantitative convergence is [189]. There certain similarities between their algorithm and Algorithm 10 as both methods are based on alternating minimization. The main difference is that in [189] the constraints on the norms $\|x\|_2$ and $\|w\|_2$ are set explicitly, while in our case they are implicitly defined by the choice of Tikhonov regularization parameters α_T, β_T and the initial guesses z^0, v^0. Furthermore, as the objective functions are not the same, the sets of critical points might be different.*
We note that it is possible to replace the gradient steps in Algorithm 11 with projection iterations corresponding to the (smoothed) Error Reduction algorithm discussed in Section 3.5.2. Because of the scaled gradient representation of Error Reduction provided in Lemma 3.5.7, the resulting algorithm should posses convergence guarantees analogously to Theorem 4.2.9. In the literature, alternating Error Reduction without reweighting is discussed in [46].

Chapter 5

Polychromatic ptychography

So far we have focused on measurement setups where monochromatic light is used to illuminate the object. In practice, however, the illuminating light is almost never monochromatic. Instead the light usually consists of several spectral components. In this situation an adaptation of the mathematical model is necessary.

5.1 Changes in measurement model

In Chapter 1, diffraction patterns obtained in ptychographic experiment were explained by studying the physical phenomena for monochromatic waves. In this section, we expand on results of Chapter 1 for polychromatic light and derive analogous characterization of measurements. Hence, we will only highlight the main differences to the monochromatic case. Similarly to Chapter 1, within this section, we will avoid technicalities and allow ourselves to be not mathematically rigorous.

Recall that the mathematical description of ptychographic experiment is based on three phenomena: propagation of wave in free space, its interaction with an object and intensity of light. Their respective formulas where given by the equations (1.9), (1.10) and (1.11). These equations slightly change as polychromatic light contains several components.

Retracing the steps of Section 1.1, we recall that for a component u of an electro-magnetic field the wave equation (1.2) is satisfied. In the case of monochromatic waves, the Fourier transform was used to transit to the Helmholtz equation (1.6). The application of the Fourier transform can also be used for polychromatic waves. That is, the wave u is decomposed into its spectral components via Fourier transform as

$$u(s,t) = \int_{\mathbb{R}} e^{2\pi i \nu t} d\sigma_s(\nu),$$

with the spectral measure satisfying $d\sigma_s(-\nu) = \overline{d\sigma_s(\nu)}$, so that u is real-valued. Furthermore, we concentrate on the situation were σ is a linear combination of Dirac measures. That means that the light consists of a finite number of well-separated spectral components. Let $L \in \mathbb{N}$ be the number of monochromatic components and denote by ν_ℓ, $\ell \in [L]$, the frequencies of the monochromatic waves ordered in decreasing order, i.e.,

$\nu_0 > \nu_1 > \ldots, \nu_{L-1}$. Then, the spectral measure is given by

$$\sigma_r(\nu) = \frac{1}{2} \sum_{\ell \in [L]} [\boldsymbol{u}_\ell(s) \mathcal{I}_{\nu=\nu_\ell} + \overline{\boldsymbol{u}}_\ell(s) \mathcal{I}_{\nu=-\nu_\ell}],$$

with $\boldsymbol{u}_\ell$ being a density function corresponding to the frequency ν_ℓ.
Consequently, u is given by

$$u(s,t) = \frac{1}{2} \sum_{\ell \in [L]} \left[\boldsymbol{u}_\ell(s) e^{2\pi i \nu_\ell t} + \overline{\boldsymbol{u}}_\ell(s) e^{-2\pi i \nu_\ell t} \right]. \tag{5.1}$$

and this decomposition of u leads to a separate Helmholtz equation for each frequency

$$\Delta \boldsymbol{u}_\ell + k^2 \boldsymbol{u}_\ell = 0, \quad \ell \in [L].$$

Therefore, the propagation of polychromatic waves in free space is approximated by (1.9) applied to each frequency component $\boldsymbol{u}_\ell$ separately. More precisely, for a polychromatic wave with density functions $(\boldsymbol{u}_0)_\ell$, $\ell \in [L]$, propagating in free space from the x-y plane with $z = 0$ to the x-y plane with $z = d$ along the z-axis, the density functions $(\boldsymbol{u}_d)_\ell$ at $z = d$ are approximated by

$$(\boldsymbol{u}_d)_\ell(s) \approx \frac{-i\nu_\ell}{d} e^{2\pi i d \nu_\ell} e^{\frac{\pi i \nu_\ell}{d}(x^2+y^2)} \mathcal{F}(\boldsymbol{u}_0)_\ell \left(\frac{\nu_\ell s}{d} \right), \tag{5.2}$$

for $s = (x, y) \in \mathbb{R}^2$ and $\ell \in [L]$. This approximation is more precise if d is sufficiently large and the far-field assumption for frequency ν_ℓ is satisfied.
An interaction of a polychromatic wave with an object is described by separate interactions of monochromatic components with the object via (1.10). That is, for an incoming wave u_i with the density functions $(\boldsymbol{u}_i)_\ell$, $\ell \in [L]$, the density functions of the exit wave satisfy

$$(\boldsymbol{u}_e)_\ell(s) = \boldsymbol{x}_\ell(s)(\boldsymbol{u}_i)_\ell(s), \quad s \in \mathbb{R}^2, \ell \in [L], \tag{5.3}$$

where $\boldsymbol{x}_\ell : \mathbb{R}^2 \to \mathbb{C}$, $\ell \in [L]$, are object transfer functions corresponding to frequencies ν_ℓ.
The last building block of the ptychographic experiment is the intensity of light, which at position $s \in \mathbb{R}^2$ for wave u is proportional to

$$\boldsymbol{I}(s) \propto \langle |u(s,t)|^2 \rangle = \frac{1}{2T} \int_{-T}^{T} |u(s,t)|^2 dt,$$

where $T > 0$ is the acquisition time. By (5.1), we have

$$\begin{aligned} |u(s,t)|^2 &= \frac{1}{4} \left| \sum_{\ell \in [L]} \boldsymbol{u}_\ell(s) e^{2\pi i \nu_\ell t} + \overline{\boldsymbol{u}_\ell}(s) e^{-2\pi i \nu_\ell t} \right|^2 \\ &= \frac{1}{4} \sum_{\ell_1, \ell_2 \in [L]} \left[\boldsymbol{u}_{\ell_1}(s) \overline{\boldsymbol{u}_{\ell_2}}(s) e^{2\pi i t (\nu_{\ell_2} - \nu_{\ell_1})} + \overline{\boldsymbol{u}_{\ell_1}}(s) \boldsymbol{u}_{\ell_2}(s) e^{2\pi i t (\nu_{\ell_1} - \nu_{\ell_2})} \right] \\ &\quad + \frac{1}{4} \sum_{\ell_1, \ell_2 \in [L]} \left[\boldsymbol{u}_{\ell_1}(s) \boldsymbol{u}_{\ell_2}(s) e^{2\pi i t (\nu_{\ell_1} + \nu_{\ell_2})} + \overline{\boldsymbol{u}_{\ell_1}}(s) \overline{\boldsymbol{u}_{\ell_2}}(s) e^{-2\pi i t (\nu_{\ell_1} + \nu_{\ell_2})} \right]. \end{aligned}$$

Averaging this expression over a period of time $(-T, T)$, we arrive at

$$\begin{aligned}\boldsymbol{I}(s) \propto \frac{1}{2} \sum_{\ell_1,\ell_2\in[L]} \operatorname{Re}(\boldsymbol{u}_{\ell_1}(s)\overline{\boldsymbol{u}_{\ell_2}}(s)) \operatorname{sinc}(2\pi T|\nu_{\ell_1} - \nu_{\ell_2}|) \\ + \frac{1}{2} \sum_{\ell_1,\ell_2\in[L]} \operatorname{Re}(\boldsymbol{u}_{\ell_1}(s)\boldsymbol{u}_{\ell_2}(s)) \operatorname{sinc}(2\pi T(\nu_{\ell_1} + \nu_{\ell_2})).\end{aligned}$$

If the acquisition time T is significantly larger than $\max_{\ell_1\neq\ell_2} |\nu_{\ell_1} - \nu_{\ell_2}|^{-1}$, the sinc function vanishes in all cases, except if its argument is zero. Therefore, the intensity of the polychromatic wave u is proportional to the sum of intensities for each monochromatic component,

$$\boldsymbol{I}(s) \propto \sum_{\ell\in[L]} |\boldsymbol{u}_\ell(s)|^2, \quad s \in \mathbb{R}^2. \tag{5.4}$$

Consequently, let us consider the setting of the ptychographic experiment described in Section 1.4. If a light source is polychromatic, the resulting window w is also polychromatic and is described by densities $\boldsymbol{w}_\ell$, $\ell \in [L]$. For instance, if a plane polychromatic wave is truncated by a circular aperture, in analogy to Section 1.5, each component $\boldsymbol{w}_\ell$ is given by (1.14) with ν_ℓ substituted for ν.
Then, by replacing the equations (1.9), (1.10) and (1.11) in Section 1.4 with (5.2), (5.3) and (5.4), respectively, the measurements obtained by the detector is essentially the sum of the diffraction patterns for each monochromatic component,

$$\boldsymbol{I}(r, s) \propto \sum_{\ell\in[L]} \left| \frac{\nu_\ell}{p} \mathcal{F}[\boldsymbol{w}_\ell \mathcal{T}_{-r} \boldsymbol{x}_\ell] \left(\frac{\nu_\ell s}{p} \right) \right|^2, \quad s, r \in \mathbb{R}^2, \tag{5.5}$$

with $\mathcal{T}_r$ being the translation operator and $p > 0$ denoting the sufficiently large distance from the object plane to the detector plane to satisfy the far-field assumption for all frequencies ν_ℓ, $\ell \in [L]$.

5.2 Discrete polychromatic ptychography

For simplicity, we will work with one-dimensional problem again. Without loss of generality, we set $p = 1$, which is the same as changing the variable s and adjusting a proportional constant in (5.4). This leads to the problem of finding a family of functions $\boldsymbol{x}_\ell : \mathbb{R} \to \mathbb{C}$, $\ell \in [L]$, from the measurements

$$\boldsymbol{I}(r, s) = \sum_{\ell\in[L]} |\nu_\ell \mathcal{F}[\boldsymbol{w}_\ell \mathcal{T}_{-r} \boldsymbol{x}_\ell] (\nu_\ell s)|^2, \quad s, r \in \mathbb{R}. \tag{5.6}$$

Furthermore, we will assume that both the object transfer functions $\boldsymbol{x}_\ell$ and the density functions $\boldsymbol{w}_\ell$ are supported on the closed interval $[0, 1]$. For a discretization of the Fourier transforms in (5.6), we are using the partition with equidistant nodes

$$\Gamma_d = \left\{ \frac{0}{d}, \frac{1}{d}, \ldots, \frac{d-1}{d} \right\},$$

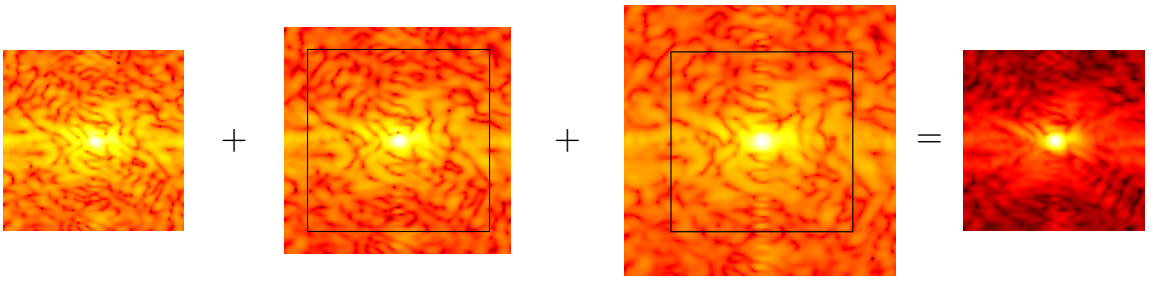


Figure 5.1: A diffraction pattern corresponding to polychromatic light consisting of three monochromatic waves. The black square shows the size of the diffraction pattern corresponding to the largest frequency.

and we consider shifts r of the form $\frac{n}{d}$, $n \in \mathbb{Z}$. This gives

$$\mathcal{F}[\boldsymbol{w}_\ell \mathcal{T}_{-r/d} \boldsymbol{x}_\ell](s\nu_\ell) \approx \frac{1}{d} \sum_{k\in[d]} \boldsymbol{x}_\ell\left(\frac{k+r}{d}\right) \boldsymbol{w}_\ell\left(\frac{k}{d}\right) \mathcal{I}_{\frac{k+r}{d}\in[0,1]} e^{-\frac{2\pi i \nu_\ell s k}{d}} \tag{5.7}$$

$$= \frac{1}{d} \sum_{k\in[d]} \boldsymbol{x}_\ell\left(\frac{k}{d}\right) \boldsymbol{w}_\ell\left(\frac{k-r}{d}\right) \mathcal{I}_{\frac{k-r}{d}\in[0,1]} e^{-\frac{2\pi i \nu_\ell s k}{d}} e^{\frac{2\pi i \nu_\ell s r}{d}}, \tag{5.8}$$

for $r \in \mathbb{Z}$. Note that unlike in Chapter 3, we are not working with cyclic shifts, but cutting out parts of $\boldsymbol{x}_\ell$ or $\boldsymbol{w}_\ell$ which lie in $[0,1]$. We assume that the measurements are only recorded for shift positions r in a set $\mathcal{R} \subseteq \{-d, -d+1, \ldots, d\}$ and denote the cardinality of this set by R. The restriction to $\{-d, -d+1, \ldots, d\}$ is a consequence of the fact that the indicator function $\mathcal{I}_{\frac{k+r}{d}\in[0,1]} = 0$ for all $r \notin \{-d, -d+1, \ldots, d\}$. If $\mathcal{R} = \{-d, -d+1, \ldots, d\}$, then all shits positions are present in the dataset.
The dual grid for the Fourier transform is dilated by ν_ℓ and in order to avoid multiple contributions from the largest frequency term in (5.8) we have to evaluate (5.8) on the dual grid $\{\frac{0}{\nu_0}, \frac{1}{\nu_0}, \ldots, \frac{d-1}{\nu_0}\}$. This is represented by the black square in Figure 5.1. With the notation

$$(x_\ell)_k = \boldsymbol{x}_\ell\left(\frac{k}{d}\right) \text{ and } (w_\ell)_k = \boldsymbol{w}_\ell\left(\frac{k}{d}\right),$$

we obtain the discretized polychromatic intensity measurements

$$\boldsymbol{I}\left(\frac{r}{d}, \frac{j}{\nu_0}\right) \approx I_{r,j} := \sum_{\ell\in[L]} \nu_\ell^2 \left| \sum_{k\in[d]} (x_\ell)_k (w_\ell)_{k-r} \mathcal{I}_{k-r\in[0,d]} e^{-\frac{2\pi i \nu_\ell k j}{\nu_0 d}} \right|^2, \quad r \in \mathcal{R},\ j \in [d].$$

Furthermore, the measurements may be corrupted by noise $N \in \mathbb{R}^{d\times R}$,

$$Y_{r,j} = I_{r,j} + N_{r,j}, \quad j \in [d],\ r \in \mathcal{R}. \tag{5.9}$$

Therefore, the *discrete polychromatic ptychographic problem* reads as follows:

Reconstruct $x_\ell \in \mathbb{C}^d,\ \ell \in [L],$ from data (5.9).

In Section 5.5 we will also consider the *blind discrete polychromatic ptychographic problem*:

Reconstruct $x_\ell, w_\ell \in \mathbb{C}^d,\ \ell \in [L],$ from data (5.9).

We note that the subsampling of frequencies may be considered similarly to the monochromatic case in order to reduce data volumes.

5.3 Ambiguities

The polychromatic ptychography can be considered as a generalization of ptychography as the both problems coincide for $L = 1$. That is, the measurements (5.6) for $L = 1$ are precisely the measurements (3.1). Consequently, polychromatic ptychography is bound to similar ambiguities.

Lemma 5.3.1. *Let $\alpha_\ell \in \mathbb{T}$, $\ell \in [L]$. Then, x_ℓ, $\ell \in [L]$ and $\alpha_\ell x_\ell$, $\ell \in [L]$ generate the same measurements.*

Proof. It follows directly from equality

$$\left| \sum_{k\in[d]} (\alpha_\ell x_\ell)_k (w_\ell)_{k-r} \mathcal{I}_{k-r\in[0,d]} e^{-\frac{2\pi i \nu_\ell kj}{\nu_0 d}} \right|^2 = \left| \sum_{k\in[d]} (x_\ell)_k (w_\ell)_{k-r} \mathcal{I}_{k-r\in[0,d]} e^{-\frac{2\pi i \nu_\ell kj}{\nu_0 d}} \right|^2 .$$

□

Analogously, in the case of blind polychromatic ptychography, the ambiguities described in Theorem 4.1.1 apply for each pair (x_ℓ, w_ℓ), $\ell \in [L]$ separately.

Theorem 5.3.2 (General ambiguities in blind polychromatic ptychography)**.** *Consider $x_\ell, w_\ell \in \mathbb{C}^d$, $\ell \in [L]$ and corresponding measurements (5.9). Then,*

1. *(global phase ambiguity) for all $\alpha_\ell, \beta_\ell \in \mathbb{T}$, $\ell \in [L]$ the pairs $(\alpha_\ell x_\ell, \beta_\ell w_\ell)$, $\ell \in [L]$ produces the same measurements (5.9),*
2. *(scaling ambiguity) for all $\gamma_\ell \in \mathbb{C}\backslash\{0\}$, $\ell \in [L]$ pairs $(\gamma_\ell x_\ell, w_\ell/\gamma_\ell)$, $\ell \in [L]$ produces the same measurements (5.9),*
3. *(linear phase ambiguity) for all $\rho_\ell \in \mathbb{R}$, $\ell \in [L]$ pairs (z_ℓ, v_ℓ), $\ell \in [L]$ with $(z_\ell)_k = e^{-i\rho_\ell k}(x_\ell)_k$ and $(v_\ell)_k = e^{i\rho_\ell k}(w_\ell)_k$, $k \in [d]$ produces the same measurements (5.9).*

Proof. The proof is analogous to Theorem 4.1.1. □

If we compare the number of unknowns in polychromatic ptychography to the monochromatic case, it is L times larger, yet the number of measurements remains the same. Consequently, it is possible that other ambiguities, such as in Example 4.1.2 may arise. However, their characterization is an open problem.

5.4 Amplitude Flow for polychromatic ptychography

In this section, we expand the Amplitude Flow algorithm for polychromatic ptychography. For doing this, we first present the polychromatic measurements in a form of quadratic

measurements as in (3.13). The first step towards the quadratic measurements is an introduction of vectors $a_{w_\ell,r,j} \in \mathbb{C}^d$ with entries

$$(a_{w_\ell,r,j})_k = \nu_\ell \, (\overline{w}_\ell)_{k-r} \, \mathcal{I}_{k-r\in[0,d]} \, e^{\frac{2\pi i \nu_\ell k j}{\nu_0 d}}.$$

Then, we can rewrite the polychromatic measurements (5.2) as

$$I_{r,j} := \sum_{\ell\in[L]} \left| a^*_{w_\ell,r,j} x_\ell \right|^2, \quad r \in \mathcal{R}, \ j \in [d].$$

This can be further rewritten as follows. Define a family $\mathcal{Q}^w = \{Q^w_{r,j}\}_{j\in[d],r\in\mathcal{R}}$ of block diagonal positive semidefinite matrices and vector

$$Q^w_{r,j} = \begin{bmatrix} a_{w_0,r,j} a^*_{w_0,r,j} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{w_{L-1},r,j} a^*_{w_{L-1},r,j} \end{bmatrix} \in \mathbb{H}^{dL}, \quad x = \begin{bmatrix} x_0 \\ \vdots \\ x_{L-1} \end{bmatrix} \in \mathbb{C}^{dL}, \tag{5.10}$$

which gives

$$Y_{r,j} = x^* Q^w_{r,j} x + N_{r,j}, \quad r \in \mathcal{R}, j \in [d].$$

Hence, x can be recovered by minimizing the amplitude-based loss function (3.13) or its smoothed version (3.14). Moreover, we include two additional regularization terms. The first is the Tikhonov regularization $\|x\|_2^2$, which leads to the smoothness of the object in the spatial domain. The second is the frequency smoothness term,

$$\mathcal{S}(x;\kappa) := \sum_{\ell\in[L-1]} \kappa_\ell \left\| x_{\ell+1} - x_\ell \right\|_2^2,$$

with parameters $\kappa_\ell > 0$, $\ell \in [L-1]$. It can be used to impose Lipschitz continuity of the object function $\boldsymbol{x}_\nu$ in frequency ν or alternatively in wavelength ν^{-1}. For instance, if the parameters κ_ℓ are set as $\frac{1}{|\nu^{-1}_{\ell+1} - \nu^{-1}_\ell|^2}$, this gives

$$\mathcal{S}(x;\kappa) := \sum_{\ell\in[L-1]} \frac{\|x_{\ell+1} - x_\ell\|_2^2}{|\nu^{-1}_{\ell+1} - \nu^{-1}_\ell|^2}.$$

A minimization of $\mathcal{S}(x;\kappa)$ leads to a smaller constant c in the inequality

$$\|x_{\ell+1} - x_\ell\|_2 \le c|\nu^{-1}_{\ell+1} - \nu^{-1}_\ell|.$$

Combining these terms, we establish an objective function

$$\mathcal{J}(x;\mathcal{Q}^w,\varepsilon,\alpha_T,\alpha_S,\kappa) := \mathcal{L}_{2,\varepsilon}(x;\mathcal{Q}^w) + \alpha_T \|x\|_2^2 + \alpha_S \mathcal{S}(x;\kappa). \tag{5.11}$$

The Amplitude Flow (AF) for polychromatic ptychography reconstructs x by minimizing $\mathcal{J}$ via gradient descent. That is, for an initial guess $z^0 \in \mathbb{C}^{dL}$ a sequence of iterates $\{z^t\}_{t\ge0} \subset \mathbb{C}^{dL}$ is determined by

$$z^{t+1} = z^t - \mu_t \nabla_z \mathcal{J}(z),$$

where $\mu_t > 0$ is a suitable learning rate.
For the convergence analysis of AF, we aim to apply Theorem 3.5.4, but before that, we derive a more convenient way to compute the gradient of $\mathcal{J}$ by introducing supplementary matrices $A_{w_\ell,r} \in \mathbb{C}^{d\times d}$ with rows

$$(A_{w_\ell,r})_{(j)} = a^*_{w_\ell,r,j}. \tag{5.12}$$

These matrices split into a product of two matrices $A_{w_\ell,r} = F_\ell D_{w_\ell,r}$, where the entries of Fourier-like matrix F_ℓ are given by

$$(F_\ell)_{j,k} := e^{-\frac{2\pi i \nu_\ell kj}{\nu_0 d}} \tag{5.13}$$

and $D_{w_\ell,r}$ is a diagonal matrix with entries

$$(D_{w_\ell,r})_{k,k} := \nu_\ell\,(w_\ell)_{k-r}\,\mathcal{I}_{k-r\in[0,d]}.$$

Furthermore, we also consider a block matrix $A_{w_\ell} \in \mathbb{C}^{dR\times d}$ and vectorize the measurements as

$$A_{w_\ell} = \begin{bmatrix} A_{w_\ell,r_1} \\ \vdots \\ A_{w_\ell,r_R} \end{bmatrix} \text{ and } y = \begin{bmatrix} Y^{(r_1)} \\ \vdots \\ Y^{(r_R)} \end{bmatrix}, \tag{5.14}$$

respectively. With these definitions we obtain the following properties for $\mathcal{J}$.

Lemma 5.4.1. *Let $\varepsilon > 0$. The function $\mathcal{J}$ is twice continuously differentiable with the gradient given by*

$$\nabla_{z_\ell}\mathcal{J}(z) = A^*_{w_\ell}\left(1 - \frac{\sqrt{y+\varepsilon}}{\sqrt{\sum_{\ell\in[L]}|A_{w_\ell}z|^2+\varepsilon}}\right)A_{w_\ell}z_\ell \\ + \alpha_T z_\ell + \alpha_S(\kappa_{\ell-1}(z_\ell - z_{\ell-1})\mathcal{I}_{\ell>0} + \kappa_\ell(z_\ell - z_{\ell+1})\mathcal{I}_{\ell<L-1}).$$

Moreover, the Hessian matrix of $\mathcal{J}$ satisfies inequality (2.17) with constant

$$L_w := \max_{\ell\in[L]}\nu_\ell^2\,\|F_\ell\|_\infty^2 \max_{k\in[d]}\sum_{r\in\mathcal{R}}|(w_\ell)_{k-r}|^2\mathcal{I}_{k-r\in[0,d]} + \alpha_T + \alpha_S\,\|K\|_\infty,$$

for all $z, u \in \mathbb{C}^{dL}$, where a matrix $K \in \mathbb{H}^L$ is defined by

$$K_{j,k} = \begin{cases} \kappa_{k-1}\mathcal{I}_{k>0} + \kappa_k\mathcal{I}_{k<L-1}, & j = k, \\ -\kappa_k, & j = k+1, \\ -\kappa_{k-1}, & j = k-1, \\ 0, & \textit{otherwise.} \end{cases} \tag{5.15}$$

Proof. By Lemma 3.5.1, the gradient is given by

$$\nabla_z\mathcal{L}_{2,\varepsilon}(z) = \sum_{j\in[d]}\sum_{r\in\mathcal{R}}\left(1 - \frac{\sqrt{y_k+\varepsilon}}{\sqrt{z^*Q^w_{r,j}z+\varepsilon}}\right)Q^w_{r,j}z.$$

By construction, we have

$$Q^w_{r,j} z = \begin{bmatrix} a_{w_0,r,j} a^*_{w_0,r,j} z_0 \\ \vdots \\ a_{w_{L-1},r,j} a^*_{w_{L-1},r,j} z_{L-1} \end{bmatrix} = \begin{bmatrix} (\overline{A}_{w_0,r})_{(j)} (A_{w_0,r} z_0)_j \\ \vdots \\ (\overline{A}_{w_{L-1},r})_{(j)} (A_{w_{L-1},r} z_{L-1})_j \end{bmatrix},$$

and

$$z^* Q^w_{r,j} z = \sum_{\ell \in [L]} |(A_{w_\ell,r} z_\ell)_j|^2 = \sum_{\ell \in [L]} |(A_{w_\ell} z_\ell)_{j,r}|^2.$$

Consequently, the gradient with respect to the component z_ℓ is given by

$$\begin{aligned} \nabla_{z_\ell} \mathcal{L}_{2,\varepsilon}(z) &= \sum_{r \in \mathcal{R}} \sum_{j \in [d]} \left(1 - \frac{\sqrt{Y_{r,j} + \varepsilon}}{\sqrt{\sum_{\ell \in [L]} |(A_{w_\ell,r} z_\ell)_j|^2 + \varepsilon}} \right) (\overline{A}_{w_\ell,r})_{(j)} (A_{w_\ell,r} z_\ell)_j \\ &= \sum_{r \in \mathcal{R}} A^*_{w_\ell,r} \left(I_d - \frac{\sqrt{Y_{(r)} + \varepsilon}}{\sqrt{\sum_{\ell \in [L]} |A_{w_\ell,r} z_\ell|^2 + \varepsilon}} \right) A_{w_\ell,r} z_\ell \\ &= A^*_{w_\ell} \left(I_{dL} - \frac{\sqrt{y + \varepsilon}}{\sqrt{\sum_{\ell \in [L]} |A_{w_\ell} z_\ell|^2 + \varepsilon}} \right) A_{w_\ell} z_\ell. \end{aligned}$$

Furthermore, by Lemma 3.5.1 we have

$$\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{L}_{2,\varepsilon}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \leq \left\| \sum_{r \in \mathcal{R}} \sum_{j \in [d]} Q^w_{r,j} \right\|_\infty \left\| \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \right\|_2^2.$$

The sum of the matrices in the spectral norm is equal to

$$\begin{aligned} \sum_{r \in \mathcal{R}} \sum_{j \in [d]} Q^w_{r,j} &= \begin{bmatrix} \sum_{r \in \mathcal{R}} \sum_{j \in [d]} a_{w_0,r,j} a^*_{w_0,r,j} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sum_{r \in \mathcal{R}} \sum_{j \in [d]} a_{w_{L-1},r,j} a^*_{w_{L-1},r,j} \end{bmatrix} \\ &= \begin{bmatrix} A^*_{w_0} A_{w_0} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & A^*_{w_{L-1}} A_{w_{L-1}}, \end{bmatrix} \end{aligned}$$

and for the block diagonal matrices the spectral norms satisfies

$$\left\| \sum_{r \in \mathcal{R}} \sum_{j \in [d]} Q^w_{r,j} \right\|_\infty = \max_{\ell \in [L]} \left\| A^*_{w_\ell} A_{w_\ell} \right\|_\infty = \max_{\ell \in [L]} \left\| A_{w_\ell} \right\|_\infty^2. \tag{5.16}$$

Each matrix A_{w_ℓ} can be decomposed as

$$A_{w_\ell} = \begin{bmatrix} F_\ell D_{w_\ell,r_1} \\ \vdots \\ F_\ell D_{w_\ell,r_R} \end{bmatrix} = \begin{bmatrix} F_\ell & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & F_\ell \end{bmatrix} \cdot \begin{bmatrix} D_{w_\ell,r_1} \\ \vdots \\ D_{w_\ell,r_R} \end{bmatrix} =: \tilde{F}_\ell D_{w_\ell}$$

and, thus, using the properties of the spectral norm and block diagonal matrices we get

$$\|A_{w_\ell}\|_\infty^2 \leq \left\|\tilde{F}_\ell\right\|_\infty^2 \|D_{w_\ell}\|_\infty^2 = \|F_\ell\|_\infty^2 \|D_{w_\ell}\|_\infty^2 . \tag{5.17}$$

Moreover,

$$\|D_{w_\ell}\|_\infty^2 = \left\|D_{w_\ell}^* D_{w_\ell}\right\|_\infty = \left\|\sum_{r\in\mathcal{R}} D_{w_\ell,r}^* D_{w_\ell,r}\right\|_\infty ,$$

and since the resulting matrix is diagonal, we have

$$\|D_{w_\ell}\|_\infty^2 = \nu_\ell^2 \max_{k\in[d]} \sum_{r\in\mathcal{R}} |(w_\ell)_{k-r}|^2 \mathcal{I}_{k-r\in[0,d]}. \tag{5.18}$$

Combining (5.16), (5.17) and (5.18), leads to

$$\left\|\sum_{r\in\mathcal{R}}\sum_{j\in[d]} Q_{r,j}^w\right\|_\infty \leq \max_{\ell\in[L]} \nu_\ell^2 \|F_\ell\|_\infty^2 \max_{k\in[d]} \sum_{r\in\mathcal{R}} |(w_\ell)_{k-r}|^2 \mathcal{I}_{k-r\in[0,d]}.$$

From the proof of Lemma 4.2.1, we deduce that the gradient of $\alpha_T \|z\|_2^2$ is given by $\alpha_T z$ and its Hessian satisfies

$$\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2(\alpha_T \|z\|_2^2) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \leq \alpha_T \left\|\begin{bmatrix} u \\ \bar{u} \end{bmatrix}\right\|_2^2 .$$

Finally, the smoothness penalty term $S(z)$ can be rewritten as

$$S(z) = \sum_{\ell\in[L-1]} \kappa_\ell (z_{\ell+1}^* - z_\ell^*)(z_{\ell+1} - z_\ell) \tag{5.19}$$

$$= \sum_{\ell\in[L-1]} \kappa_\ell (z_{\ell+1}^* z_{\ell+1} + z_\ell^* z_\ell - z_\ell^* z_{\ell+1} - z_{\ell+1}^* z_\ell) = \sum_{\ell\in[L-1]} \kappa_\ell z^* (K^\ell \otimes I_d) z, \tag{5.20}$$

where $\otimes$ denotes the tensor product (2.3) and $K^\ell \in \mathbb{R}^{L\times L}$ is a matrix with four non-zero entries

$$K_{\ell,\ell}^\ell = 1, \quad K_{\ell+1,\ell+1}^\ell = 1, \quad K_{\ell,\ell+1}^\ell = -1, \quad K_{\ell+1,\ell}^\ell = -1.$$

From (5.19), we compute the gradient with respect to z_ℓ as

$$\nabla_{z_\ell} S(z) = \kappa_{\ell-1}(z_\ell - z_{\ell-1})\mathcal{I}_{\ell>0} + \kappa_\ell (z_\ell - z_{\ell+1})\mathcal{I}_{\ell<L-1}.$$

For the bound (2.17), the sum in (5.20) is combined into a single matrix

$$S(z) = z^* \left(\sum_{\ell\in[L-1]} \kappa_\ell K^\ell \otimes I_d\right) z = z^* (K \otimes I_d) z,$$

and the application of Lemma 4.2.2 gives

$$\begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{S}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \leq \left\|\begin{bmatrix} K\otimes I_d & O_{dL} \\ O_{dL} & K\otimes I_d \end{bmatrix}\right\|_\infty \left\|\begin{bmatrix} u \\ \bar{u} \end{bmatrix}\right\|_2^2 .$$

By the properties of block diagonal matrices and Proposition 2.1.2, we have

$$\left\| \begin{bmatrix} K \otimes I_d & O_{dL} \\ O_{dL} & K \otimes I_d \end{bmatrix} \right\|_\infty = \|K \otimes I_d\|_\infty = \|K\|_\infty \|I_d\|_\infty = \|K\|_\infty .$$

Consequently, using the linearity of the Wirtinger derivatives, we obtain the desired gradient formula and the estimate

$$\begin{aligned} \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{J}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} &= \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \mathcal{L}_{2,\varepsilon}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} + \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 (\alpha_T \|z\|_2^2) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} + \begin{bmatrix} u \\ \bar{u} \end{bmatrix}^* \nabla^2 \alpha_S \mathcal{S}(z) \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \\ &\leq L_w \left\| \begin{bmatrix} u \\ \bar{u} \end{bmatrix} \right\|_2^2, \end{aligned}$$

for all $z, u \in \mathbb{C}^{dL}$. □

Combining the results of Lemma 5.4.1 with Theorem 3.5.4, we obtain the following corollary regarding the convergence of Amplitude Flow for polychromatic ptychography.

Corollary 5.4.2. *Consider the family $\mathcal{Q}^w = \{Q^w_{r,j}\}_{j\in[d],r\in\mathcal{R}}$ defined by (5.14) and the objective function $\mathcal{J} = \mathcal{J}(\,\cdot\,; \mathcal{Q}^w, \varepsilon, \alpha_T, \alpha_S, \kappa)$ given by (5.11) with parameters $\varepsilon, \alpha_T, \alpha_S, \kappa_\ell \geq 0$, $\ell \in [L-1]$. Fix a constant learning rate $0 < \mu_c \leq L_w^{-1}$ with L_w defined in Lemma 5.4.1. Let $z^0 \in \mathbb{C}^{dL}$ be arbitrary. Then, for a sequence $\{z^t\}_{t\geq 0}$ given by*

$$z^{t+1} = z^t - \mu_t \nabla_z \mathcal{J}(z^t),$$

with learning rates $\mu_t = \mu(\mathcal{J}, z^t, \tau, \mu_c, N)$ determined by Algorithm 1, we have

$$\mathcal{J}(z^{t+1}) - \mathcal{J}(z^t) \leq -\mu_t \left\| \nabla_z \mathcal{J}(z^t) \right\|_2^2,$$

for all $t \geq 0$. In particular,

$$\lim_{t\to\infty} \left\| z^{t+1} - z^t \right\|_2^2 = 0 \quad \text{and} \quad \min_{t\in[T]} \left\| z^{t+1} - z^t \right\|_2^2 \leq \frac{\mathcal{J}(z^0)}{T L_w},$$

for all $T \geq 1$.

5.5 Alternating Amplitude Flow for blind polychromatic ptychography

The Amplitude Flow algorithm can also be combined with an alternating minimization to address blind polychromatic reconstruction as in Section 4.2. That is, we consider a loss function

$$\begin{aligned} \mathcal{K}(z, v) := \sum_{r\in\mathcal{R}} \sum_{j\in[d]} \left| \sqrt{\sum_{\ell\in[L]} \nu_\ell^2 \left| \sum_{k\in[d]} (z_\ell)_k (v_\ell)_{k-r} \mathcal{I}_{k-r\in[0,d]} e^{-\frac{2\pi i \nu_\ell k j}{\nu_0 d}} \right|^2 + \varepsilon} - \sqrt{Y_{r,j} + \varepsilon} \right|^2 \\ + \alpha_T \|z\|_2^2 + \beta_T \|v\|_2^2 + \alpha_S S(z; \kappa^\alpha) + \beta_S S(v; \kappa^\beta), \end{aligned} \tag{5.21}$$

with parameters $\varepsilon, \alpha_T, \alpha_S, \beta_T, \beta_S \geq 0$ and $v = (v_0^T, \dots, v_{L-1}^T)^T$ representing all components of the window as one long vector. The frequency smoothness penalty for object $S(z)$ and window $S(w)$ use parameters $\kappa_\ell^\alpha \geq 0$ and $\kappa_\ell^\beta \geq 0$, $\ell \in [L-1]$, respectively. The corresponding matrices (5.15) are denoted by K^α and K^β.
For a minimization of $\mathcal{K}$ with respect to the object variable, we observe that

$$\mathcal{K}(z, v) = \mathcal{J}(z; \mathcal{Q}^v, \varepsilon, \alpha_T, \alpha_S, \kappa^\alpha) + \beta_T \|v\|_2^2 + \beta_S S(v; \kappa^\beta), \tag{5.22}$$

and, thus, minimization of $\mathcal{K}(z, v)$ as a function of the object z is equivalent to the minimization of $\mathcal{J}(z; \mathcal{Q}^v, \varepsilon, \alpha_T, \alpha_S, \kappa^\alpha)$, which was discussed in the previous section.
For a minimization of $\mathcal{K}$ with respect to the window variable, the measurements are first rewritten as

$$I_{r,j} = \sum_{\ell \in [L]} \nu_\ell^2 \left| \sum_{k \in [d]} (x_\ell)_{k+r} (w_\ell)_k \mathcal{I}_{k+r \in [0,d]} e^{-\frac{2\pi i \nu_\ell k j}{\nu_0 d}} \right|^2, \quad r \in \mathcal{R}, j \in [d].$$

Then, in analogy to the previous section, $I_{j,r}$ is presented as a set of quadratic measurements of the form $w^* Q_{r,j}^x w$ with $w = (w_0^T, \dots, w_{L-1}^T)^T$,

$$Q_{j,r}^x = \begin{bmatrix} a_{x_0,r,j} a_{x_0,r,j}^* & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{x_{L-1},r,j} a_{x_{L-1},r,j}^* \end{bmatrix} \in \mathbb{H}^{dL}, \tag{5.23}$$

and

$$(a_{x_\ell,r,j})_k = \nu_\ell (\overline{x}_\ell)_{k+r} \mathcal{I}_{k+r \in [0,d]} e^{\frac{2\pi i \nu_\ell k j}{\nu_0 d}}, \quad k \in [d].$$

Consequently, for the minimization with respect to the window, the loss function $\mathcal{K}$ is rewritten as

$$\mathcal{K}(z, v) = \mathcal{J}(v; \mathcal{Q}^z, \varepsilon, \beta_T, \beta_S, \kappa^\beta) + \alpha_T \|z\|_2^2 + \alpha_S S(z; \kappa^\alpha), \tag{5.24}$$

where $\mathcal{Q}^z = \{Q_{r,j}^z\}_{r \in \mathcal{R}, j \in [d]}$ is the family of positive semidefinite defined via (5.23) for z. Therefore, the minimization of $\mathcal{K}(z, v)$ with respect to the window is equivalent to a minimization of $\mathcal{J}(v; \mathcal{Q}^z, \varepsilon, \beta_T, \beta_S, \kappa^\beta)$ and can be performed via gradient descent. Its convergence guarantees are analogous to Corollary 5.4.2.

Corollary 5.5.1. *Consider the family $\mathcal{Q}^z = \{Q_{j,r}^z\}_{j \in [d], r \in \mathcal{R}}$ defined by (5.23) and the loss function $\mathcal{J} = \mathcal{J}(\cdot; \mathcal{Q}^z, \varepsilon, \beta_T, \beta_S)$ given by (5.11) with parameters $\varepsilon, \beta_T, \beta_S, \kappa_\ell^\beta \geq 0$, $\ell \in [L-1]$. Let*

$$L_z := \max_{\ell \in [L]} \nu_\ell^2 \|F_\ell\|_\infty^2 \max_{k \in [d]} \sum_{r \in \mathcal{R}} |(z_\ell)_{k+r}|^2 \mathcal{I}_{k+r \in [0,d]} + \beta_T + \beta_S \left\| K^\beta \right\|_\infty,$$

with matrices F_ℓ and K^β defined as in (5.13) and (5.15), respectively, and fix a constant learning rate $0 < \nu_c \leq L_z^{-1}$. Then, for a sequence $\{v^t\}_{t \geq 0}$ given by

$$v^{t+1} = v^t - \nu_t \nabla_v \mathcal{J}(v^t),$$

with arbitrary $v^0 \in \mathbb{C}^{dL}$ *and learning rates* $\nu_t = \nu(\mathcal{J}, v^t, \tau, \nu_c, N)$ *determined by Algorithm 1, we have*

$$\mathcal{J}(v^{t+1}) - \mathcal{J}(v^t) \leq -\nu_t \left\|\nabla_v \mathcal{J}(v^t)\right\|_2^2,$$

for all $t \geq 0$. *In particular,*

$$\lim_{t\to\infty} \left\|v^{t+1} - v^t\right\|_2^2 = 0 \quad \textit{and} \quad \min_{t\in[T]} \left\|v^{t+1} - v^t\right\|_2^2 \leq \frac{\mathcal{J}(v^0)}{TL_z},$$

for all $T \geq 1$.

We note that in view of (5.22) and (5.24),

$$\begin{bmatrix}\nabla_z \mathcal{K}(z,v)\\ \nabla_v \mathcal{K}(z,v)\end{bmatrix} = \begin{bmatrix}\nabla_z J(z;\mathcal{Q}^v,\varepsilon,\alpha_T,\alpha_S,\kappa^\alpha)\\ \nabla_v \mathcal{J}(v;\mathcal{Q}^z,\varepsilon,\beta_T,\beta_S,\kappa^\beta)\end{bmatrix},$$

where in analogy to Lemma 5.4.1, the gradient of $\mathcal{J}(v;\mathcal{Q}^z,\varepsilon,\beta_T,\beta_S,\kappa^\beta)$ with respect to components v_ℓ is given by

$$\begin{aligned}\nabla_{v_\ell}\mathcal{J}(v;\mathcal{Q}^z,\varepsilon,\beta_T,\beta_S,\kappa^\beta) = A_{z_\ell}^*\left(1 - \frac{\sqrt{y+\varepsilon}}{\sqrt{\sum_{\ell\in[L]}|A_{z_\ell}v|^2+\varepsilon}}\right)A_{z_\ell}v_\ell\\ + \beta_T v_\ell + \beta_S(\kappa_{\ell-1}^\beta(v_\ell - v_{\ell-1})\mathcal{I}_{\ell>0} + \kappa_\ell^\beta(v_\ell - v_{\ell+1})\mathcal{I}_{\ell<L-1}),\end{aligned}$$

where $A_{z_\ell} \in \mathbb{C}^{dR\times d}$ are matrices with rows $a_{z_\ell,r,j}^*$.
Now, as the minimization with respect to each components is established, we can formulate the alternating Amplitude Flow for blind polychromatic ptychography.

Algorithm 13: Alternating Amplitude Flow for blind polychromatic ptychography

Input : Measurements Y as in (5.9), number of iterations $T \in \mathbb{N}$,
number of object and window subiterations $T_z \in \mathbb{N}$ and $T_v \in \mathbb{N}$,
parameters $\varepsilon, \alpha_T, \alpha_S, \beta_T, \beta_S, \kappa_\ell^z, \kappa_\ell^v \geq 0$, $\ell \in [L-1]$,
initial guesses $z^0, v^0 \in \mathbb{C}^{dL}$, AG parameters $0 < \tau < 1$, $N \in \mathbb{N}\cup\{0\}$.
Output: $z, v \in \mathbb{C}^{dL}$.
for $t = 1, \ldots, T$ **do**
  Let $z^{t,0} = z^{t-1}$.
  Set $\mu_{t,c} = L_{v^{t-1}}^{-1}$ as in Lemma 5.4.1.
  for $i \in [T_z]$ **do**
    Determine $\mu_{t,i} = \mu_{t,i}(\mathcal{K}, z^{t,i}, \tau, \mu_{t,c}, N)$ via Algorithm 1.
    Update $z^{t,i+1} = z^{t,i} - \mu_{t,i}\nabla_z \mathcal{K}(z^{t,i}, v^{t-1})$.
  Let $z^t = z^{t,T_z}$ and $v^{t,0} = v^{t-1}$
  Set $\nu_{t,c} = L_{z^t}^{-1}$ as in Corollary 5.5.1.
  for $j \in [T_v]$ **do**
    Determine $\nu_{t,j} = \nu_{t,j}(\mathcal{K}, v^{t,j}, \tau, \nu_{t,c}, N)$ via Algorithm 1.
    Update $v^{t,j+1} = v^{t,j} - \nu_{t,j}\nabla_v \mathcal{K}(z^t, v^{t,j})$.
  Let $v^t = v^{t,T_v}$.
return $z = z^T, v = v^T$.

For Algorithm 13 similarly to Algorithm 10, we are able to show that the learning rates are always finite.

Lemma 5.5.2. *Let $\varepsilon, \alpha_T, \beta_T > 0$. Then, for all $z, v \in \mathbb{C}^{dL}$ we have $L_v \geq \alpha_T > 0$ and $L_z \geq \beta_T > 0$. Furthermore, the learning rates $\mu_{t,i}$ and $\nu_{t,j}$, $i \in [T_z]$, $j \in [T_v]$, $t \geq 1$, determined by Algorithm 13 are bounded by τ^{-N}/α_T and τ^{-N}/β_T, respectively.*

Furthermore, we derive the convergence guarantees for Algorithm 13.

Theorem 5.5.3. *Let $\mathcal{K} : \mathbb{C}^{dL} \times \mathbb{C}^{dL} \to [0, \infty)$ be defined as in (5.21) with $\varepsilon, \alpha_T, \beta_T > 0$ and $\alpha_S, \beta_S, \kappa_\ell^\alpha, \kappa_\ell^\beta$, $\ell \in [L-1]$. Consider the two sequences $\{z^t\}_{t\geq 0}, \{v^t\}_{t\geq 0}$ generated by Algorithm 13 with arbitrary starting points $z^0, v^0 \in \mathbb{C}^{dL}$ and let $\mu_{t,i}$ and $\nu_{t,i}$ be learning rates determined by Algorithm 13. Then, for each subiteration of Algorithm 13 we have*

$$\begin{aligned}\mathcal{K}(z^{t,i+1}, v^{t-1}) - \mathcal{K}(z^{t,i}, v^{t-1}) &\leq -\mu_{t,i} \left\|\nabla_z \mathcal{K}(z^{t,i}, v^{t-1})\right\|_2^2 \\ \mathcal{K}(z^t, v^{t,j+1}) - \mathcal{K}(z^t, v^{t,j}) &\leq -\nu_{t,j} \left\|\nabla_v \mathcal{K}(z^t, v^{t,j})\right\|_2^2 .\end{aligned}$$

for every $t \geq 1$ and $i \in [T_z]$, $j \in [T_v]$.
Moreover,

$$\lim_{t\to\infty} \left\|\nabla_z \mathcal{K}(z^t, v^t)\right\|_2^2 + \left\|\nabla_v \mathcal{K}(z^t, v^t)\right\|_2^2 = 0,$$

where the rate of convergence is dominated by

$$\frac{\max\{\alpha_T^{-1}, \beta_T^{-1}\} \max_{\ell\in[L]} \nu_\ell^2 \|F_\ell\|_\infty^2 \mathcal{K}^2(z^0, v^0) + \max\{\alpha, \beta\}\mathcal{K}(z^0, v^0)}{T \min\{T_z, T_v\}},$$

with $\alpha = \alpha_T + \alpha_S \|K^\alpha\|_\infty$, $\beta = \beta_T + \beta_S \|K^\beta\|_\infty$ and the matrices K^α, K^β as in (5.15).

The proofs are analogous to Lemma 4.2.5 and Theorem 4.2.6.

Notes and References. *Polychromatic ptychography is significantly less studied in the literature compared to monochromatic ptychography. Several algorithms [24, 190, 191, 192, 193, 194] were proposed for polychromatic reconstruction, all based either on minimization of the loss function $\mathcal{K}$ without regularization or its analogue for Poisson maximum likelihood function and employ gradient-based approaches. Algorithm in [190] approaches polychromatic intensity as a convolution of separate monochromatic intensities. In [192], the gradient descent is applied for the reconstruction of the exit waves, which are later decoupled into the object and the window via least squares minimization. The gradient descent is combined with the independent component analysis in [191]. In [24] the authors establish the generalization of ePIE algorithm discussed in Section 4.3 for polychromatic light. We note that it is equivalent to the stochastic gradient descent, in analogy to Theorem 4.3.1. In [193], the idea of ePIE is extended by an inclusion of additional assumptions on the object. Another extension of polychromatic ePIE can be found in [150] for illuminations with multiple orthogonal probes. There also exists a deep learning approach [194] based on generative priors. In comparison to these works, the main advantage of our results, based on [195], are the supporting convergence guarantees.*
The measurements of the form (5.5) arise not only in the case of polychromatic light, but also when the multiple spatially incoherent beams are used in the ptychographic experiment [196].
Note that we mostly build up upon the results established in Section 3.5.1 and Section 4.2.3, which was the main motivation in considering the general quadratic measurements in the

first place. In comparison to blind monochromatic ptychography in Section 4.2.4, we do not provide a Tikhonov regularization-free algorithm for blind polychromatic ptychography. While it is possible to derive an analogy of Algorithm 11 and Theorem 4.2.9 if $\alpha_S = \beta_S = 0$*, for other cases the scale invariance of the loss function* (4.21) *does not hold for* $\mathcal{K}$*. Finally, the main assumption behind the measurement formula* (5.5) *is that the acquisition time of the detector* T *is significantly larger than* $\max_{\ell_1 \neq \ell_2} |\nu_{\ell_1} - \nu_{\ell_2}|^{-1}$*. If it is false, the intensity will include extra terms, but the problem can still be transformed into a recovery from quadratic measurements of the form* $x^* Q_{r,j} x$*.*

Chapter 6

Numerical experiments

6.1 Monochromatic ptychography

In this section we perform numerical experiments to study the performance of algorithms for monochromatic ptychography discussed in this thesis.

6.1.1 Experimental setup

Our experiments are based on synthetic objects. That is we generate piecewise constant objects in $\mathbb{C}^d$ with dimension $d = 256$ consisting of $b \leq d$ constant intervals. In order to do so, we sample $b-1$ indices $\{j_k\}_{k\in[b-1]}$ from $[d]$ uniformly at random without replacement. The indices are ordered in increasing order $\{j_{(k)}\}_{k\in[b-1]}$ and we additionally assign $j_{(-1)} = 0$ and $j_{(b-1)} = d$. These indices form intervals $[j_{(k-1)}, j_{(k)})$, $k \in [b]$, each corresponding to a constant value of the generated object x. The values $v \in \mathbb{C}^b$ for the intervals are generated by sampling b independent random complex Gaussian variables, $v_k = \mathcal{N}(0,1) + i\mathcal{N}(0,1)$. Then, the object x is compiled by

$$x_j = 2v_k / \|v\|_\infty , \quad j \in [j_{(k-1)}, j_{(k)}) \text{ for some } k \in [b].$$

An example of such an object can be seen in Figure 6.1a.
As a window (Figure 6.1b), we use the discrete localized analogue of (1.15), which is given by

$$w_j = \begin{cases} \exp\left(-\frac{(j-\mu)^2}{2\sigma^2\delta^2} + i\frac{\pi(j-\mu)^2}{2\sigma^2\delta^2}\right), & j \in [\delta], \\ 0, & j \notin [\delta], \end{cases} \tag{6.1}$$

(a) Generated object with 16 intervals and $d = 256$.

(b) Window, $\delta = 32$.

Figure 6.1: An example of an object and the window (6.1). The top line represents the amplitudes with white corresponding to the largest magnitudes and black to zero. The bottom line represents the phases on the RGB color wheel as in Figure 6.2.

Figure 6.2: The representation of the phases $e^{i\varphi}$ by the color wheel for angles $\varphi \in [0, 2\pi)$.

with $\mu = -0.5 + 0.5\delta$ and $\sigma^2 = 0.05$.
The ptychographic measurements are then computed according to (PTY) with a frequency subsampling parameter m and a set of shifts $\mathcal{R}$, which will vary between experiments and will be specified separately. For most experiments the data is corrupted by Poisson noise, which models random behavior of photons during the experiment. More precisely, we will set

$$Y_{j,r} = \frac{m \|w\|_2^2}{K} \mathrm{Pois}\left(\frac{K}{m \|w\|_2^2} I_{j,r}^m \right),$$

where K denotes the number of photons used for a single illumination (shift position) during the experiment (Figure 6.3). The number of photons K takes values $\{10^2, 10^3, \ldots, 10^9\}$ and serves as a measure of the noise level, in the sense that a higher K corresponds to less noisy measurements. The number of photons can be roughly translated to the signal-to-noise ratio (SNR)

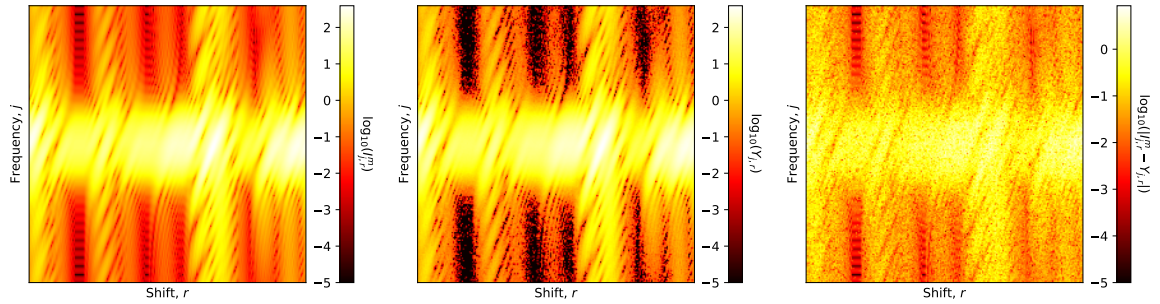
$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\|I^m\|_F^2}{\|Y - I^m\|_F^2} \right)$$

often used in the literature. In our case, the values of K roughly correspond to the values $\{5, 15, \ldots, 75\}$ of SNR.
As a measure of reconstruction quality, the relative error (RE)

$$\mathrm{RE} = \frac{\mathrm{dist}(z, x)}{\|x\|_2}$$

is used, where $\mathrm{dist}(z, x)$ is given by (3.11) and accounts for the global phase ambiguity. Another relevant metric is the relative measurement error (RME)

$$\mathrm{RME} = \frac{\sum_{j\in[m]} \sum_{r\in\mathcal{R}} ||F_m P_m [S_{-r} z \circ w]| - \sqrt{Y}_{j,r}|^2}{\sum_{j\in[m]} \sum_{r\in\mathcal{R}} |\sqrt{Y}_{j,r}|^2} = \frac{\mathcal{L}_2(z; A)}{\mathcal{L}_2(\mathbb{0}_d; A)},$$



(a) Noiseless I^m. (b) Noisy Y. (c) Noise $|N|$.

Figure 6.3: Ptychographic measurements (PTY) for the object and the window from Figure 6.1 with $m = d$ and $\mathcal{R} = [d]$ and number of photons $K = 10^5$ and SNR ≈ 35.

Table 6.1: The distribution of the singular values in the inversion step of BPR for $d = 256$ and the window (6.1).

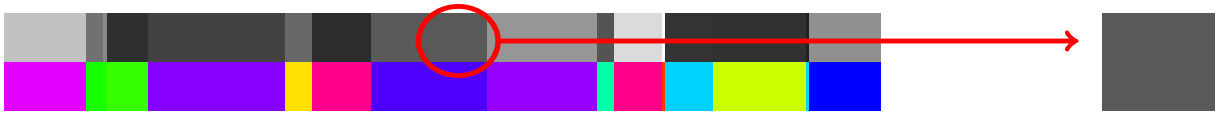
$\{0\}$	$[10^{-12}, 10^{-3})$	$[10^{-3}, 10^{-2.5})$	$[10^{-2.5}, 10^{-2})$	$[10^{-2}, 10^{-1.5})$	$[10^{-1.5}, 0.1)$	$[0.1, 1)$	$[1, 13)$	Total
16	116	254	1080	2075	2799	1177	675	8192

where $\mathcal{L}_2(\,\cdot\,; A)$ is the amplitude-based squared loss (3.13) with the measurement matrix A given by (3.9) and $\mathbb{0}_d$ denotes the zero vector in $\mathbb{C}^d$.
In the upcoming plots each data point is the average of the chosen metric based on 30 trials. All experiments are performed in Python on a laptop running Windows 10 Pro with an Intel(R) Core(TM) i7-8550U processor and with 16 GB RAM.

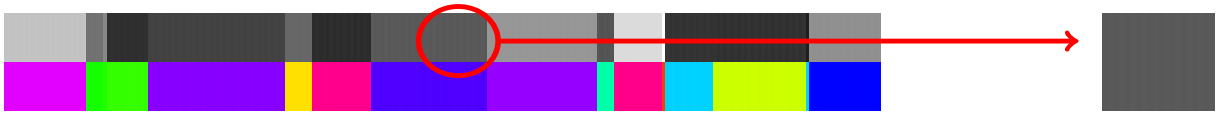
6.1.2 Block Phase Retrieval

6.1.2.1 Noiseless reconstruction

The first experiment explores reconstruction capabilities of the Block Phase Retrieval algorithm (BPR) (see Section 3.6) with Diagonal Magnitude Estimation and unweighted angular synchronization, which was proposed in [40]. For these trials no subsampling in frequencies is performed and all possible shifts positions are observed, so that $m = d$ and $\mathcal{R} = [d]$.
We start by performing a reconstruction with BPR in the noiseless case (Figure 6.4). Note that despite the fact that the window (6.1) does not fit into the scope of Example 3.6.9, singularities occur during the inversion step, which can also be seen in Table 6.1. Therefore, we use BPR with the subspace completion technique ($\mathrm{BPR} + \mathrm{SC}_\varepsilon$) discussed in Section 3.6.2.2 for the perfect reconstruction in the noiseless case. For comparison, the reconstruction with the truncation of singular values $\mathrm{BPR} + \mathrm{TR}_\varepsilon$ is also shown. For both methods, $\varepsilon \geq 0$ denotes the truncation threshold.
The reconstruction $\mathrm{BPR} + \mathrm{SC}_0$ in Figure 6.4a is perfect and coincides with the original object depicted in Figure 6.1a, which works as a proof of concept for the subspace recovery. In contrast, minor high frequency artifacts can be seen in Figure 6.4b for reconstruction with $\mathrm{BPR} + \mathrm{TR}_0$ only using the truncation procedure.
Despite the perfect reconstruction in the noiseless case, $\mathrm{BPR} + \mathrm{SC}_0$ with diagonal esti-

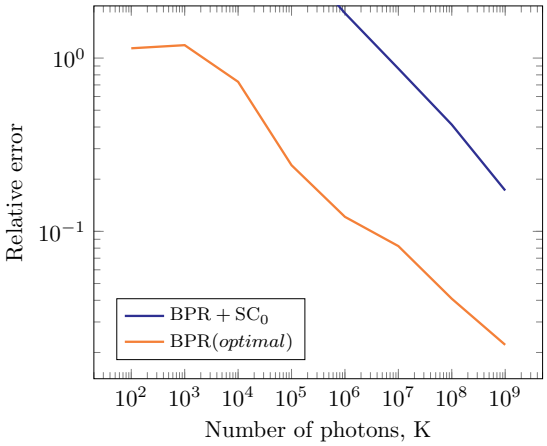


(a) $\mathrm{BPR} + \mathrm{SC}_0$.

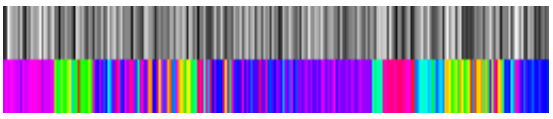


(b) $\mathrm{BPR} + \mathrm{TR}_0$.

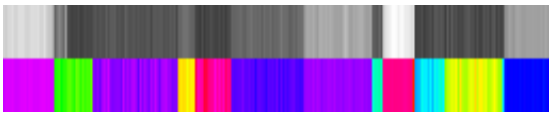
Figure 6.4: Reconstruction of the object with two versions of the BPR algorithm in the noiseless case.



(a) Relative error.



(b) Reconstruction with BPR + SC_0 for $K = 10^6$.



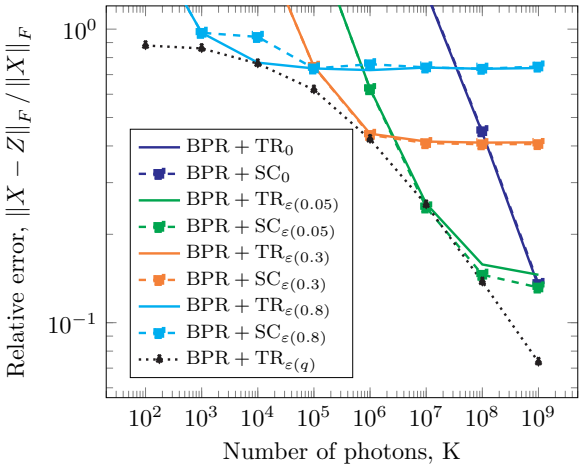
(c) Reconstruction with BPR(*optimal*) for $K = 10^6$.

Figure 6.5: Reconstruction with BPR in the presence of noise.

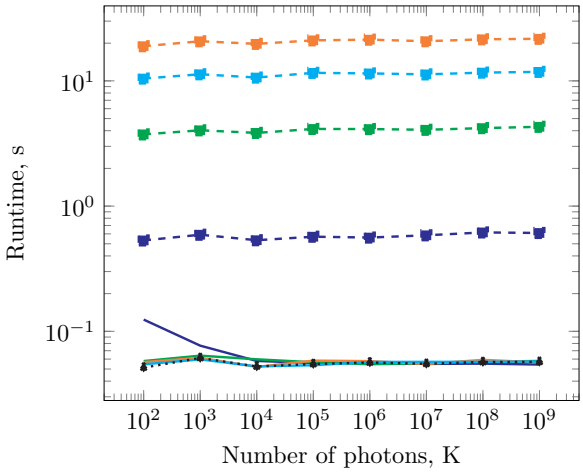
mation and unweighted angular synchronization is unstable in the presence of noise. In Figure 6.5a we observe that the relative error for BPR as in Algorithm 5 does not decrease until small noise levels. However, if the alternatives for the magnitude and the phase reconstruction are chosen and the regularization parameter is optimized, BPR achieves better noise stability as can be seen in Figure 6.5. We will specify the optimal choices for each component of BPR as we proceed with the numerical investigation.

6.1.2.2 Inversion step

The first component of the BPR algorithm is the inversion step, which reconstructs the matrix Z, an approximation to the matrix X defined in (3.31), based on the pseudoinverse operator or equivalently on the Wigner Distribution Deconvolution as discussed in



(a) Relative error.



(b) Runtime.

Figure 6.6: Performance of the truncation and subspace completion techniques for BPR. BPR + $TR_{\varepsilon(q)}$ selects q via (6.2).

Section 3.6.2.1. Zero singular values are present for the window (6.1), see Table 6.1. Thus, the inversion step can be performed with the truncation step ($\mathrm{BPR} + \mathrm{TR}_\varepsilon$ as in Algorithm 4) or with the additional subspace completion ($\mathrm{BPR} + \mathrm{SC}_\varepsilon$ as in Algorithm 5). While setting $\varepsilon = 0$ already suffices for a reconstruction with singularities, for larger values of ε both versions of BPR act as a regularization techniques. In the following ε will be defined by parameter q, so that $\varepsilon(q)$ is the empirical q-quantile and truncates $q \cdot 100\%$ of all singular values.

We note that assumption of the subspace completion stated in Theorem 3.6.11 may no longer hold and the subspace completion acts as a heuristic. That is, if more than one Fourier coefficient is lost for a single diagonal of Z, all lost coefficients are initially set to zero and then one by one are reconstructed via (3.49).

In Figure 6.6, we observe how different values of $\varepsilon(q)$ impact the reconstruction of X in terms of the relative error $\|X - Z\|_F / \|X\|_F$ and the runtime. The first outcome of Figure 6.6a is that the subspace completion has no significant improvement over truncation unless the noise is small. Furthermore, in Figure 6.6b the computation time of $\mathrm{BPR} + \mathrm{SC}_{\varepsilon(q)}$ is at least ten times larger than for $\mathrm{BPR} + \mathrm{TR}_{\varepsilon(q)}$, which is caused by the construction of the linear system (3.49). Moreover, the runtime increases with the percentage q, i.e., the number of Fourier coefficients to be recovered via the subspace completion.

Secondly, as the value of parameter q increases, the algorithm becomes more robust to noise. However, we also observe that for larger q, e.g., 0.3, if the noise level decreases, the relative error flattens. This is caused by the incorrect estimation of the lost Fourier coefficients of the diagonals. The larger is the value of q, the higher is the percentage of the truncated singular values and, thus, the larger is the loss of information and the faster the flattening occurs. In the extreme case, for $q = 0.8$, 80% of the singular values are below $\varepsilon(0.8)$ and the good reconstruction quality is no longer possible. These observations suggest that the optimal choice of q depends on the noise level represented by the number of photons K. If K is assumed to be known, e.g., from experimental metadata, the choice

$$q = 0.9e^{-\frac{|\log_{10} K - 2|^3}{45}} + 0.005 \max\{11 - \log_{10} K, 0\} \tag{6.2}$$

is approximately optimal. The selection of parameter q in (6.2) is such that for a fixed K the resulting q approximately coincides with the value providing the minimal relative error in Table B.1.

Therefore, for the rest of the experiments, we will only use truncation with the choice of $\varepsilon(q)$ with q selected via (6.2).

In Figure 6.7 we investigate how the errors $|X - Z|$ are distributed depending on the noise level and the truncation parameter. For a high regularization threshold, the errors are mainly concentrated far from the main diagonal. However, as ε decreases, the concentration of the errors shifts from the distant off-diagonals closer to the main diagonal. This behavior does not depend on the number of photons K. For a noise-dependent choice of ε this effect causes high errors for the off-diagonals and as the noise decreases, more and more off-diagonals are estimated correctly.

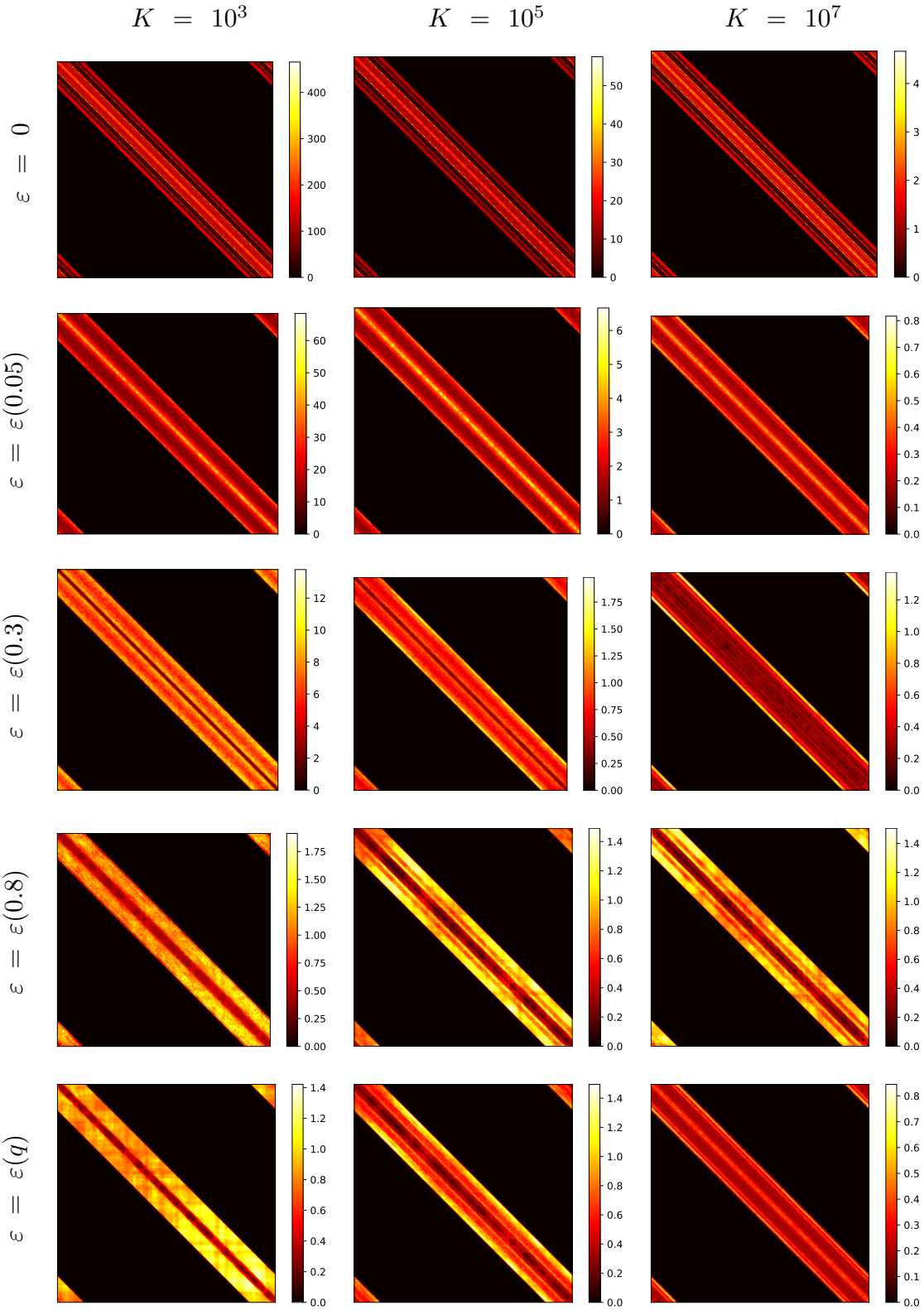
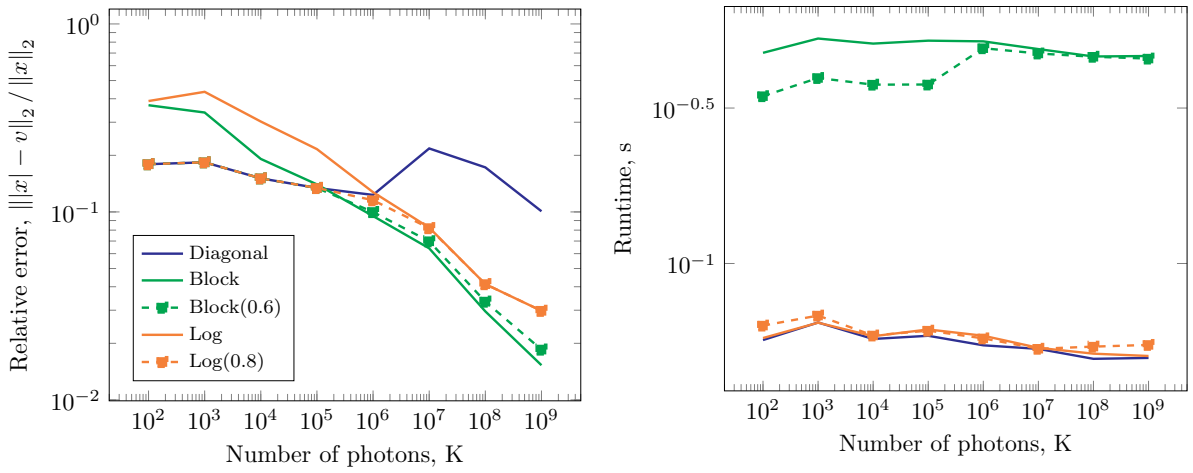


Figure 6.7: The average error $|X - Z|$ resulting from the inversion step.



(a) Relative error.

(b) Runtime.

Figure 6.8: Performance of different magnitude estimation techniques for BPR.

6.1.2.3 Magnitude estimation

For the magnitude estimation, we will consider the three methods discussed in Section 3.6.3: Diagonal, Block and Log Magnitude Estimation techniques. For the latter two methods, we will consider two variants: one, which uses all diagonals of Z and the other, which only uses a subset of diagonals.
For Block Magnitude Estimation, we will use the covering $\{\mathcal{J}_\ell^\gamma\}_{\ell\in[d]}$, where sets $\mathcal{J}_\ell^\gamma$ are defined as in (3.56). The parameter γ identifies how many diagonals are used for Block Magnitude Estimation and we will consider two choices. The first uses all diagonals $\gamma = \delta$ and, the second is an adaptive choice based on the percentage of the recovered Fourier coefficients of the diagonals. That is for each possible $\gamma = 1, \dots, \delta$ the fraction

$$\frac{\text{number of recovered Fourier coefficients for first } \gamma \text{ diagonals}}{\gamma d}$$

is computed and the value of γ corresponding to the maximal value of the fraction is selected. In the case that the maximal value of the fraction is below 0.6, we set γ as 1, so that Diagonal Magnitude estimation is used. The choice of the threshold 0.6 is based on addition numerical trials exploring reconstruction errors for different values of the threshold parameter shown in Table B.2.
Similarly, for Log Magnitude Estimation, either all diagonals are used or a subset $\mathcal{J} \subset [\delta]$, where for each diagonal $d^j(Z)$, $j \in \mathcal{J}$, at least 80% of Fourier coefficients were recovered. If none of the diagonals satisfy the 80% condition, only the main diagonal is used. Again, the choice of the threshold 80% is based on additional trials summarized in Table B.3.
Figure 6.8a shows the relative error between the magnitude estimate v and actual magnitude $|x|$ for five selected reconstruction techniques. We observe that if the number of photons K is low and, respectively, high noise is present, methods using all diagonals show worse reconstruction than those depending only on the small subset of diagonals. In fact, the adaptive methods coincide with Diagonal Magnitude Estimation for high noise level. As noise decreases, the benefit from the inclusion of more diagonals is prominent with

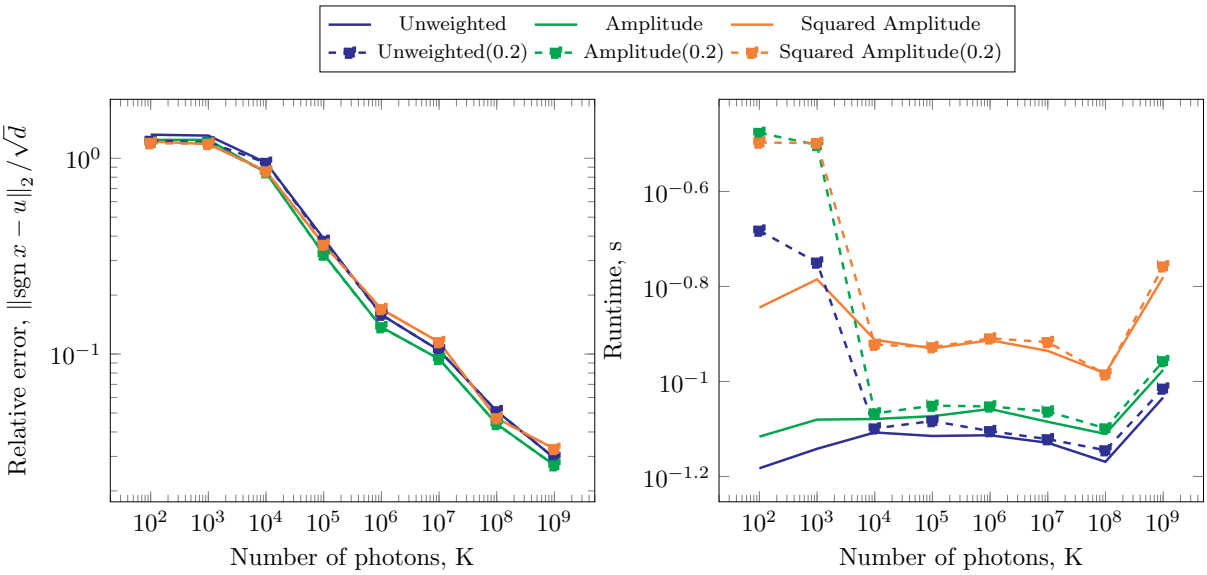


Figure 6.9: Performance of different weights for phase estimation within BPR

slightly better relative errors for Block Magnitude Estimation than for Log Magnitude Estimation. In terms of runtime (Figure 6.8b), Block Magnitude Estimation consumes more time than the other two methods, however, it is takes below one second to produce the estimate. Notably, the runtime of Log Magnitude Estimation, which uses all diagonals, is just slightly slower than the runtime of Diagonal Magnitude Estimation based only on the main diagonal of the matrix Z.
In view of the discussion above, the optimum in terms of the relative error is Block Magnitude Estimation with the adaptive selection of diagonals.

6.1.2.4 Phase estimation

For the phase estimation, we solve the angular synchronization as discussed in Section 3.6.4. We will only use the eigenvalue relaxation EIG for our experiments as the semidefinite relaxation SDP requires an impractical amount of time for a single reconstruction. In terms of weight choices, six scenarios are considered. The first three are the unweighted case $W_{k,\ell} = \mathcal{I}_{(k,\ell)\in E}$, the amplitude weights $W_{k,\ell} = |Z_{k,\ell}|\mathcal{I}_{(k,\ell)\in E}$ and the squared amplitude weights $W_{k,\ell} = |Z_{k,\ell}|^2\mathcal{I}_{(k,\ell)\in E}$, which were introduced in Section 3.6.4.5. The second triple are the same weight scenarios, however, we only use diagonals where at least 20% of the Fourier coefficients were recovered. If none of the diagonals satisfy this condition only $d^1(Z)$ is used for the phase estimation. The choice of the threshold 20% is based on additional trials with different thresholds presented in Table B.4.
Figure 6.9 shows the relative error between the obtained estimate u and the true vector of phases $\operatorname{sgn} x$ given by $\operatorname{dist}(\operatorname{sgn} x, u)/\sqrt{d}$ and the corresponding runtimes. We observe that there is no significant difference between the angular synchronization applied to a subset of the diagonals and to all diagonals. In terms of the weight choice, the smallest errors are achieved for the amplitude weights $W_{k,\ell} = |Z_{k,\ell}|\mathcal{I}_{(k,\ell)\in E}$, and while the improvement is mediocre, the runtime is only slightly longer than for the unweighted scenario. That is why we choose the amplitude weights for the rest of the experiments.

Consequently, the "optimal" BPR algorithm in Figure 6.5a is BPR + TR_ε with the regularization parameter $\varepsilon(q)$ with the quantile value q selected via (6.2), adaptive Block Magnitude Estimation and the phases, obtained via the eigenvalue relaxation of the angular synchronization with the amplitude weights. We note that the optimality has to be understood purely in the sense of empirical observations. While the "optimal" algorithm is noise-aware, i.e., depends on K, it shows the reconstruction possibilities of BPR. In the case where the number of photons is not available, the proper choice of q can be achieved by trial-and-error with the relative measurement error as a performance metric.

6.1.2.5 Heuristics for larger shift

Let us now consider the ptychographic measurements with equidistant shifts, as discussed in Section 3.6.5. That is the set of shifts $\mathcal{R} = \mathcal{R}_s$ is given by (3.111), so that each time the object is shifted by s pixels. In this case, the BPR algorithm as in Section 3.6 cannot be applied, however, its adjustment can be applied under the assumption that the object x belongs to the space $\mathbb{C}^d_s$. This assumption is not true for our generated synthetic objects and, thus, BPR as in Algorithm 7 is applied as a heuristics.
The resulting reconstruction errors for $s \in \{1, 2, 4, 8, 16\}$ are summarized in Table 6.2. We observe that for larger shift sizes s, the reconstruction error deteriorates and already for $s = 4$ is at least 0.899. Furthermore, the relative error improves as K increases from 10^2 to 10^6 and then deteriorates, which suggests that the choice of the regularization parameter $\varepsilon(q)$ with quantile q selected via (6.2) is suboptimal. In addition to the relative error presented in Table 6.2, examples of reconstructed objects are shown in Table 6.3. While the phase reconstruction quickly deteriorates with increase of s, the magnitude estimation is still quite accurate.
Additional numerical trials summarized in Table B.5 justified the hypothesis that the choice of the regularization parameter q via is suboptimal. Thus, for the experiments with $\mathcal{R}_s$, we will instead select q via

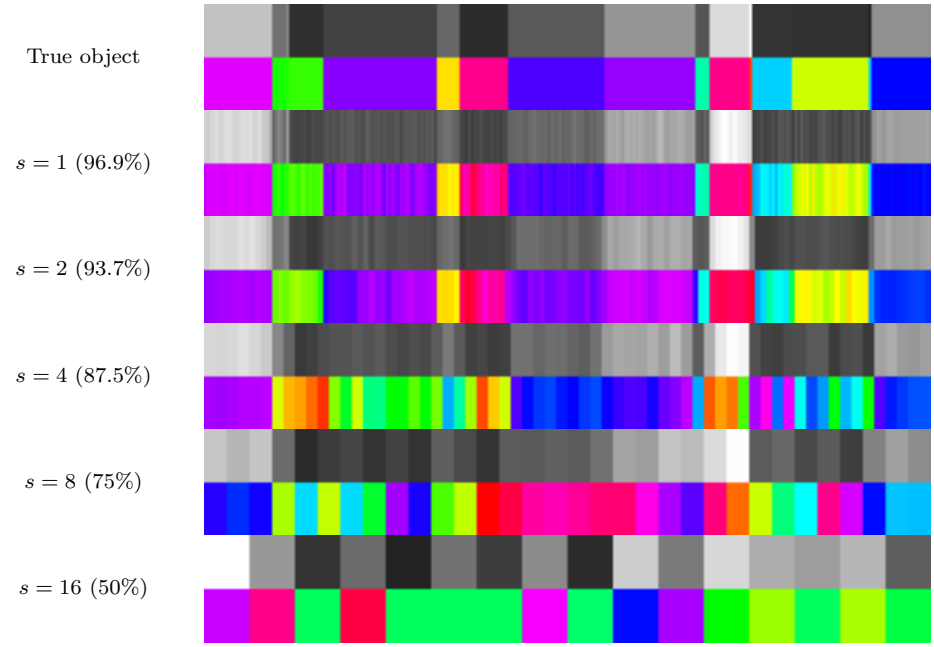
$$q = \max\{0.8 - 0.13(\log_{10} K - 2), 0.4\} \tag{6.3}$$

for $s = 4$. Despite the optimality of q, the errors are still high, since BPR still acts as a heuristic. We also note that the formula for q depends on the window w and the shift size

Table 6.2: The impact of the shift size s on the reconstruction error of BPR. The last row shows the average runtime.

K	$s = 1$	$s = 2$	$s = 4$	$s = 8$	$s = 16$
10^2	1.139	1.206	1.239	1.194	1.162
10^3	1.187	1.170	1.276	1.189	1.242
10^4	0.730	1.095	1.109	1.258	1.237
10^5	0.241	0.535	0.930	1.233	1.280
10^6	0.121	0.242	0.899	1.257	1.263
10^7	0.082	0.201	1.019	1.205	1.273
10^8	0.041	0.204	1.066	1.208	1.380
10^9	0.022	0.225	1.131	1.313	1.392
avg. times, s	0.349	0.075	0.030	0.013	0.004

Table 6.3: An example of reconstructions for different shift sizes with Algorithm 7. In brackets is the overlap, given by $(\delta - s)/\delta$. $K = 10^6$.



s and there is no unified approach towards its selection, which is an interesting direction of future research.

6.1.3 Iterative methods

In this section, we will explore the performance of the iterative methods discussed in Chapter 3. For all experiments, the shift size will be set to $s = 4$, so that $\mathcal{R} = \mathcal{R}_s$ as in Section 6.1.2.5 above. As the iterative methods are sensitive to the initial guess, we will consider two possibilities. The first is a random guess, for which each entry of the object is independently sampled from a distribution

$$z_j^{0,rand} \sim \mathcal{N}(0,1) + i\mathcal{N}(0,1), \quad j \in [d].$$

The second option is initialization with BPR (acting as a heuristic), which was numerically explored in Section 6.1.2.5. In order to identify the initialization in the plots, we will write (rand) next to the algorithm's name for the random initialization and (BPR) for the BPR heuristic.

The first iterative algorithm is Amplitude Flow (AF), which was covered in Section 3.5.1. Both for initialization of AF with random guess and with BPR, the learning rate is constant and set to $\mu_c = \|A\|_\infty^{-2}$, which is the maximal choice of a constant learning rate with guaranteed convergence according to Theorem 3.5.5. Moreover, for AF with BPR initialization, we will also select the learning rate μ_t via Armijo-Goldstein (AG) condition (Algorithm 1) with $\tau = 0.5$ and $N = 2$. This case will be denoted by AF(BPR)+AG. In all three cases, AF runs for at most $T = 5000$ iterations or until the norm of the gradient becomes smaller than 1.

The second algorithm is Error Reduction (ER) from Section 3.5.2. Similarly to AF, the ER methods stops computations after $T = 5000$ iteration or once the norm of the difference between the iterates $\|z^{t+1} - z^t\|_2$ is smaller than 10^{-3}.

The last algorithm is Ptychographic Iterative Engine (PIE) introduced in Section 3.5.3. We consider two versions of the PIE algorithm. The first, abbreviated as PIE$_{unif}$, selects the index r^t uniformly at random and runs for $T = 150000$ iterations with the maximal possible choice of parameter α according to Theorem 3.5.13. The second version uses sampling with probabilities (3.25) based on the norms of the gradients. Since the computational complexity is approximately $(2\delta - 1)/s = 15.75$ times larger, the number of iterations is reduced to $T = 15000$. The parameter α is set $\alpha = \|w\|_\infty^2 / 200 \left\| \sum_{r \in \mathcal{R}} |S_r w|^2 \right\|_\infty$.

First, let us consider how the number of photons impacts the reconstruction (Figure 6.10). Starting with the relative measurement error, there is only a small difference between random initialization and initialization with BPR for AF and ER. In contrast, PIE with random initialization stops to improve around RME of $10^{-2.4}$, while PIE combined with BPR shows RME of 10^{-3}. In both cases, RME for PIE is significantly higher than for AF and ER when the noise level is medium or low.

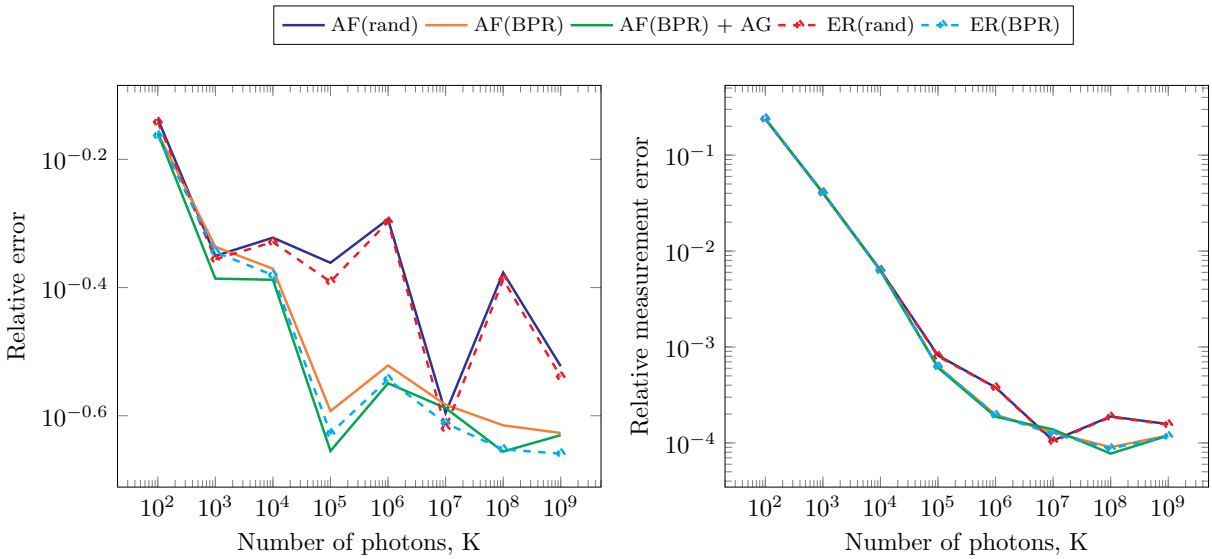
If we compare the algorithms in terms of the relative error, BPR initialization leads to a lower relative error than random initialization. As in the case of RME, the relative error for the versions of PIE is somewhat the same and is higher than for AF and ER. Comparing AF to ER, the reconstruction with the latter is slightly better in terms of RE. An additional usage of the Armijo-Goldstein condition for AF improves upon the reconstruction error, so that the reconstruction error for AF(BPR)+AG are similar to the error for ER(BPR) .

Turning to computation times (Table 6.4), we observe that there is no significant difference between the two initializations for a high noise level. However, as the noise level decreases, initialization with the BPR heuristic leads to a faster convergence of the AF and ER algorithms. The inclusion of the AG condition for AF reduces the runtime by a factor of two. The runtime of PIE$_{norm}$ is approximately the same as of PIE$_{unif}$, since the number of iterations was decreased by a factor of ten. Finally, we note that the computational cost of BPR can be neglected compared to the runtime of iterative algorithms.

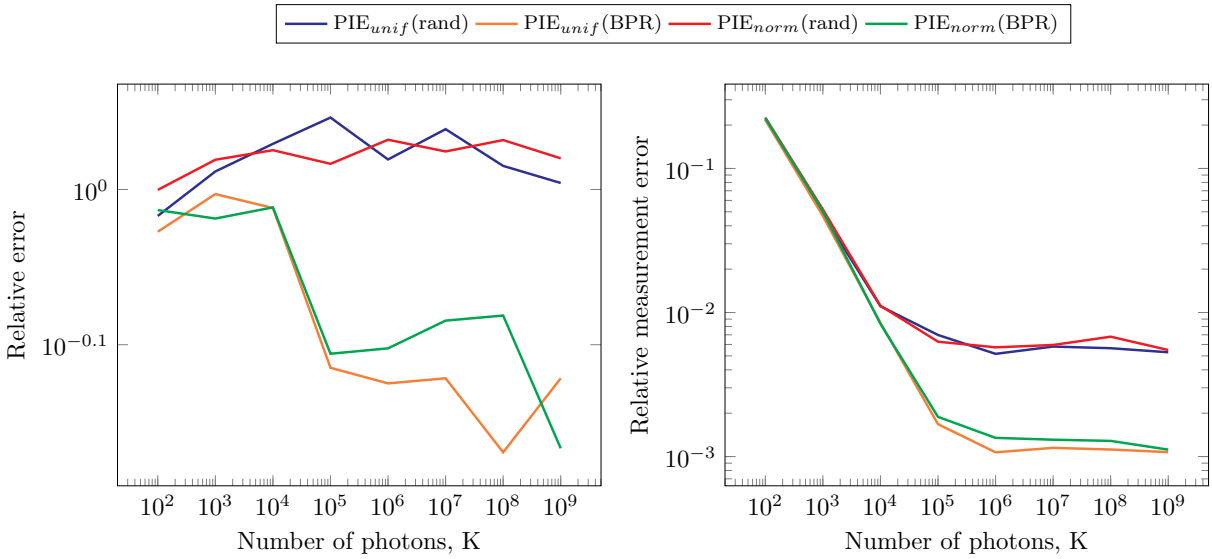
Figure 6.11 shows how the errors progress for a single reconstruction process with each

Table 6.4: The average computation times in seconds of iterative methods.

K	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
BPR	0.026	0.024	0.032	0.035	0.033	0.033	0.032	0.037
AF(rand)	25.637	26.085	25.286	24.184	25.438	27.665	29.975	33.371
AF(BPR)	26.876	25.867	24.550	23.012	20.053	21.595	23.029	22.877
AF(BPR)+AG	13.646	14.064	12.213	11.818	9.981	10.498	11.443	10.745
ER(rand)	27.898	28.467	27.418	27.140	26.187	29.335	31.039	31.244
ER(BPR)	28.806	28.476	26.609	25.784	21.503	23.754	24.771	22.478
PIE$_{unif}$(rand)	29.811	30.620	32.085	30.229	30.388	31.574	30.292	32.562
PIE$_{unif}$(BPR)	29.818	30.673	31.523	29.905	29.952	31.039	29.514	32.422
PIE$_{norm}$(rand)	30.268	29.566	28.861	29.408	30.446	29.692	30.241	30.266
PIE$_{norm}$(BPR)	30.142	29.330	28.656	29.850	30.014	29.646	30.503	30.150



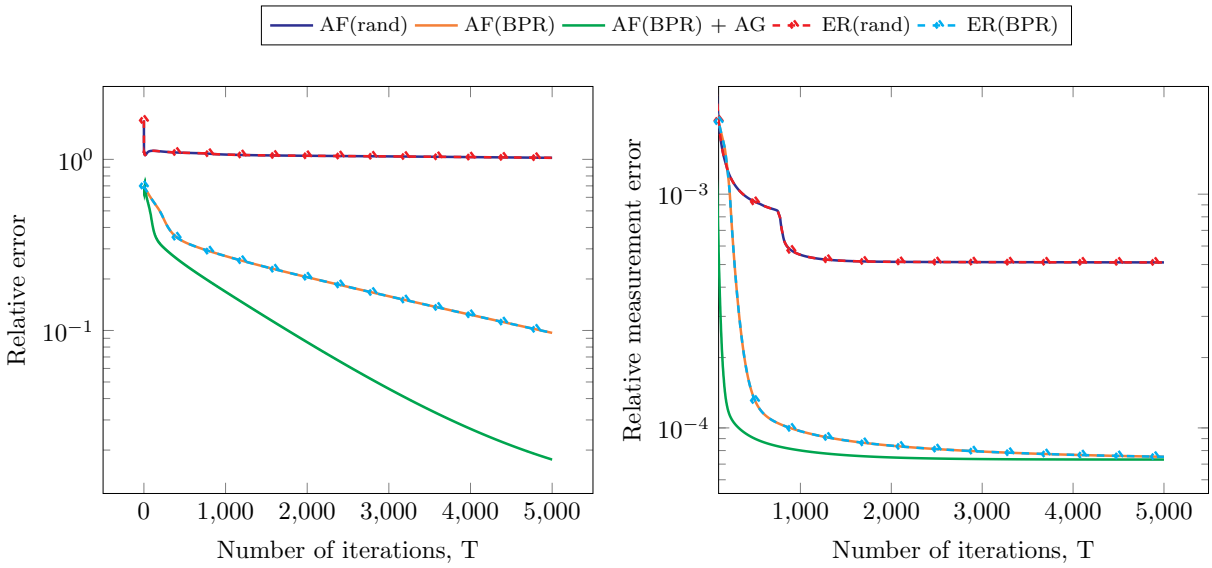
(a) Amplitude Flow and Error Reduction.



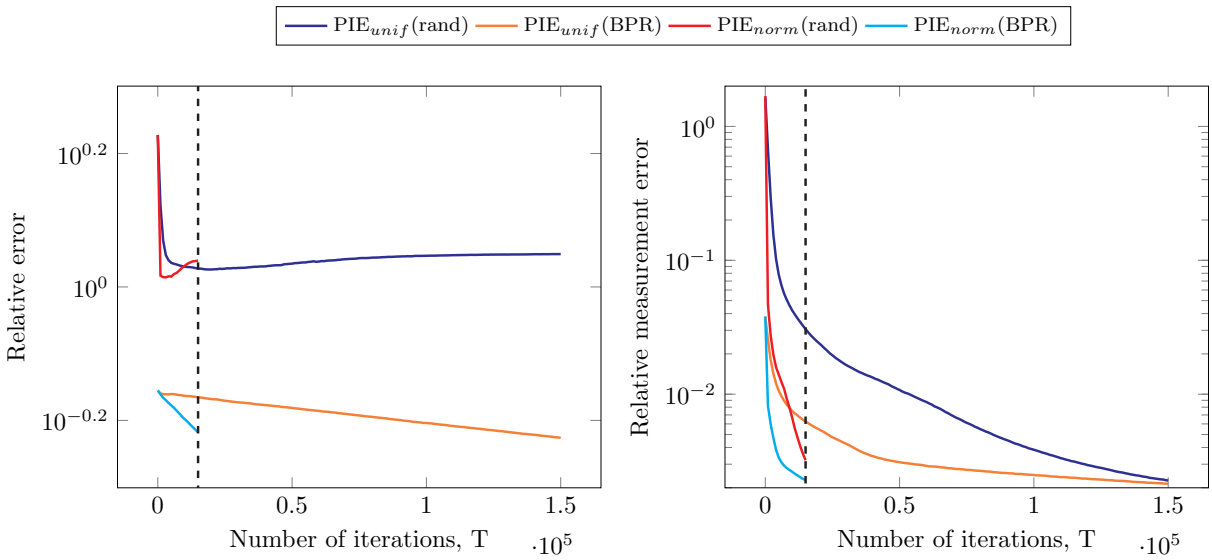
(b) Ptychographic Iterative Engine.

Figure 6.10: Performance of different iterative methods in the presence of noise.

iteration for the object shown in Figure 6.1a. We observe that with each iteration of AF and ER the relative measurement error improves, which is in line with Theorems 3.5.5 and 3.5.8. Also the same holds for PIE$_{unif}$, which is only guaranteed on average by Theorem 3.5.13. For this particular object, the errors for ER and AF without AG are almost the same. Furthermore, the measurement error for the random initialization stagnates after 1000 iterations and the relative error suggests that the algorithms converge to wrong fixed points. Similarly, a contrast between the initializations is observed for PIE algorithms. This can also be seen by looking at the phases of the reconstructed object with the random initialization in Table 6.5. We also note that reconstructions via PIE still contain phase errors from the initialization with BPR, which suggests that more iterations could lead to a better performance of the algorithms.



(a) Amplitude Flow and Error Reduction.



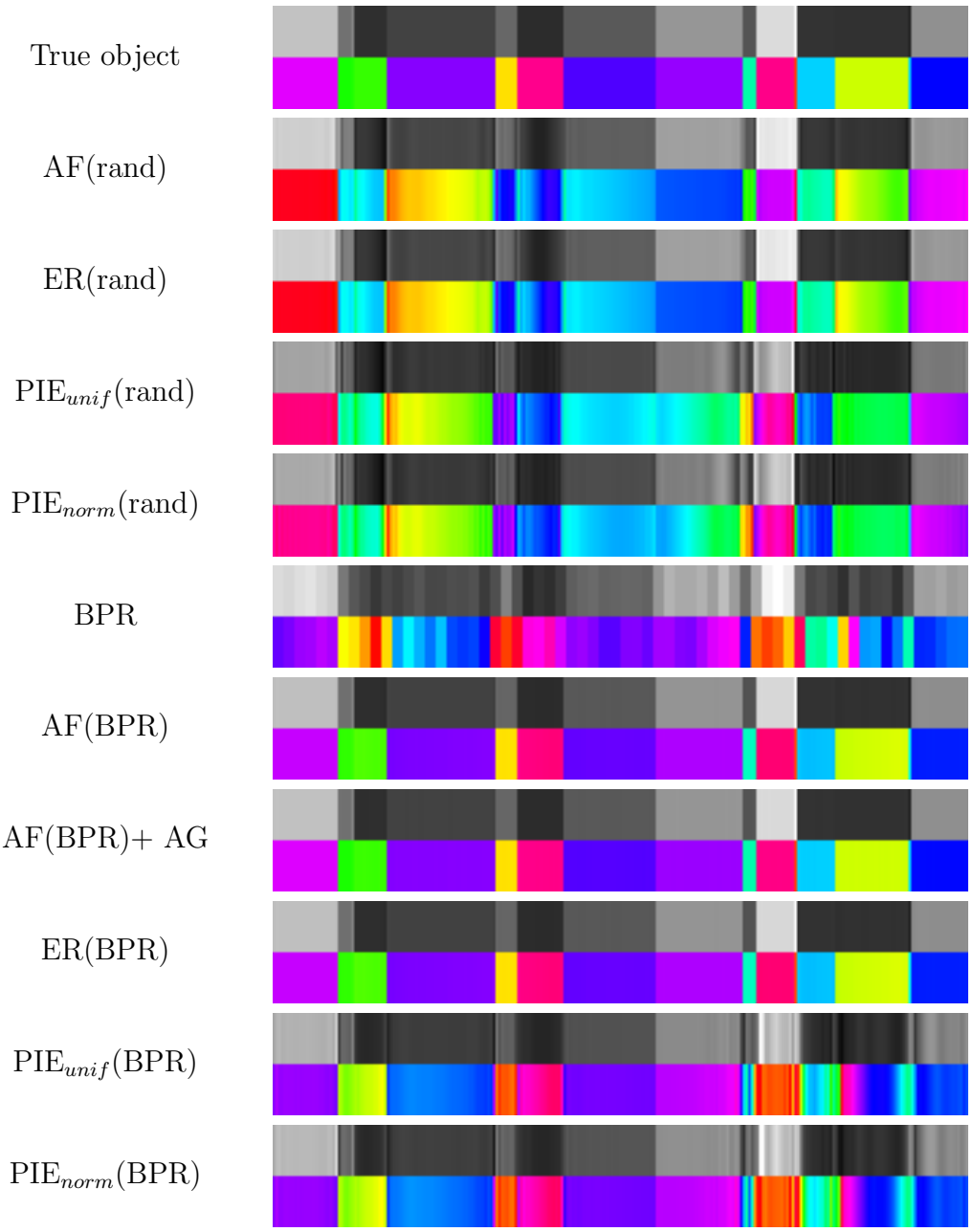
(b) Ptychographic Iterative Engine. Black dashed line shows $T = 15000$, the stopping point of PIE$_{norm}$.

Figure 6.11: Evolution of the errors with respect to the number of performed iterations for different iterative methods. $K = 10^6$.

6.1.4 Subsampling of frequencies

In this section, the subsampling of frequencies is considered, which allows to reduce the memory requirements and the computational complexity of the algorithms. For our first experiment regarding the subsampling, we consider six choices of the parameter m, equally spaced from the minimal theoretical choice $2\delta - 1$ (Theorem 3.6.3) to the maximal possible $m = d$. For each choice of m we generate measurements with $\mathcal{R} = [d]$ and perform the reconstruction with BPR. The resulting reconstruction errors are shown in Table 6.6. Note that the number of photons K and corresponding SNR remains the same for all m.

Table 6.5: Example of reconstructions with different iterative methods. $K = 10^6$.

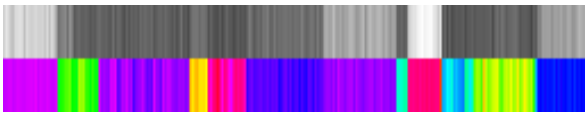


In the second experiment, we consider $\mathcal{R} = \mathcal{R}_s$ with $s = 4$, as in Section 6.1.3 and perform thirty reconstructions for each value of m and K via AF(BPR)+AG, discussed in the previous section. The averaged errors of these trials are summarized in Table 6.7.

The main outcome of both experiments is that by decreasing m, i.e., the quality of diffraction patterns, the reconstruction error increases, however, simultaneously the runtime decreases. As it is seen in Figure 6.12, despite the decrease of quality due to subsampling, the main features of the object are visible. The trade-off between the runtime and the quality of reconstruction provides an opportunity to obtain a quick noisy glance at the object from the subsampled measurements and then use it as an initialization for an

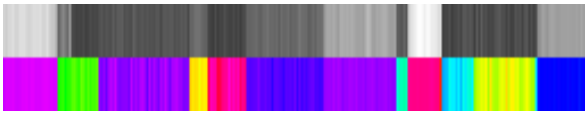
Table 6.6: The impact of the number of subsampled frequencies on the reconstruction error for BPR. The last row shows the average runtime.

K	$m = 63$	$m = 102$	$m = 140$	$m = 179$	$m = 217$	$m = 256$
10^2	1.228	1.227	1.196	1.194	1.184	1.169
10^3	1.121	1.129	1.134	1.144	1.077	1.141
10^4	0.821	0.750	0.736	0.779	0.821	0.646
10^5	0.372	0.297	0.287	0.296	0.239	0.262
10^6	0.216	0.181	0.161	0.137	0.134	0.126
10^7	0.124	0.117	0.085	0.084	0.066	0.086
10^8	0.071	0.062	0.046	0.049	0.040	0.046
10^9	0.044	0.031	0.025	0.024	0.020	0.020
avg. times, s	0.346	0.355	0.354	0.373	0.382	0.407



(a) With.

(c) With.



(b) Without.

(d) Without.

Figure 6.12: Reconstruction with BPR for $\mathcal{R} = [d]$ (left) and with AF(BPR)+AG for $\mathcal{R} = \mathcal{R}_s$ (right) with and without subsampling of frequencies in the presence of noise ($K = 10^6$).

iterative method with unsubsampled diffraction patterns. This could be faster than running an iterative method initialized with BPR on the unsubsampled diffraction patterns. When dimensions are large, this becomes even more crucial.

Table 6.7: The impact of the number of subsampled frequencies on the reconstruction error for AF(BPR)+AG.

K	$m = 63$	$m = 102$	$m = 140$	$m = 179$	$m = 217$	$m = 256$
10^2	0.865	0.874	0.781	0.696	0.667	0.710
10^3	0.812	0.651	0.558	0.533	0.490	0.429
10^4	0.698	0.443	0.468	0.502	0.358	0.451
10^5	0.385	0.433	0.467	0.383	0.254	0.363
10^6	0.579	0.523	0.527	0.295	0.340	0.246
10^7	0.592	0.445	0.397	0.339	0.282	0.376
10^8	0.471	0.516	0.446	0.364	0.340	0.294
10^9	0.511	0.517	0.331	0.295	0.300	0.323
times	3.238	6.170	8.066	12.382	12.909	13.439

6.2 Blind Ptychography

In this section we consider blind ptychographic reconstruction. We keep the same experimental setup as in the previous sections, so that $m = d$ and $\mathcal{R} = \mathcal{R}_s$ with $s = 4$.
The first algorithm is alternating Amplitude Flow (Algorithm 10) with smoothing parameter $\varepsilon = 10^{-12}$, Tikhonov regularization $\alpha_T = 10^{-2}$ and $\beta_T = 10^{-4}$. It performs $T = 250$ iterations with $T_z = T_{\hat{v}} = 10$ object and window subiterations, which in total gives 5000 gradient steps. The learning rates for object and window subiterations are selected via AG condition (Algorithm 1) with $\tau = 0.5$ and $N = 2$.
The second algorithm is alternating Amplitude Flow with reweighting (Algorithm 11) with the same parameters, except that $\alpha_T = \beta_T = 0$.
The third and last method is extended Ptychographic Iterative Engine in the form of the stochastic gradient descent (Theorem 4.3.1) with learning rates $\mu_t = \nu_t = 10^{-2.5}$ and $T = 150000$ iterations.
We denote these algorithms as AAF, AAF+RW and ePIE, respectively. Furthermore we consider the following initialization for the object and the window. The window initialization is a rough approximation of the shape of the window (see Figure 6.13b) and is given by

$$v_j^1 = \begin{cases} 0.05, & 0 \le j < 4 \text{ or } 28 \le j < 32, \\ 0.3, & 4 \le j < 8 \text{ or } 24 \le j < 28, \\ 0.6, & 8 \le j < 12 \text{ or } 20 \le j < 24, \\ 1, & 12 \le j < 20, \\ 0, & 20 \le j < 256. \end{cases} \tag{6.4}$$

The object is either randomly initialized or it is the result of the BPR reconstruction, which uses v^1 as window.
Figure 6.14 shows the resulting reconstruction errors for the different algorithms. We observe that AAF and AAF+RW provide almost the same results, while ePIE performs significantly worse. There is little difference between random initialization of the object and initialization via BPR. Moreover, the relative error for the object reconstruction does not drop below 0.5 and the relative error for the window reconstruction reaches the minimum of 0.1.
This suggests that the algorithms often converge to fixed points which are not the global minimizers. As the gradient methods are sensitive to the initialization, the reconstruction quality may be improved by a better initialization. In order to prove this assumption, we repeat the experiment with two different windows: v^1 as before, for which the relative error is 0.688 and a better approximation v^2 from the space $\mathbb{C}_4^d$ defined in (3.116). That is, we construct v^2 by setting up its isomorphic $\tilde{v}^2 \in \mathbb{C}^{d/4}$, which is defined via (6.1) with

(a) True window. (b) Window initialization v^1. (c) Window initialization v^2.

Figure 6.13: Window initializations for blind ptychographic reconstruction.

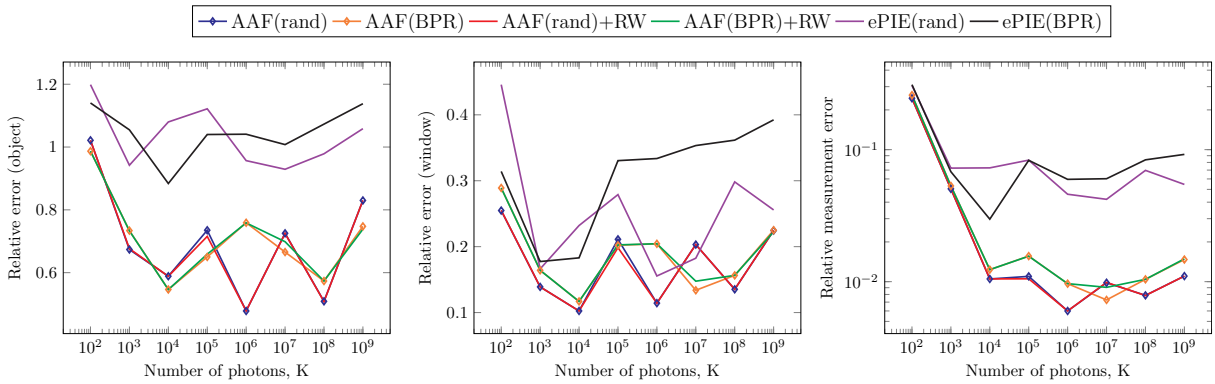


Figure 6.14: Performance of different iterative methods for blind ptychographic reconstruction in the presence of noise.

$\delta = 8$, $\mu = -0.5 + 0.5\delta$ and $\sigma^2 = 0.05$. The resulting window has a relative error of 0.161 and is shown in Figure 6.13c. For both windows, we use the BPR initialization as a starting point for AAF+RW and ePIE. We exclude AAF, since its performance is very similar to AAF+RW. The resulting reconstruction errors are shown in Figure 6.15 and the initialization with v^2 significantly improves all errors. The improvement of the reconstruction quality can also be seen in Table 6.8. Note that for this generated object and this noise vector, AAF with random initialization and v^1 performs as well as AFF with BPR initialization and v^2. Furthermore, in Table 6.8 we observe that minor shift ambiguities as in Example 4.1.3 occur despite the fact that the condition $\delta = d$ is not satisfied.

Finally, according to Table 6.9, the runtime of AAF is slightly longer than of ePIE, while its non-blind analogue AF+AG was much faster than PIE (Table 6.4). This is a consequence of an exclusion of the gradient norm stopping criterion.

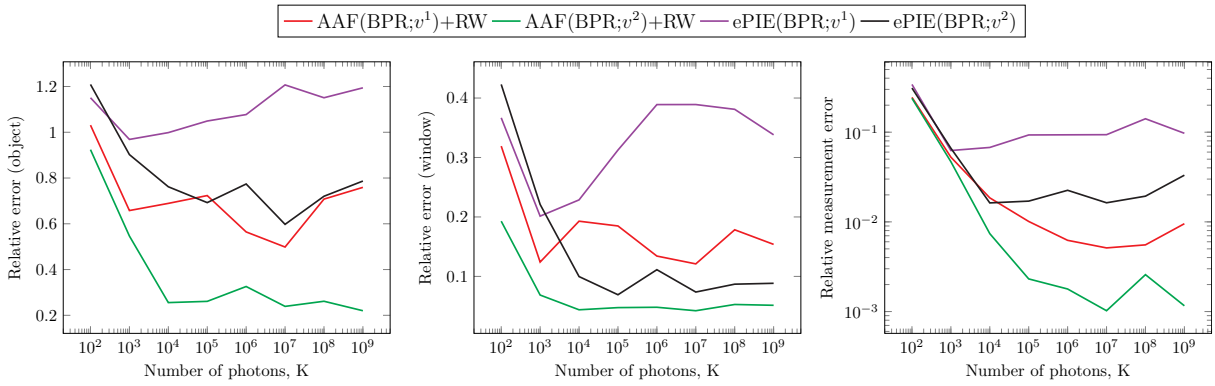
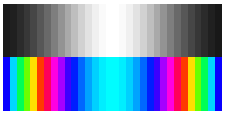
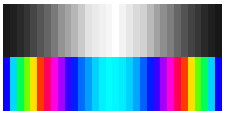
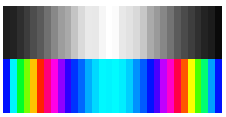
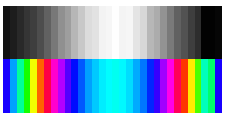
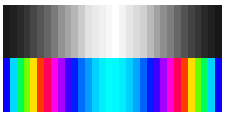


Figure 6.15: Impact of initialization on reconstruction quality via iterative methods for blind ptychography in the presence of noise.

Table 6.8: Example of blind reconstruction with different iterative methods. $K = 10^6$.

Algorithm	Object	Window
True		
AAF(rand, v^1)		
ePIE(rand, v^1)		
BPR, v^1		
AAF(BPR, v^1)		
ePIE(BPR, v^1)		
BPR, v^2		
AAF(BPR, v^2)		
ePIE(BPR, v^2)		

6.3 Polychromatic ptychography

In this section we perform numerical experiments to explore the performance of gradient descent methods for polychromatic ptychography.

6.3.1 Experimental setup

All our experiments will be performed on two-dimensional synthetic data within the following setup. We will consider polychromatic light with $L = 3$ frequencies $\nu = (1, 4/5, 2/3)$. For an object x an image of size $d = 100 \times 100$ is used, where the real and the imaginary parts are scaled images of the Shepp-Logan phantom and the cameraman, respectively. We slightly alternate the real and imaginary parts for different $\ell \in [L]$

Table 6.9: The average runtime of different algorithms and initializations for blind ptychography.

K	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
AAF(rand, v^1)	48.553	46.891	47.147	47.025	46.219	46.676	47.332	46.734
AAF(rand, v^1)+RW	48.306	46.851	46.875	46.474	45.732	46.583	47.116	46.295
AAF(BPR, v^1)	48.733	46.906	47.027	46.806	46.119	46.696	47.308	46.500
AAF(BPR, v^1)+RW	47.892	46.818	46.830	46.503	45.823	46.673	47.071	46.216
AF(BPR, v^2)+RW	51.583	47.956	44.912	45.635	46.484	45.826	45.813	45.496
ePIE(rand, v^1)	39.834	38.113	38.043	38.240	38.159	38.275	38.153	38.076
ePIE(BPR, v^1)	40.083	37.937	38.072	38.161	38.005	38.007	38.163	37.963
ePIE(BPR, v^2)	43.553	41.659	38.948	39.508	39.588	39.640	39.530	39.322

to imitate the dependence of the object on the frequencies as described in Figure 6.16. The window w is assumed to be locally supported with $\mathrm{supp}(w) = [\delta]^2$, $\delta = 40$. Within the support, its values are sampled from a Gaussian function,

$$(\hat{w}_\ell)_k = e^{-\|k-\mu\|_2^2/2\sigma^2}, \quad \ell \in [L], k \in [\delta]^2$$

where $\mu = ((\delta-1)/2, (\delta-1)/2)$ and $\sigma^2 = \delta^2/20$ and then scaled so that $w_\ell = \sqrt{\eta_\ell}\hat{w}/\|\hat{w}\|_2$ with $\eta = (0.2, 0.5, 0.3)$. For visualization, we refer the reader to Figure 6.17.
The set of shifts is selected by moving the center of the window along the Fermat spiral as discussed in [197]. That is, in the polar coordinate system (ρ, ϕ) the center of the window satisfies

$$\rho_k = c_{sp}\sqrt{k}, \ \phi_k = k\phi_0, \ 0 \le k \le \lceil 0.5((d-\delta)/c_{sp})^2 \rceil \tag{6.5}$$

where the $c_{sp} = 4.9$ is the scaling factor of the radius and the initial angle ϕ_0 is given by $\phi_0 = 2\pi(\frac{2}{1+\sqrt{5}})^2 \approx 137.508°$. Then, the pairs (ρ_k, ϕ_k) are transformed into a Cartesian coordinate system as $r = (r_k^1, r_k^2)$ and the set $\mathcal{R}$ only contains those points $r = (r_k^1 + d/2, r_k^2 + d/2)$, for which all non-zero entries of w are contained inside the object domain $[d]^2$ as depicted in Figure 6.17.
Figure 5.1 shows the simulated measurements $I_{r,j}$ given by (5.9). Furthermore, the measurements are corrupted by Poisson noise, so that

$$Y_{r,j} = \frac{d}{K}\mathrm{Pois}\left(\frac{KI_{r,j}}{d}\right),$$

$x_\ell = 0.2\ell +$  $+(1-0.3\ell)i$  $=$  $, \ell = 0$ $, \ell = 1$ $, \ell = 2$

Figure 6.16: The synthetic object in polychromatic light, $d = 100 \times 100$, $L = 3$.

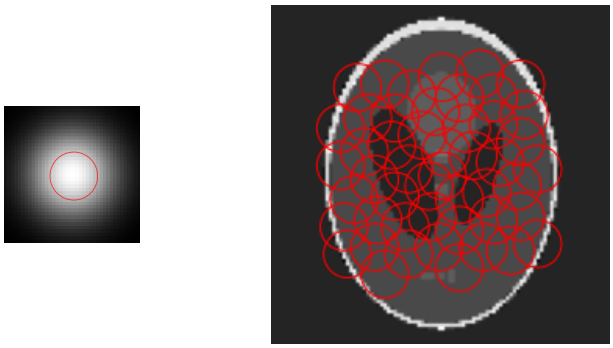


Figure 6.17: The localized window of size 40×40 and its shifts. Each red circle indicates position of the window on the Fermat spiral (6.5).

where $K = 10^6$ represents the number of photons used for the experiment. Since we fix a random seed for reproducible results, the SNR for all experiments is 38.69.
In order to measure the performance we will consider loss functions $\mathcal{L}_{2,\varepsilon}, \mathcal{J}, \mathcal{K}$ defined in equations (3.14), (5.11) and (5.21), respectively, with $\varepsilon = 10^{-8}$. In addition, with true object x known for synthetic data, the total and componentwise relative object errors

$$\mathrm{TRE} = \left[\sum_{\ell\in[L]} \mathrm{dist}^2(x_\ell, z_\ell)\right]^{1/2} / \|x\|_2 , \quad \text{and} \quad \mathrm{RE}_\ell = \mathrm{dist}(x_\ell, z_\ell)/\|z_\ell\|_2$$

can be evaluated and will be used for comparisons.

6.3.2 Amplitude Flow

We start with the non-blind polychromatic ptychography. In order to reconstruct the object x, the Amplitude Flow algorithm discussed in Section 5.4 is employed. That is, we perform the gradient descent minimization of the loss function $\mathcal{J}$ with parameters $\alpha_T = 10^{-2}$ and $\alpha_S = 10^{-1}$. The weights for the smoothness penalty $\mathcal{S}$ are set to impose the Lipschitz continuity in wavelength ν^{-1}, that is $\kappa_\ell = |\nu_{\ell+1}^{-1} - \nu_\ell^{-1}|^{-2}$.

Figure 6.18: Reconstruction of the object with a known window. Each row corresponds to a single frequency $\ell = 0, 1, 2$. The two consecutive columns are the real and imaginary parts of the object. In the figure, we show the true object x and iterates z^t for $t = 0, 100, 500, 2000$.

For an initial guess of the object z^0 the flat object is used, i.e., $(z^0)_k = 1$ for all $k \in [d]^2$. The learning rate is set to be a constant $\mu_t = \mu_c = L_w^{-1}$, where L_w is defined in Lemma 5.4.1 (with appropriate parameters for the two-dimensional case). The outcome of 2000 iterations of gradient descent is presented in Figure 6.18. We observe that already after 100 iterations, a blurry object is visible, after 500 iterations the object is more prominent and after 2000 iterations the edges of the Shepp-Logan phantom are smoothed. The reconstruction is more precise for the entries of the object, which belong to a larger number of regions. For those entries with lower number of overlaps, artifacts start to occur due to the small amount of available information.
Note that in Figure 6.18, the objects z_ℓ, $\ell \in [L]$ are similar, but not the same, which suggest that the smoothness penalty parameter $\alpha_S = 10^{-1}$ was chosen small enough to prevent the object being a constant function of frequencies, but large enough to impose continuity in ν^{-1}. In order to highlight the importance of the smoothness penalty, we repeat the reconstruction with $\alpha_S = 0$ and plot the resulting reconstructed objects in Figure 6.19. For $\alpha_S = 0$ the reconstruction is worse and more blurry.
Turning to numerical comparison, in both cases gradient descent decreases the loss $\mathcal{J}$ on each step (Figure 6.20a) as guaranteed by Corollary 5.4.2. Note that for $\alpha_S = 0$ the loss function $\mathcal{L}_{2,\varepsilon}$ is better optimized, which points towards the overfitting phenomena and its prevention by the inclusion of the smoothness penalty. This hypothesis is further supported by the relative errors in Figure 6.20b. In contrast to the loss function, the relative errors for the non-penalized solution are generally higher than for penalized.

6.3.3 Alternating Amplitude Flow

In the next experiment, we assume that the window w is unknown and the alternating Amplitude Flow (Algorithm 13) is considered for the blind polychromatic reconstruction.

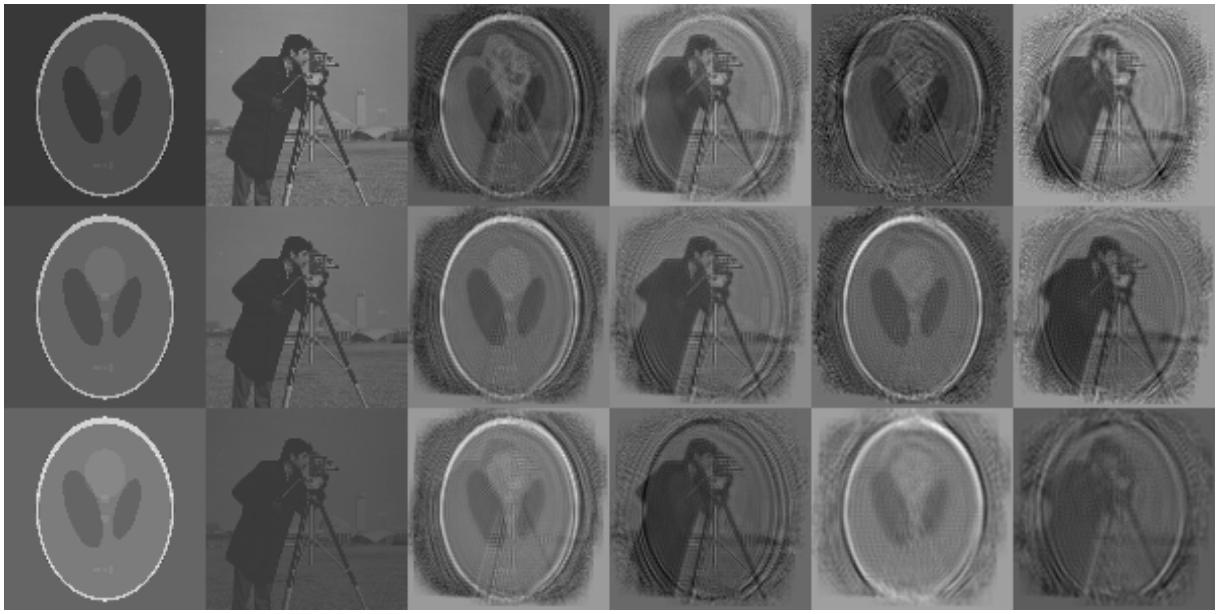
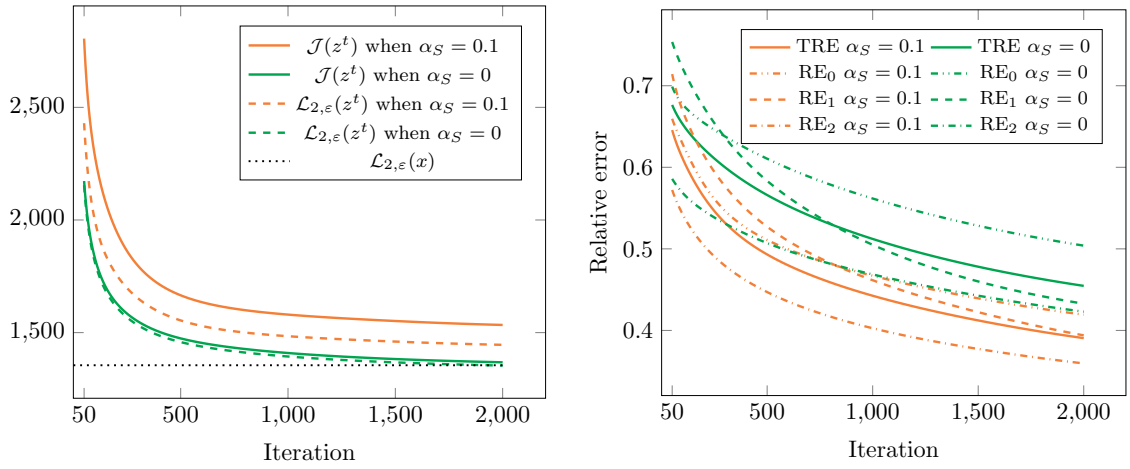


Figure 6.19: Comparison of the true object and the reconstructions with parameter $\alpha_S = 0.1$ and $\alpha_S = 0$.



(a) Objective function.

(b) Relative object error.

Figure 6.20: Comparison of reconstructions with parameter $\alpha_S = 0.1$ and $\alpha_S = 0$. First 50 iterations are excluded for a better visualization.

The number of iterations is set to $T = 200$ with object and window subiterations $T_z = T_v = 10$. This corresponds to a total of 4000 gradient steps, 2000 for each the object and the window. The object regularization parameters α_T, α_S and parameter ε are chosen as for the non-blind experiment above and the window regularization parameters are set to $\beta_T = 0.1, \beta_S = 10$.
For the object initialization, the flat starting point z^0 is used and the initial guess for the window is given by $v_\ell^0 = \sqrt{\eta_\ell}\hat{v}_\ell^0 / \left\|\hat{v}_\ell^0\right\|_2$ with

$$(\hat{v}_\ell^0)_k = \begin{cases} 2.3, & \|k-\mu\|_2 \leq \sqrt{0.3}\delta/2, \\ 1.3, & \|k-\mu\|_2 \leq \sqrt{0.6}\delta/2, \\ 0.3, & \|k-\mu\|_2 > \sqrt{0.6}\delta/2, k \in [\delta]^2, \\ 0, & \text{otherwise.} \end{cases}$$

Figure 6.21: Reconstruction of the object for blind polychromatic ptychography. Each row corresponds to a single frequency $\ell = 0, 1, 2$. The two consecutive columns are the real and imaginary parts of the object. In the figure, we show the true object x and iterates z^t of Algorithm 13 for $t = 0, 10, 50, 200$.

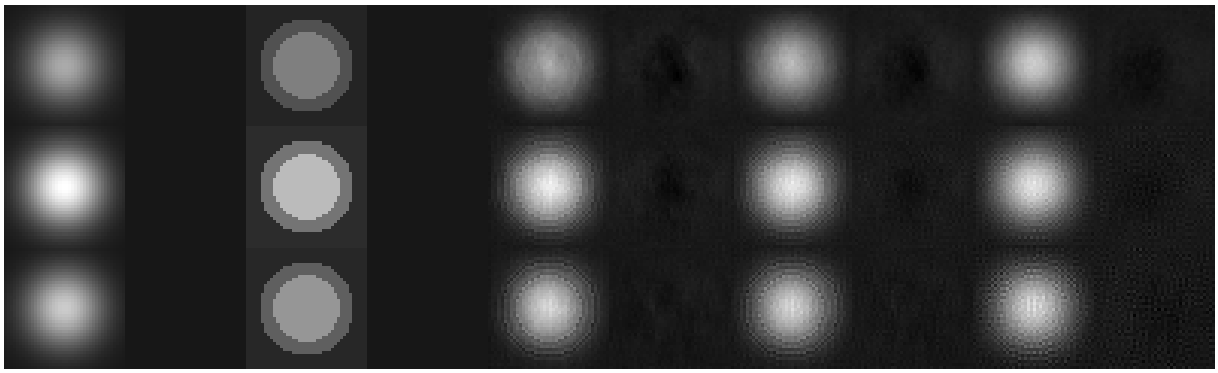
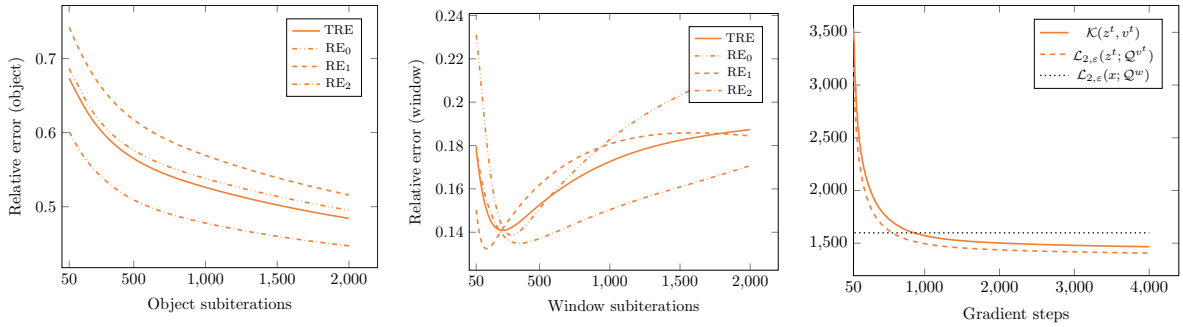


Figure 6.22: Reconstruction of the window for blind polychromatic ptychography. Each row corresponds to a single frequency $\ell = 1, 2, 3$. The two consecutive columns are the real and imaginary parts of the window. In the figure, we show the true window w and iterates v^t of Algorithm 13 for $t = 0, 10, 50, 200$.

The motivation behind this construction is to roughly imitate the shape of the true window w, which would be sufficient to ensure a fast convergence to the true window.
Let us explore the performance of Algorithm 13. The reconstruction of the object gradually improves over the number of performed object subiterations, which can be observed both in actual pictures provided by Figure 6.21 and in terms of errors in Figure 6.23a. On the other hand, according to Figure 6.22 the reconstruction of the window visually quickly stagnates. According to Figure 6.23b, the relative errors for the window reconstruction only improve in the beginning and grow back to their initial values. From Figure 6.23c, the stagnation in the reconstruction of the window may result from the overfitting, as the values of $\mathcal{L}_{2,\varepsilon}(z^t; \mathcal{Q}^{v^t})$ drop below $\mathcal{L}_{2,\varepsilon}(x; \mathcal{Q}^w)$.



(a) Relative object errors. (b) Relative window errors. (c) Loss functions.

Figure 6.23: Errors and loss functions during the blind polychromatic ptychographic reconstruction.

Chapter 7

Outlook and future research

In the conclusion, we discuss the main outcomes of this thesis, open problems and potential directions of future research.

By investigating Amplitude Flow, Error Reduction and Ptychographic Iterative Engine, we observed that these algorithms can be seen as generalized gradient methods applied to a non-convex and non-smooth loss function $\mathcal{L}_2$. While we established the sublinear convergence of the algorithms to a fixed point, whether it can be strengthened to a linear convergence rate or to a guaranteed convergence to a global minimum is unclear, but numerical examples suggest otherwise. This is a common issue of gradient methods applied to non-convex functions, however some studies for phase retrieval algorithms [86, 146] guarantee linear convergence to a global minimum under additional assumptions. The applicability of these assumptions to the ptychographic reconstruction could be an interesting direction of future research.

The convergence properties are even less understood for blind ptychographic reconstruction. We were able to establish the convergence of alternating Amplitude Flow to a fixed point with sublinear rate. However, the convergence analysis of extended Ptychographic Iterative Engine as stochastic gradient descent requires a significant advances in available methods and remains an open problem.

We note that these iterative algorithms for monochromatic ptychography and their analysis can be extended to other measurements scenarios, such as polychromatic ptychography in Chapter 5 or tomographic ptychography [198].

In Sections 6.1.3 and 6.2 we explored the use of Block Phase Retrieval algorithm as an initialization for iterative methods in comparison to a random guess. The obtained reconstruction errors identify the starting point as an important parameter for the fast convergence of the gradient methods. That is why the development of initialization algorithms is crucial for a good and fast reconstruction. Moreover, the usage of so-called pipelines of algorithms improves upon the reconstruction with a single method.

The numerical experiments for Block Phase Retrieval also highlighted several weak spots of the algorithm, the first of which is the inversion step. While the regularization via truncation, introduced in Section 3.6.2.2, decreases the relative error, the optimal choice of the truncation threshold depends on the unknown noise level. Furthermore, for a better noise robustness more advanced deconvolution methods may be used for the recovery of the diagonals in the inversion step.

The second weakness are the restrictions, which Block Phase Retrieval imposes on the

experimental setup. The extensions of Block Phase Retrieval to larger shifts between illumination regions was addressed in Section 3.6.5.1 for piecewise constant objects as well as in [141] for bandlimited objects. However, in Section 6.1.2.5, we applied the algorithm outside these classes, i.e., as a heuristic, with limited success in phase reconstruction. Thus, an extension to other classes of objects, e.g., with bounded total variation, may be beneficial.

Chapter 8

Bibliography

[1] R. Beinert, G. Plonka, Sparse Phase Retrieval of One-Dimensional Signals by Prony's Method, Frontiers in Applied Mathematics and Statistics 3 (2017). doi:10.3389/fams.2017.00005.

[2] W. Hoppe, Beugung im inhomogenen Primärstrahlwellenfeld. I. Prinzip einer Phasenmessung von Elektronenbeugungsinterferenzen, Acta Crystallographica Section A 25 (4) (1969) 495–501. doi:10.1107/S0567739469001045.

[3] R. Hegerl, W. Hoppe, Dynamische Theorie der Kristallstrukturanalyse durch Elektronenbeugung im inhomogenen Primärstrahlwellenfeld, Berichte der Bunsengesellschaft für physikalische Chemie 74 (11) (1970) 1148–1154. doi:10.1002/bbpc.19700741112.

[4] F. Pfeiffer, X-ray ptychography, Nature Photonics 12 (1) (2018) 9–17. doi:10.1038/s41566-017-0072-5.

[5] H. N. Chapman, Phase-retrieval X-ray microscopy by Wigner-distribution deconvolution, Ultramicroscopy 66 (3-4) (1996) 153–172. doi:10.1016/S0304-3991(96)00084-8.

[6] P. Thibault, M. Dierolf, A. Menzel, O. Bunk, C. David, F. Pfeiffer, High-resolution scanning x-ray diffraction microscopy, Science 321 (5887) (2008) 379–382. doi:10.1126/science.1158573.

[7] C. M. Kewish, P. Thibault, M. Dierolf, O. Bunk, A. Menzel, J. Vila-Comamala, K. Jefimovs, F. Pfeiffer, Ptychographic characterization of the wavefield in the focus of reflective hard X-ray optics, Ultramicroscopy 110 (4) (2010) 325–329. doi:10.1016/j.ultramic.2010.01.004.

[8] V. Piazza, B. Weinhausen, A. Diaz, C. Dammann, C. Maurer, M. Reynolds, M. Burghammer, S. Köster, Revealing the structure of stereociliary actin by X-ray nanoimaging, ACS nano 8 (12) (2014) 12228–12237. doi:10.1021/nn5041526.

[9] K. Høydalsvik, J. Bø Fløystad, T. Zhao, M. Esmaeili, A. Diaz, J. W. Andreasen, R. H. Mathiesen, M. Rønning, D. W. Breiby, In situ X-ray ptychography imaging

of high-temperature CO 2 acceptor particle agglomerates, Applied Physics Letters 104 (24) (2014) 241909. doi:10.1063/1.4884598.

[10] M. Esmaeili, J. B. Fløystad, A. Hipp, M. Willner, M. Bech, A. Diaz, A. Røyset, J. W. Andreasen, F. Pfeiffer, D. W. Breiby, Monitoring moisture distribution in textile materials using grating interferometry and ptychographic X-ray imaging, Textile Research Journal 85 (1) (2015) 80–90. doi:10.1177/0040517514538693.

[11] X. Zhu, A. P. Hitchcock, D. A. Bazylinski, P. Denes, J. Joseph, U. Lins, S. Marchesini, H.-W. Shiu, T. Tyliszczak, D. A. Shapiro, Measuring spectroscopy and magnetism of extracted and intracellular magnetosomes using soft X-ray ptychography, Proceedings of the National Academy of Sciences of the United States of America 113 (51) (2016) E8219–E8227. doi:10.1073/pnas.1610260114.

[12] S. Lazarev, I. Besedin, A. V. Zozulya, J.-M. Meijer, D. Dzhigaev, O. Y. Gorobtsov, R. P. Kurta, M. Rose, A. G. Shabalin, E. A. Sulyanova, I. Zaluzhnyy, A. P. Menushenkov, M. Sprung, A. V. Petukhov, I. A. Vartanyants, Ptychographic X-Ray Imaging of Colloidal Crystals, Small (Weinheim an der Bergstrasse, Germany) 14 (3) (2018). doi:10.1002/smll.201702575.

[13] M. Dierolf, A. Menzel, P. Thibault, P. Schneider, C. M. Kewish, R. Wepf, O. Bunk, F. Pfeiffer, Ptychographic X-ray computed tomography at the nanoscale, Nature 467 (7314) (2010) 436–439. doi:10.1038/nature09419.

[14] B. Chen, M. Guizar-Sicairos, G. Xiong, L. Shemilt, A. Diaz, J. Nutter, N. Burdet, S. Huo, J. Mancuso, A. Monteith, F. Vergeer, A. Burgess, I. Robinson, Three-dimensional structure analysis and percolation properties of a barrier marine coating, Scientific reports 3 (2013) 1177. doi:10.1038/srep01177.

[15] M. Esmaeili, J. B. Fløystad, A. Diaz, K. Høydalsvik, M. Guizar-Sicairos, J. W. Andreasen, D. W. Breiby, Ptychographic X-ray Tomography of Silk Fiber Hydration, Macromolecules 46 (2) (2013) 434–439. doi:10.1021/ma3021163.

[16] M. Holler, A. Diaz, M. Guizar-Sicairos, P. Karvinen, E. Färm, E. Härkönen, M. Ritala, A. Menzel, J. Raabe, O. Bunk, X-ray ptychographic computed tomography at 16 nm isotropic 3D resolution, Scientific reports 4 (2014) 3857. doi:10.1038/srep03857.

[17] J. C. da Silva, K. Mader, M. Holler, D. Haberthür, A. Diaz, M. Guizar-Sicairos, W.-C. Cheng, Y. Shu, J. Raabe, A. Menzel, J. A. van Bokhoven, Assessment of the 3 D Pore Structure and Individual Components of Preshaped Catalyst Bodies by X-Ray Imaging, ChemCatChem 7 (3) (2015) 413–416. doi:10.1002/cctc.201402925.

[18] K. Giewekemeyer, C. Hackenberg, A. Aquila, R. N. Wilke, M. R. Groves, R. Jordanova, V. S. Lamzin, G. Borchers, K. Saksl, A. V. Zozulya, M. Sprung, A. P. Mancuso, Tomography of a Cryo-immobilized Yeast Cell Using Ptychographic Coherent X-Ray Diffractive Imaging, Biophysical journal 109 (9) (2015) 1986–1995. doi:10.1016/j.bpj.2015.08.047.

[19] M. Holler, M. Guizar-Sicairos, E. H. R. Tsai, R. Dinapoli, E. Müller, O. Bunk, J. Raabe, G. Aeppli, High-resolution non-destructive three-dimensional imaging of integrated circuits, Nature 543 (7645) (2017) 402–406. doi:10.1038/nature21698.

[20] A. M. Maiden, M. J. Humphry, J. M. Rodenburg, Ptychographic transmission microscopy in three dimensions using a multi-slice approach, Journal of the Optical Society of America. A, Optics, image science, and vision 29 (8) (2012) 1606–1614. doi:10.1364/JOSAA.29.001606.

[21] P. Li, A. Maiden, Multi-slice ptychographic tomography, Scientific reports 8 (1) (2018) 2049. doi:10.1038/s41598-018-20530-x.

[22] M. Kahnt, L. Grote, D. Brückner, M. Seyrich, F. Wittwer, D. Koziej, C. G. Schroer, Multi-slice ptychography enables high-resolution measurements in extended chemical reactors, Scientific reports 11 (1) (2021) 1500. doi:10.1038/s41598-020-80926-6.

[23] Z. Chen, Y. Jiang, Y.-T. Shao, M. E. Holtz, M. Odstrčil, M. Guizar-Sicairos, I. Hanke, S. Ganschow, D. G. Schlom, D. A. Muller, Electron ptychography achieves atomic-resolution limits set by lattice vibrations, Science 372 (6544) (2021) 826–831. doi:10.1126/science.abg2533.

[24] D. J. Batey, D. Claus, J. M. Rodenburg, Information multiplexing in ptychography, Ultramicroscopy 138 (2014) 13–21. doi:10.1016/j.ultramic.2013.12.003.

[25] X. Huang, K. Lauer, J. N. Clark, W. Xu, E. Nazaretski, R. Harder, I. K. Robinson, Y. S. Chu, Fly-scan ptychography, Scientific reports 5 (2015) 9074. doi:10.1038/srep09074.

[26] D. Griffin, J. Lim, Signal estimation from modified short-time Fourier transform, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (2) (1984) 236–243. doi:10.1109/TASSP.1984.1164317.

[27] J. R. Fienup, Reconstruction of an object from the modulus of its Fourier transform, Optics letters 3 (1) (1978) 27–29. doi:10.1364/ol.3.000027.

[28] V. Elser, Phase retrieval by iterated projections, Journal of the Optical Society of America. A, Optics, image science, and vision 20 (1) (2003) 40–55. doi:10.1364/josaa.20.000040.

[29] H. H. Bauschke, P. L. Combettes, D. R. Luke, Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization, Journal of the Optical Society of America. A, Optics, image science, and vision 19 (7) (2002) 1334–1345. doi:10.1364/josaa.19.001334.

[30] D. R. Luke, Relaxed averaged alternating reflections for diffraction imaging, Inverse Problems 21 (1) (2005) 37–50. doi:10.1088/0266-5611/21/1/004.

[31] E. J. Candes, X. Li, M. Soltanolkotabi, Phase Retrieval via Wirtinger Flow: Theory and Algorithms, IEEE Transactions on Information Theory 61 (4) (2015) 1985–2007. doi:10.1109/TIT.2015.2399924.

[32] G. Wang, G. B. Giannakis, Y. C. Eldar, Solving Systems of Random Quadratic Equations via Truncated Amplitude Flow, IEEE Transactions on Information Theory 64 (2) (2018) 773–794. doi:10.1109/TIT.2017.2756858.

[33] A. M. Maiden, J. M. Rodenburg, An improved ptychographical phase retrieval algorithm for diffractive imaging, Ultramicroscopy 109 (10) (2009) 1256–1262. doi:10.1016/j.ultramic.2009.05.012.

[34] E. J. Candès, Y. C. Eldar, T. Strohmer, V. Voroninski, Phase Retrieval via Matrix Completion, SIAM Journal on Imaging Sciences 6 (1) (2013) 199–225. doi:10.1137/110848074.

[35] M. A. Iwen, A. Viswanathan, Y. Wang, Fast Phase Retrieval from Local Correlation Measurements, SIAM Journal on Imaging Sciences 9 (4) (2016) 1655–1688. doi:10.1137/15M1053761.

[36] R. Xu, M. Soltanolkotabi, J. P. Haldar, W. Unglaub, J. Zusman, A. F. J. Levi, R. M. Leahy, Accelerated Wirtinger Flow: A fast algorithm for ptychography. URL https://arxiv.org/pdf/1806.05546

[37] Gerchberg R.W., Saxton W.O., A practical algorithm for the determination of phase from image and diffraction plane pictures, Optik 35 (1972) 237.

[38] S. Marchesini, Y.-C. Tu, H.-T. Wu, Alternating projection, ptychographic imaging and phase synchronization, Applied and Computational Harmonic Analysis 41 (3) (2016) 815–851. doi:10.1016/j.acha.2015.06.005.

[39] J. M. Rodenburg, H. M. L. Faulkner, A phase retrieval algorithm for shifting illumination, Applied Physics Letters 85 (20) (2004) 4795–4797. doi:10.1063/1.1823034.

[40] M. A. Iwen, B. Preskitt, R. Saab, A. Viswanathan, Phase retrieval from local measurements: Improved robustness via eigenvector-based angular synchronization, Applied and Computational Harmonic Analysis 48 (1) (2020) 415–444. doi:10.1016/j.acha.2018.06.004.

[41] M. Born, E. Wolf, Principles of optics: Electromagnetic theory of propagation, interference and diffraction of light / by Max Born and Emil Wolf with contributions by A. B. Bhatia ...[et al.], 6th Edition, Cambridge University Press, Cambridge, 1997, 1980.

[42] B. E. A. Saleh, M. C. Teich, Fundamentals of photonics, 2nd Edition, Wiley Series in Pure and Applied Optics, Wiley, Chicester, 2013.

[43] J. C. Maxwell, T. F. Torrance, A dynamical theory of the electromagnetic field, The Torrance collection, Wipf & Stock, Eugene, Oregon, 1996.

[44] K. Gröchenig, Foundations of time-frequency analysis, softcover repr. of the hardcover 1. ed. Edition, Applied and numerical harmonic analysis, Springer Science + Business Media, New York, NY, 2001.

[45] P. Thibault, V. Elser, C. Jacobsen, D. Shapiro, D. Sayre, Reconstruction of a yeast cell from X-ray diffraction data, Acta crystallographica. Section A, Foundations of crystallography 62 (Pt 4) (2006) 248–261. doi:10.1107/S0108767306016515.

[46] P. Thibault, M. Dierolf, O. Bunk, A. Menzel, F. Pfeiffer, Probe retrieval in ptychographic coherent diffractive imaging, Ultramicroscopy 109 (4) (2009) 338–343. doi:10.1016/j.ultramic.2008.12.011.

[47] R. A. Horn, C. R. Johnson, Matrix analysis, 2nd Edition, Cambridge University Press, New York, op. 2013. doi:10.1017/CBO9780511810817.

[48] R. Bhatia, Matrix Analysis, softcover reprint of the original 1st ed. 1997 Edition, Vol. 169 of Graduate Texts in Mathematics, Springer-Verlag New York Inc, [s.l.], 2012. doi:10.1007/978-1-4612-0653-8.

[49] J.-P. Aubin, Applied Functional Analysis, 2nd Edition, Pure and applied mathematics, John Wiley & Sons, Inc, Hoboken, NJ, USA, 2000. doi:10.1002/9781118032725.
URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781118032725

[50] W. Wirtinger, Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen, Mathematische Annalen 97 (1) (1927) 357–375. doi:10.1007/BF01447872.

[51] R. Hunger, An Introduction to Complex Differentials and Complex Differentiability (2008).
URL https://mediatum.ub.tum.de/doc/631019/631019.pdf

[52] P. Bouboulis, Wirtinger's Calculus in general Hilbert Spaces.
URL https://arxiv.org/pdf/1005.5170

[53] A. Khaled, P. Richtárik, Better theory for SGD in the nonconvex world, Transactions on Machine Learning Research (2023).
URL https://openreview.net/forum?id=AU4qHN2VkS

[54] O. Sebbouh, R. M. Gower, A. Defazio, Almost sure convergence rates for Stochastic Gradient Descent and Stochastic Heavy Ball, in: Proceedings of Thirty Fourth Conference on Learning Theory, Vol. 134, pp. 3935–3971.
URL https://proceedings.mlr.press/v134/

[55] J. Liu, Y. Yuan, On almost sure convergence rates of stochastic gradient methods, in: P.-L. Loh, M. Raginsky (Eds.), Proceedings of Thirty Fifth Conference on Learning Theory, Vol. 178 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 2963–2983.
URL https://proceedings.mlr.press/v178/liu22d.html

[56] G. Thakur, Reconstruction of Bandlimited Functions from Unsigned Samples, Journal of Fourier Analysis and Applications 17 (4) (2011) 720–732. doi:10.1007/s00041-010-9144-3.

[57] B. A. Shenoy, S. Mulleti, C. S. Seelamantula, Exact Phase Retrieval in Principal Shift-Invariant Spaces, IEEE Transactions on Signal Processing 64 (2) (2016) 406–416. doi:10.1109/TSP.2015.2481871.

[58] V. Pohl, N. Li, H. Boche, Phase retrieval in spaces of analytic functions on the unit disk, in: G. Anbarjafari, I. C. o. S. T. a. Applications (Eds.), 2017 International Conference on Sampling Theory and Applications (SampTA), IEEE, Piscataway, NJ, 2017, pp. 336–340. doi:10.1109/SAMPTA.2017.8024411.

[59] R. Alaifari, I. Daubechies, P. Grohs, G. Thakur, Reconstructing Real-Valued Functions from Unsigned Coefficients with Respect to Wavelet and Other Frames, Journal of Fourier Analysis and Applications 23 (6) (2017) 1480–1494. doi:10.1007/s00041-016-9513-7.

[60] R. Alaifari, P. Grohs, Phase Retrieval In The General Setting Of Continuous Frames For Banach Spaces, SIAM Journal on Mathematical Analysis 49 (3) (2017) 1895–1911. doi:10.1137/16M1071481.

[61] Y. F. Li, D. G. Han, Phase Retrieval of Real-valued Functions in Sobolev Space, Acta Mathematica Sinica, English Series 34 (12) (2018) 1778–1794. doi:10.1007/s10114-018-7422-1.

[62] R. Alaifari, I. Daubechies, P. Grohs, R. Yin, Stable Phase Retrieval in Infinite Dimensions, Foundations of Computational Mathematics 19 (4) (2019) 869–900. doi:10.1007/s10208-018-9399-7.

[63] Y. Chen, C. Cheng, Q. Sun, H. Wang, Phase retrieval of real-valued signals in a shift-invariant space, Applied and Computational Harmonic Analysis 49 (1) (2020) 56–73. doi:10.1016/j.acha.2018.11.002.

[64] K. Gröchenig, Phase-Retrieval in Shift-Invariant Spaces with Gaussian Generator, Journal of Fourier Analysis and Applications 26 (3) (2020). doi:10.1007/s00041-020-09755-5.

[65] P. Jaming, K. Kellay, R. Perez, On the effect of zero-flipping on the stability of the phase retrieval problem in the Paley-Wiener class, Monatshefte für Mathematik 198 (4) (2022) 757–776. doi:10.1007/s00605-022-01716-y.

[66] S. Merhi, A. Viswanathan, M. Iwen, Recovery of compactly supported functions from spectrogram measurements via lifting, in: G. Anbarjafari, I. C. o. S. T. a. Applications (Eds.), 2017 International Conference on Sampling Theory and Applications (SampTA), IEEE, Piscataway, NJ, 2017, pp. 538–542. doi:10.1109/SAMPTA.2017.8024397.

[67] P. Grohs, L. Liehr, Injectivity of gabor phase retrieval from lattice measurements, Applied and Computational Harmonic Analysis 62 (2023) 173–193. doi:10.1016/j.acha.2022.09.001.
URL https://www.sciencedirect.com/science/article/pii/S106352032200077X

[68] P. Grohs, S. Koppensteiner, M. Rathmair, Phase Retrieval: Uniqueness and Stability, SIAM Review 62 (2) (2020) 301–350. doi:10.1137/19M1256865.

[69] M. Perlmutter, N. Sissouno, A. Viswantathan, M. Iwen, A Provably Accurate Algorithm for Recovering Compactly Supported Smooth Functions from Spectrogram Measurements, in: A. Marques, B. Hunyadi (Eds.), 28th European Signal Processing Conference (EUSIPCO 2020), IEEE, [Piscataway, NJ], 2020, pp. 970–974. doi:10.23919/Eusipco47968.2020.9287698.

[70] R. Alaifari, M. Wellershoff, Uniqueness of STFT phase retrieval for bandlimited functions, Applied and Computational Harmonic Analysis 50 (2021) 34–48. doi:10.1016/j.acha.2020.08.003.

[71] R. Li, B. Liu, Q. Zhang, Uniqueness of STFT phase retrieval in shift-invariant spaces, Applied Mathematics Letters 118 (2021) 107131. doi:10.1016/j.aml.2021.107131.

[72] P. Grohs, M. Rathmair, Stable Gabor Phase Retrieval and Spectral Clustering, Communications on Pure and Applied Mathematics 72 (5) (2019) 981–1043. doi:10.1002/cpa.21799.

[73] P. Grohs, M. Rathmair, Stable Gabor phase retrieval for multivariate functions, Journal of the European Mathematical Society 24 (5) (2022) 1593–1615. doi:10.4171/JEMS/1114.

[74] P. Jaming, K. Kellay, R. Perez, Phase Retrieval for Wide Band Signals, Journal of Fourier Analysis and Applications 26 (4) (2020). doi:10.1007/s00041-020-09767-1.

[75] T. Claasen, W. Mecklenbräuker, The Wigner Distribution - A Tool for Time-Frequency Signal Analysis, Part III: Relations with other Time-Frequency Signal Transformations, Philips Journal of Research 35 (1980) 372–389.
URL http://publik.tuwien.ac.at/files/PubDat_249647.pdf

[76] L. Cohen, Time-frequency distributions-a review, Proceedings of the IEEE 77 (7) (1989) 941–981. doi:10.1109/5.30749.

[77] T. Claasen, W. Mecklenbräuker, The Wigner Distribution - A Tool for Time-Frequency Signal Analysis, Part I: Continuous-Time Signals, Philips Journal of Research 35 (1980) 217–250.
URL https://publik.tuwien.ac.at/files/PubDat_249637.pdf

[78] T. Claasen, W. Mecklenbräuker, The Wigner Distribution - A Tool for Time-Frequency Signal Analysis, Part II: Discrete-Time Signals, Philips Journal of Research 35 (1980) 276–300.
URL https://publik.tuwien.ac.at/files/PubDat_249646.pdf

[79] J. M. Rodenburg, R. Bates, The theory of super-resolution electron microscopy via Wigner-distribution deconvolution, Philosophical Transactions of the Royal Society

of London. Series A: Physical and Engineering Sciences 339 (1655) (1992) 521–553. doi:10.1098/rsta.1992.0050.

[80] F. Filbir, L. Liehr, Phase Distortion by Linear Signal Transforms, Frontiers in Applied Mathematics and Statistics 6 (2020). doi:10.3389/fams.2020.556585.

[81] R. Balan, P. Casazza, D. Edidin, On signal reconstruction without phase, Applied and Computational Harmonic Analysis 20 (3) (2006) 345–356. doi:10.1016/j.acha.2005.07.001.

[82] A. S. Bandeira, J. Cahill, D. G. Mixon, A. A. Nelson, Saving phase: Injectivity and stability for phase retrieval, Applied and Computational Harmonic Analysis 37 (1) (2014) 106–125. doi:10.1016/j.acha.2013.10.002.

[83] A. Conca, D. Edidin, M. Hering, C. Vinzant, An algebraic characterization of injectivity in phase retrieval, Applied and Computational Harmonic Analysis 38 (2) (2015) 346–356. doi:10.1016/j.acha.2014.06.005.

[84] T. Bendory, C.-y. Cheng, D. Edidin, Near-Optimal Bounds for Signal Recovery from Blind Phaseless Periodic Short-Time Fourier Transform, Journal of Fourier Analysis and Applications 29 (1) (2022). doi:10.1007/s00041-022-09983-x.

[85] I. Bojarovska, A. Flinth, Phase Retrieval from Gabor Measurements, Journal of Fourier Analysis and Applications 22 (3) (2016) 542–567. doi:10.1007/s00041-015-9431-0.

[86] T. Bendory, Y. C. Eldar, N. Boumal, Non-Convex Phase Retrieval From STFT Measurements, IEEE Transactions on Information Theory 64 (1) (2018) 467–484. doi:10.1109/TIT.2017.2745623.

[87] R. Alaifari, M. Wellershoff, Stability Estimates for Phase Retrieval from Discrete Gabor Measurements, Journal of Fourier Analysis and Applications 27 (2) (2021). doi:10.1007/s00041-020-09802-1.

[88] J. Cahill, P. G. Casazza, I. Daubechies, Phase retrieval in infinite-dimensional Hilbert spaces, Transactions of the American Mathematical Society, Series B 3 (3) (2016) 63–76. doi:10.1090/btran/12.

[89] M. A. Iwen, S. Merhi, M. Perlmutter, Lower Lipschitz bounds for phase retrieval from locally supported measurements, Applied and Computational Harmonic Analysis 47 (2) (2019) 526–538. doi:10.1016/j.acha.2019.01.004.

[90] Y. C. Eldar, S. Mendelson, Phase retrieval: Stability and recovery guarantees, Applied and Computational Harmonic Analysis 36 (3) (2014) 473–494. doi:10.1016/j.acha.2013.08.003.

[91] F. Krahmer, Y.-K. Liu, Phase Retrieval Without Small-Ball Probability Assumptions, IEEE Transactions on Information Theory 64 (1) (2018) 485–500. doi:10.1109/TIT.2017.2757520.

[92] M. Kabanava, R. Kueng, H. Rauhut, U. Terstiege, Stable low-rank matrix recovery via null space properties, Information and Inference: A Journal of the IMA 5 (4) (2016) 405–441. doi:10.1093/imaiai/iaw014.

[93] R. Kueng, H. Zhu, D. Gross, Low rank matrix recovery from Clifford orbits. URL http://arxiv.org/pdf/1610.08070v1

[94] P. Salanevich, Stability of Phase Retrieval Problem, in: 2019 13th International Conference on Sampling Theory and Applications (SampTA), IEEE, Piscataway, NJ, 2019, pp. 1–4. doi:10.1109/SampTA45681.2019.9031013.

[95] Y. C. Eldar, P. Sidorenko, D. G. Mixon, S. Barel, O. Cohen, Sparse Phase Retrieval from Short-Time Fourier Measurements, IEEE Signal Processing Letters 22 (5) (2015) 638–642. doi:10.1109/LSP.2014.2364225.

[96] L. Li, C. Cheng, D. Han, Q. Sun, G. Shi, Phase Retrieval From Multiple-Window Short-Time Fourier Measurements, IEEE Signal Processing Letters 24 (4) (2017) 372–376. doi:10.1109/LSP.2017.2663668.

[97] S. Nawab, T. Quatieri, J. Lim, Signal reconstruction from short-time Fourier transform magnitude, IEEE Transactions on Acoustics, Speech, and Signal Processing 31 (4) (1983) 986–998. doi:10.1109/TASSP.1983.1164162.

[98] R. Balan, On signal reconstruction from its spectrogram, in: 2010 44th Annual Conference on Information Sciences and Systems, IEEE, Piscataway, N.J., 2010, pp. 1–4. doi:10.1109/CISS.2010.5464828.

[99] K. Jaganathan, Y. C. Eldar, B. Hassibi, STFT Phase Retrieval: Uniqueness Guarantees and Recovery Algorithms, IEEE Journal of Selected Topics in Signal Processing 10 (4) (2016) 770–781. doi:10.1109/JSTSP.2016.2549507.

[100] T. Bendory, R. Beinert, Y. C. Eldar, Fourier Phase Retrieval: Uniqueness and Algorithms, in: H. Boche, G. Caire, R. Calderbank, M. März, G. Kutyniok, R. Mathar (Eds.), Compressed sensing and its applications, Applied and numerical harmonic analysis, Birkhäuser, Cham, Switzerland, 2017, pp. 55–91. doi:10.1007/978-3-319-69802-1_2.

[101] R. Alaifari, P. Grohs, Gabor phase retrieval is severely ill-posed, Applied and Computational Harmonic Analysis 50 (2021) 401–419. doi:10.1016/j.acha.2019.09.003.

[102] P. Grohs, M. Rathmair, L^2-stability analysis for Gabor phase retrieval. URL https://arxiv.org/pdf/2108.06154

[103] J. R. Fienup, Phase retrieval algorithms: a comparison, Applied optics 21 (15) (1982) 2758–2769. doi:10.1364/AO.21.002758.

[104] A. Levi, H. Stark, Image restoration by the method of generalized projections with application to restoration from magnitude, in: ICASSP '84. IEEE International Conference on Acoustics, Speech, and Signal Processing, Institute of Electrical and Electronics Engineers, San Diego, CA, USA, 1984, pp. 88–91. doi:10.1109/ICASSP.1984.1172785.

[105] A. Tsipenyuk, Variational Approach to Fourier Phase Retrieval, Doctoral thesis, Technical University of Munich (2023).
URL https://mediatum.ub.tum.de/1638097

[106] J. Douglas, H. H. Rachford, On the numerical solution of heat conduction problems in two and three space variables, Transactions of the American Mathematical Society 82 (2) (1956) 421–439. doi:10.1090/S0002-9947-1956-0084194-4.

[107] D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Computers & Mathematics with Applications 2 (1) (1976) 17–40. doi:10.1016/0898-1221(76)90003-1.

[108] P. L. Lions, B. Mercier, Splitting Algorithms for the Sum of Two Nonlinear Operators, SIAM Journal on Numerical Analysis 16 (6) (1979) 964–979. doi:10.1137/0716071.

[109] A. Fannjiang, Z. Zhang, Fixed Point Analysis of Douglas–Rachford Splitting for Ptychography and Phase Retrieval, SIAM Journal on Imaging Sciences 13 (2) (2020) 609–650. doi:10.1137/19M128781X.

[110] Y. Chen, E. Candes, Solving random quadratic systems of equations is nearly as easy as solving linear systems, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 28, Curran Associates, Inc., 2015.
URL https://proceedings.neurips.cc/paper_files/paper/2015/file/7380ad8a673226ae47fce7bff88e9c33-Paper.pdf

[111] G. Wang, G. B. Giannakis, Y. Saad, J. Chen, Phase Retrieval via Reweighted Amplitude Flow, IEEE Transactions on Signal Processing 66 (11) (2018) 2818–2833. doi:10.1109/TSP.2018.2818077.

[112] H. Zhang, Y. Liang, Reshaped wirtinger flow for solving quadratic system of equations, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 29, Curran Associates, Inc., 2016.
URL https://proceedings.neurips.cc/paper_files/paper/2016/file/83adc9225e4deb67d7ce42d58fe5157c-Paper.pdf

[113] G. Wang, G. B. Giannakis, J. Chen, Solving large-scale systems of random quadratic equations via stochastic truncated amplitude flow, in: I. Staff (Ed.), 2017 25th European Signal Processing Conference (EUSIPCO), IEEE, Piscataway, Aug. 2017, pp. 1420–1424. doi:10.23919/EUSIPCO.2017.8081443.

[114] J.-W. Liu, Z.-J. Cao, J. Liu, X.-L. Luo, W.-M. Li, N. Ito, L.-C. Guo, Phase Retrieval via Wirtinger Flow Algorithm and Its Variants, in: Proceedings of 2019 International Conference on Machine Learning and Cybernetics, IEEE, [Piscataway, New Jersey], 2019, pp. 1–9. doi:10.1109/ICMLC48188.2019.8949170.

[115] A. Maiden, D. Johnson, P. Li, Further improvements to the ptychographical iterative engine, Optica 4 (7) (2017) 736. doi:10.1364/OPTICA.4.000736.

[116] S. Kandel, S. Maddali, Y. S. G. Nashed, S. O. Hruszkewycz, C. Jacobsen, M. Allain, Efficient ptychographic phase retrieval via a matrix-free Levenberg-Marquardt algorithm, Optics express 29 (15) (2021) 23019–23055. doi:10.1364/OE.422768.

[117] Z. Li, K. Lange, J. A. Fessler, Poisson phase retrieval in very low-count regimes, IEEE Transactions on Computational Imaging 8 (2022) 838–850. doi:10.1109/TCI.2022.3209936.

[118] Z. Wen, C. Yang, X. Liu, S. Marchesini, Alternating direction methods for classical and ptychographic phase retrieval, Inverse Problems 28 (11) (2012) 115010. doi:10.1088/0266-5611/28/11/115010.

[119] H. Chang, Y. Lou, Y. Duan, S. Marchesini, Total Variation–Based Phase Retrieval for Poisson Noise Removal, SIAM Journal on Imaging Sciences 11 (1) (2018) 24–55. doi:10.1137/16M1103270.

[120] G. Fatima, Z. Li, A. Arora, P. Babu, PDMM: A Novel Primal-Dual Majorization-Minimization Algorithm for Poisson Phase-Retrieval Problem, IEEE Transactions on Signal Processing 70 (2022) 1241–1255. doi:10.1109/TSP.2022.3156014.

[121] G. Fatima, P. Babu, PGPAL: A Monotonic Iterative Algorithm for Phase-Retrieval Under the Presence of Poisson-Gaussian Noise, IEEE Signal Processing Letters 29 (2022) 533–537. doi:10.1109/LSP.2022.3143469.

[122] E. J. Candès, T. Strohmer, V. Voroninski, PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming, Communications on Pure and Applied Mathematics 66 (8) (2013) 1241–1274. doi:10.1002/cpa.21432.

[123] E. J. Candès, X. Li, Solving Quadratic Equations via PhaseLift When There Are About as Many Equations as Unknowns, Foundations of Computational Mathematics 14 (5) (2014) 1017–1026. doi:10.1007/s10208-013-9162-z.

[124] D. Gross, F. Krahmer, R. Kueng, A Partial Derandomization of PhaseLift Using Spherical Designs, Journal of Fourier Analysis and Applications 21 (2) (2015) 229–266. doi:10.1007/s00041-014-9361-2.

[125] D. Gross, F. Krahmer, R. Kueng, Improved recovery guarantees for phase retrieval from coded diffraction patterns, Applied and Computational Harmonic Analysis 42 (1) (2017) 37–64. doi:10.1016/j.acha.2015.05.004.

[126] R. Kueng, H. Rauhut, U. Terstiege, Low rank matrix recovery from rank one measurements, Applied and Computational Harmonic Analysis 42 (1) (2017) 88–116. doi:10.1016/j.acha.2015.07.007.

[127] T. Goldstein, C. Studer, PhaseMax: Convex Phase Retrieval via Basis Pursuit, IEEE Transactions on Information Theory 64 (4) (2018) 2675–2689. doi:10.1109/TIT.2018.2800768.

[128] R. Ghods, A. S. Lan, T. Goldstein, C. Studer, PhaseLin: Linear phase retrieval, in: 52nd Annual Conference on Information Sciences and Systems (CISS), IEEE, Princeton, NJ, 2018, pp. 1–6. doi:10.1109/CISS.2018.8362270.

[129] F. Krahmer, C. Kümmerle, O. Melnyk, On the robustness of noise-blind low-rank recovery from rank-one measurements, Linear Algebra and its Applications 652 (2022) 37–81. doi:10.1016/j.laa.2022.07.002.

[130] A. Goy, K. Arthur, S. Li, G. Barbastathis, Low Photon Count Phase Retrieval Using Deep Learning, Physical review letters 121 (24) (2018) 243902. doi:10.1103/PhysRevLett.121.243902.

[131] P. Hand, O. Leong, V. Voroninski, Phase Retrieval Under a Generative Prior, in: S. Bengio and H. Wallach and H. Larochelle and K. Grauman and N. Cesa-Bianchi and R. Garnett (Ed.), Advances in Neural Information Processing Systems, Vol. 31, 2018.
URL https://proceedings.neurips.cc/paper/2018/file/1bc2029a8851ad344a8d503930dfd7f7-Paper.pdf

[132] Metzler, Christopher and Schniter, Phillip and Veeraraghavan, Ashok and Richard Baraniuk, prDeep: Robust Phase Retrieval with a Flexible Deep Network, in: Jennifer Dy and Andreas Krause (Ed.), Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 3501–3510.
URL https://proceedings.mlr.press/v80/metzler18a.html

[133] C. Işil, F. S. Oktem, A. Koç, Deep Learning-Based Hybrid Approach for Phase Retrieval, in: Imaging and Applied Optics 2019 (COSI, IS, MATH, pcAOP), OSA, Washington, D.C., 2019, p. CTh2C.5. doi:10.1364/COSI.2019.CTh2C.5.

[134] Y. Nishizaki, R. Horisaki, K. Kitaguchi, M. Saito, J. Tanida, Analysis of non-iterative phase retrieval based on machine learning, Optical Review 27 (1) (2020) 136–141. doi:10.1007/s10043-019-00574-8.

[135] Kazemi, Samia and Yonel, Bariscan and Yazici, Birsen, Unrolled Wirtinger Flow With Deep Decoding Priors for Phaseless Imaging, IEEE Transactions on Computational Imaging 8 (2022) 609–625. doi:10.1109/TCI.2022.3189217.

[136] B. Alexeev, A. S. Bandeira, M. Fickus, D. G. Mixon, Phase Retrieval with Polarization, SIAM Journal on Imaging Sciences 7 (1) (2014) 35–66. doi:10.1137/12089939X.

[137] G. E. Pfander, P. Salanevich, Robust Phase Retrieval Algorithm for Time-Frequency Structured Measurements, SIAM Journal on Imaging Sciences 12 (2) (2019) 736–761. doi:10.1137/18M1205522.

[138] B. Preskitt, Phase Retrieval from Locally Supported Measurements, Doctoral thesis, University of California San Diego, San Diego, California (2018).
URL https://escholarship.org/uc/item/97v5k8j9

[139] O. Melnyk, F. Filbir, F. Krahmer, Phase retrieval from local correlation measurements with fixed shift length, in: 2019 13th International Conference on Sampling Theory and Applications (SampTA), IEEE, Piscataway, NJ, 2019, pp. 1–5. doi:10.1109/SampTA45681.2019.9030967.

[140] A. Forstner, F. Krahmer, O. Melnyk, N. Sissouno, Well-conditioned ptychographic imaging via lost subspace completion, Inverse Problems 36 (10) (2020) 105009. doi:10.1088/1361-6420/abaf3a.

[141] M. Perlmutter, S. Merhi, A. Viswanathan, M. Iwen, Inverting spectrogram measurements via aliased Wigner distribution deconvolution and angular synchronization, Information and Inference: A Journal of the IMA (2020). doi:10.1093/imaiai/iaaa023.

[142] C. Cordor, B. Williams, Y. Hristova, A. Viswanathan, Fast 2D Phase Retrieval using Bandlimited Masks, in: A. Marques, B. Hunyadi (Eds.), 28th European Signal Processing Conference (EUSIPCO 2020), IEEE, [Piscataway, NJ], 2020, pp. 980–984. doi:10.23919/Eusipco47968.2020.9287439.

[143] P. Chen, A. Fannjiang, G.-R. Liu, Phase Retrieval by Linear Algebra, SIAM Journal on Matrix Analysis and Applications 38 (3) (2017) 854–868. doi:10.1137/16M1107747.

[144] R. Chandra, Z. Zhong, J. Hontz, V. McCulloch, C. Studer, T. Goldstein, PhasePack: A Phase Retrieval Library.
URL https://arxiv.org/pdf/1711.10175

[145] H. H. Bauschke, J. M. Borwein, On Projection Algorithms for Solving Convex Feasibility Problems, SIAM Review 38 (3) (1996) 367–426. doi:10.1137/S0036144593251710.

[146] N. Hieu Thao, O. Soloviev, R. Luke, M. Verhaegen, Projection methods for high numerical aperture phase retrieval, Inverse Problems 37 (12) (2021) 125005. doi:10.1088/1361-6420/ac3322.

[147] O. Melnyk, On connections between Amplitude Flow and Error Reduction for phase retrieval, in: Online International Conference on Computational Harmonic Analysis (Online-ICCHA2021), 2021.
URL https://www.univie.ac.at/projektservice-mathematik/e/talks/Melnyk_2021-06_melnyk_ICCHA.pdf

[148] O. Melnyk, On connections between Amplitude Flow and Error Reduction for phase retrieval and ptychography, Sampling Theory, Signal Processing, and Data Analysis 20 (2) (2022) 16. doi:10.1007/s43670-022-00035-5.

[149] E. Levin, T. Bendory, A note on Douglas-Rachford, gradients, and phase retrieval. URL https://arxiv.org/pdf/1911.13179

[150] P. Thibault, A. Menzel, Reconstructing state mixtures from diffraction measurements, Nature 494 (7435) (2013) 68–71. doi:10.1038/nature11806.

[151] G. Wang, G. B. Giannakis, J. Chen, Scalable Solvers of Random Quadratic Equations via Stochastic Truncated Amplitude Flow, IEEE Transactions on Signal Processing 65 (8) (2017) 1961–1974. doi:10.1109/TSP.2017.2652392.

[152] Z. Xiao, Zhang, Yerong and Yang, Jie, Large-Scale Phase Retrieval via Stochastic Reweighted Amplitude Flow, KSII Transactions on Internet and Information Systems 14 (11) (2020) 4355–4371. doi:10.3837/tiis.2020.11.006.

[153] S. Kaczmarz, Angenäherte Auflösung von Systemen linearer Gleichungen, Bulletin International de l'Académie Polonaise des Sciences et des Lettres (1937) 355–357.

[154] K. Wei, Solving systems of phaseless equations via Kaczmarz methods: a proof of concept study, Inverse Problems 31 (12) (2015) 125008. doi:10.1088/0266-5611/31/12/125008.

[155] Y. S. Tan, R. Vershynin, Phase retrieval via randomized Kaczmarz: theoretical guarantees, Information and Inference: A Journal of the IMA 8 (1) (2019) 97–123. doi:10.1093/imaiai/iay005.

[156] T. Zhang, Y. Feng, Phase retrieval of complex-valued objects via a randomized Kaczmarz method, Information and Inference: A Journal of the IMA (2021). doi:10.1093/imaiai/iaab017.

[157] P. Römer, F. Filbir, F. Krahmer, On the randomized Kaczmarz algorithm for phase retrieval, in: 2021 55th Asilomar Conference on Signals, Systems, and Computers, IEEE, 10/31/2021 - 11/3/2021, pp. 847–851. doi:10.1109/IEEECONF53345.2021.9723291.

[158] O. Melnyk, Stochastic Amplitude Flow for phase retrieval, its convergence and doppelgängers (2022).
URL https://arxiv.org/abs/2212.04916

[159] B. Preskitt, R. Saab, Admissible Measurements and Robust Algorithms for Ptychography, Journal of Fourier Analysis and Applications 27 (2) (2021). doi:10.1007/s00041-021-09811-8.

[160] R. Saab, B. Preskitt, M. Iwen, A. Viswanathan, Phase retrieval from local measurements in two dimensions, in: Y. M. Lu, D. van de Ville, M. Papadakis (Eds.), Wavelets and sparsity XVII, Proceedings of SPIE, 0277-786X, SPIE, Bellingham, Washington, 2017, p. 30. doi:10.1117/12.2274355.

[161] P. Li, T. B. Edo, J. M. Rodenburg, Ptychographic inversion via Wigner distribution deconvolution: noise suppression and probe design, Ultramicroscopy 147 (2014) 106–113. doi:10.1016/j.ultramic.2014.07.004.

[162] M. S. Richman, T. W. Parks, R. G. Shenoy, Discrete-time, discrete-frequency, time-frequency analysis, IEEE Transactions on Signal Processing 46 (6) (1998) 1517–1527. doi:10.1109/78.678465.

[163] J. O'Toole, M. Mesbah, B. Boashash, A Discrete Time and Frequency Wigner Distribution: Properities and Implementation, in: International Symposium on Digital Signal Processing and Communication Systems, 2005.
URL https://eprints.qut.edu.au/2607/

[164] F. Filbir, F. Krahmer, O. Melnyk, On Recovery Guarantees for Angular Synchronization, Journal of Fourier Analysis and Applications 27 (2) (2021). doi:10.1007/s00041-021-09834-1.

[165] E. J. Candès, X. Li, M. Soltanolkotabi, Phase retrieval from coded diffraction patterns, Applied and Computational Harmonic Analysis 39 (2) (2015) 277–299. doi:10.1016/j.acha.2014.09.004.

[166] I. Waldspurger, A. d'Aspremont, S. Mallat, Phase recovery, MaxCut and complex semidefinite programming, Mathematical Programming 149 (1-2) (2015) 47–81. doi:10.1007/s10107-013-0738-9.

[167] F. R. K. Chung, Spectral graph theory, American Mathematical Society, Providence, RI, 1997.

[168] A. S. Bandeira, A. Singer, D. A. Spielman, A Cheeger Inequality for the Graph Connection Laplacian, SIAM Journal on Matrix Analysis and Applications 34 (4) (2013) 1611–1630. doi:10.1137/120875338.

[169] M. Fazel, Matrix Rank Minimization with Applications, Ph.D. thesis, Stanford University (March 2002).
URL https://faculty.washington.edu/mfazel/thesis-final.pdf

[170] M. X. Goemans, D. P. Williamson, Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, Journal of the ACM (JACM) 42 (6) (1995) 1115–1145. doi:10.1145/227683.227684.

[171] A. M.-C. So, J. Zhang, Y. Ye, On approximating complex quadratic optimization problems via semidefinite programming relaxations, Mathematical Programming 110 (1) (2007) 93–110. doi:10.1007/s10107-006-0064-6.

[172] N. Boumal, Nonconvex Phase Synchronization, SIAM Journal on Optimization 26 (4) (2016) 2355–2377. doi:10.1137/16M105808X.

[173] A. S. Bandeira, N. Boumal, A. Singer, Tightness of the maximum likelihood semidefinite relaxation for angular synchronization, Mathematical Programming 163 (1-2) (2017) 145–167. doi:10.1007/s10107-016-1059-6.

[174] H. Liu, M.-C. Yue, A. Man-Cho So, On the Estimation Performance and Convergence Rate of the Generalized Power Method for Phase Synchronization, SIAM Journal on Optimization 27 (4) (2017) 2426–2446. doi:10.1137/16M110109X.

[175] Y. Zhong, N. Boumal, Near-Optimal Bounds for Phase Synchronization, SIAM Journal on Optimization 28 (2) (2018) 989–1016. doi:10.1137/17M1122025.

[176] A. Singer, Angular Synchronization by Eigenvectors and Semidefinite Programming, Applied and Computational Harmonic Analysis 30 (1) (2011) 20–36. doi:10.1016/j.acha.2010.02.001.

[177] G. Lerman, Y. Shi, Robust Group Synchronization via Cycle-Edge Message Passing, Foundations of Computational Mathematics (2021). doi:10.1007/s10208-021-09532-w.

[178] Y. Shi, G. Lerman, Message Passing Least Squares Framework and its Application to Rotation Synchronization, in: III, Hal Daumé and Singh, Aarti (Ed.), Proceedings of Machine Learning Research, Vol. 119, PMLR, 2020, pp. 8796–8806.
URL https://proceedings.mlr.press/v119/shi20b.html

[179] H. Liu, M.-C. Yue, A. M.-C. So, A Unified Approach to Synchronization Problems over Subgroups of the Orthogonal Group. doi:10.2139/ssrn.4120722.

[180] T. Maunu, G. Lerman, Depth Descent Synchronization in SO(D), International Journal of Computer Vision 131 (4) 968–986. doi:10.1007/s11263-022-01686-6.

[181] E. J. Candès, B. Recht, Exact Matrix Completion via Convex Optimization, Foundations of Computational Mathematics 9 (6) (2009) 717–772. doi:10.1007/s10208-009-9045-5.

[182] A. Fannjiang, P. Chen, Blind ptychography: uniqueness and ambiguities, Inverse Problems 36 (4) (2020) 045005. doi:10.1088/1361-6420/ab6504.

[183] T. Bendory, D. Edidin, Y. C. Eldar, Blind Phaseless Short-Time Fourier Transform Recovery, IEEE Transactions on Information Theory 66 (5) (2020) 3232–3241. doi:10.1109/TIT.2019.2947056.

[184] A. Beck, On the Convergence of Alternating Minimization for Convex Programming with Applications to Iteratively Reweighted Least Squares and Decomposition Schemes, SIAM Journal on Optimization 25 (1) (2015) 185–209. doi:10.1137/13094829X.

[185] N. Bissantz, L. Dümbgen, A. Munk, B. Stratmann, Convergence Analysis of Generalized Iteratively Reweighted Least Squares Algorithms on Convex Function Spaces, SIAM Journal on Optimization 19 (4) (2009) 1828–1845. doi:10.1137/050639132.

[186] I. Daubechies, R. DeVore, M. Fornasier, C. S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, Communications on Pure and Applied Mathematics 63 (1) (2010) 1–38. doi:10.1002/cpa.20303.

[187] A. Beck, L. Tetruashvili, On the Convergence of Block Coordinate Descent Type Methods, SIAM Journal on Optimization 23 (4) (2013) 2037–2060. doi:10.1137/120887679.

[188] H. Chang, P. Enfedaque, S. Marchesini, Blind Ptychographic Phase Retrieval via Convergent Alternating Direction Method of Multipliers, SIAM Journal on Imaging Sciences 12 (1) (2019) 153–185. doi:10.1137/18M1188446.

[189] R. Hesse, D. R. Luke, S. Sabach, M. K. Tam, Proximal Heterogeneous Block Implicit-Explicit Method and Application to Blind Ptychographic Diffraction Imaging, SIAM Journal on Imaging Sciences 8 (1) (2015) 426–457. doi:10.1137/14098168X.

[190] N. Burdet, X. Shi, D. Parks, J. N. Clark, X. Huang, S. D. Kevan, I. K. Robinson, Evaluation of partial coherence correction in X-ray ptychography, Optics express 23 (5) (2015) 5452–5467. doi:10.1364/OE.23.005452.

[191] Y. Guo, A. Wang, W. Wang, Multi-source phase retrieval from multi-channel phaseless STFT measurements, Signal Processing 144 (2018) 36–40. doi:10.1016/j.sigpro.2017.09.026.

[192] M. Odstrčil, A. Menzel, M. Guizar-Sicairos, Iterative least-squares solver for generalized maximum-likelihood ptychography, Optics express 26 (3) (2018) 3108–3123. doi:10.1364/OE.26.003108.

[193] X. Wei, P. Urbach, Ptychography with multiple wavelength illumination, Optics express 27 (25) (2019) 36767–36789. doi:10.1364/OE.27.036767.

[194] C. A. Metzler, G. Wetzstein, Deep S 3 PR: Simultaneous Source Separation and Phase Retrieval Using Deep Generative Models, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 6/6/2021 - 6/11/2021, pp. 1370–1374. doi:10.1109/ICASSP39728.2021.9413714.

[195] F. Filbir, O. Melnyk, Image Recovery for Blind Polychromatic Ptychography (2022). URL https://arxiv.org/abs/2210.01626

[196] M. Hirose, T. Higashino, N. Ishiguro, Y. Takahashi, Multibeam ptychography with synchrotron hard X-rays, Optics express 28 (2) (2020) 1216–1224. doi:10.1364/OE.378083.

[197] X. Huang, H. Yan, R. Harder, Y. Hwu, I. K. Robinson, Y. S. Chu, Optimization of overlap uniformness for ptychography, Optics express 22 (10) (2014) 12634–12644. doi:10.1364/OE.22.012634.

[198] Z. Fabian, J. Haldar, R. Leahy, M. Soltanolkotabi, 3D Phase Retrieval at Nano-Scale via Accelerated Wirtinger Flow, in: A. Marques, B. Hunyadi (Eds.), 28th European Signal Processing Conference (EUSIPCO 2020), IEEE, [Piscataway, NJ], 2020, pp. 2080–2084. doi:10.23919/Eusipco47968.2020.9287703.

[199] W. Gautschi, On inverses of Vandermonde and confluent Vandermonde matrices, Numerische Mathematik 4 (1) (1962) 117–123. doi:10.1007/BF01386302.

[200] D. Nagel, The condition number of Vandermonde matrices and its application to the stability analysis of a subspace method, Doctoral thesis, University of Osnabrück, Osnabrück (19.03.2021).
URL https://osnadocs.ub.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-202103194121

Appendix A

Proof of Lemma 3.6.48

The proof of Lemma 3.6.48 consists of several steps. The first step is to decompose the matrix $M^{j,r}$ defined in (3.121) into several simple components. Let us start with the entries of the vector $\overline{w}^{(q)} \circ S_j w^{(q)}$. We observe that for all $j \in [\delta]$ and $q \in [s]$ we have

$$[\overline{w}^{(q)} \circ S_j w^{(q)}]_t = \overline{w}_t^{(q)} w_{t-j}^{(q)} = e^{-t\alpha_q} \mathcal{I}_{t\in[\delta]} e^{-(t-j)\alpha_q} \mathcal{I}_{t-j\in[\delta]} = e^{-2t\alpha_q} e^{j\alpha_q} \mathcal{I}_{t\in\{j,j+1,\ldots,\delta-1\}}.$$

Consequently, the entries of the matrices $M^{j,r}$, $r \in [d/s]$, are given by

$$\begin{aligned} M_{q,k}^{j,r} &= \overline{F_d[\overline{w}^{(q)} \circ S_j w^{(q)}]}_{r-kd/s} = \sum_{t\in[d]} \overline{[\overline{w}^{(q)} \circ S_j w^{(q)}]}_t e^{\frac{2\pi i(r-kd/s)t}{d}} \\ &= e^{j\alpha_q} \sum_{t=j}^{\delta-1} e^{-2t\alpha_q} e^{\frac{2\pi i(r-kd/s)t}{d}}, \end{aligned} \tag{A.1}$$

for all $k \in [s]$. Let us define

$$\gamma_j := \left\lfloor \frac{\delta - j}{s} \right\rfloor, \quad \text{and} \quad \beta_j := \delta - j - \gamma_j s.$$

Then, the proof splits into two cases.

A.1 Case $\gamma_j = 0$

Let us separately consider the case $\gamma_j = 0$, which is equivalent to $\delta - j < s$. In view of Remark 3.6.46 and by construction of the space $\mathbb{T}_{\delta,s}$ we have

$$d^j(U)_{t+ps} = 0, \quad 0 \le t < s - (\delta - j),\ p \in [d/s].$$

Consequently, we can reduce the number of unknowns in the linear systems (3.122). That is, we consider the vectors $v^{j,r}$ given by

$$v_k^{j,r} = F_d[d^j(U)]_{r-kd/s}, \quad k \in [s],\ r \in [d/s],$$

and substitute the identity for diagonals above. This yields

$$\begin{aligned} v_k^{j,r} &= \sum_{t\in[s]} \sum_{p\in[d/s]} d^j(U)_{t+ps} e^{-\frac{2\pi i(r-kd/s)(t+ps)}{d}} \\ &= \sum_{t=s-(\delta-j)}^{s-1} \sum_{p\in[d/s]} d^j(U)_{t+ps} e^{-\frac{2\pi irt}{d}} e^{\frac{2\pi ikt}{s}} e^{-\frac{2\pi irps}{d}} e^{2\pi ikp} \\ &= \sum_{t\in[\delta-j]} \sum_{p\in[d/s]} d^j(U)_{t+(p+1)s-\delta+j} e^{-\frac{2\pi irt}{d}} e^{\frac{2\pi ikt}{s}} e^{-\frac{2\pi ir[(p+1)s-\delta+j]}{d}} e^{\frac{2\pi ik[s(p+1)-\delta+j]}{s}} \\ &= e^{\frac{2\pi ikj}{s}} \sum_{t\in[\delta-j]} \left[\sum_{p\in[d/s]} d^j(U)_{t+ps} e^{-\frac{2\pi ir[(p+1)s-\delta+j]}{d}} \right] e^{-\frac{2\pi irt}{d}} e^{\frac{2\pi ikt}{s}}. \end{aligned}$$

With supporting vectors

$$\begin{aligned} u_t^{j,r} &:= \sum_{p\in[d/s]} d^j(U)_{t+(p+1)s-\delta+j} e^{-\frac{2\pi ir[(p+1)s-\delta+j]}{d}}, \quad u^r \in \mathbb{C}^{\delta-j}, \\ f_t^r &:= e^{-\frac{2\pi irt}{d}}, \quad f^r \in \mathbb{C}^s, \end{aligned}$$

we can rewrite the entries of $v_k^{j,r}$ as

$$v_k^{j,r} = e^{\frac{2\pi ikj}{s}} \sum_{t\in[s]} [P_{\delta-j}^* u^{j,r} \circ f^r]_t e^{\frac{2\pi ikt}{s}} = e^{\frac{2\pi ikj}{s}} F_s^* [P_{\delta-j}^* u^{j,r} \circ f^r]_k = [M_j F_s^* \operatorname{diag}(f^r) P_{\delta-j}^* u^{j,r}]_k,$$

where M_j is the modulation operator (2.7). Returning to $M^{j,r}$ and its representation (A.1), we note that it simplifies to

$$\begin{aligned} M_{q,k}^{j,r} &= e^{j\alpha_q} \sum_{t=j}^{\delta-1} e^{-2t\alpha_q} e^{\frac{2\pi i(r-kd/s)t}{d}} \\ &= e^{-j\alpha_q} e^{\frac{2\pi irj}{d}} e^{-\frac{2\pi ikj}{s}} \sum_{t\in[\delta-j]} e^{-2t\alpha_q} e^{\frac{2\pi irt}{d}} e^{-\frac{2\pi ikt}{s}}. \end{aligned}$$

For further convenience, we introduce the following notation

$$\begin{aligned} a_q &:= e^{-2\alpha_q}, \quad 1 > a_0 > a_1 > \ldots > a_{s-1} > 0, \\ c^{q,n} &:= (1, a_q^1, a_q^2, \ldots, a_q^{n-1})^T \in \mathbb{R}^n, \\ d_q^{j,r} &:= e^{-j\alpha_q} e^{\frac{2\pi irj}{d}}, \quad d^{j,r} \in \mathbb{C}^s, \end{aligned} \tag{A.2}$$

Consequently, the entries of the matrix $M_{q,k}^{j,r}$ with new notation are given by

$$M_{q,k}^{j,r} = (M_{-j})_{k,k} d_q^{j,r} \sum_{t\in[s]} [P_{\delta-j}^* c^{q,\delta-j} \circ \overline{f^r}]_t e^{-\frac{2\pi ikt}{s}},$$

and $M^{j,r}$ is then

$$\begin{aligned} M^{j,r} &= \operatorname{diag}(d^{j,r}) \begin{bmatrix} -P_{\delta-j}^* c^{0,\delta-j} - \\ \vdots \\ -P_{\delta-j}^* c^{s,\delta-j} - \end{bmatrix} \operatorname{diag}(\overline{f^r}) F_s M_{-j} \\ &=: \operatorname{diag}(d^{j,r}) C^{\delta-j} P_{\delta-j} \operatorname{diag}(\overline{f^r}) F_s M_{-j}. \end{aligned}$$

Then, the linear system (3.122) transforms into

$$b_q^{j,r} = \frac{1}{s}\operatorname{diag}(d^{j,r})C^{\delta-j}P_{\delta-j}\operatorname{diag}(\overline{f^r})F_sM_{-j}M_jF_s^*\operatorname{diag}(f^r)P_{\delta-j}^*u^{j,r} + n^{j,r}.$$

By Propositions 2.2.1 and 2.2.2 we have $M_{-j}M_j = I_s$ and $\frac{1}{s}F_sF_s^* = I_s$. Also, vectors f^r satisfy $|f_t^r| = 1$, $t \in [s]$, so that

$$\operatorname{diag}(\overline{f^r})\operatorname{diag}(f^r) = \operatorname{diag}(|f^r|^2) = I_s.$$

Moreover, by (2.11), $P_{\delta-j}P_{\delta-j}^* = I_{\delta-j}$, and thus, the system simplifies to

$$b_q^{j,r} = \operatorname{diag}(d^{j,r})C^{\delta-j}u^{j,r} + n^{j,r}.$$

Note that $\operatorname{diag}(d^{j,r})$ is diagonal with nonzero entries and, thus, it is invertible. The $s \times (\delta - j)$ matrix $C^{\delta-j}$ is the tall Vandermonde matrix with unique generating entries a_q, $q \in [s]$, and, hence, injective. Consequently, the vectors $u^{j,r}$ can be recovered from the measurements by an application of the pseudoinverse

$$u^{j,r} = (C^{\delta-j})^\dagger \operatorname{diag}(d^{j,r})^{-1}b_q^{j,r}.$$

The last step of the proof for the case $\gamma_j = 0$ is to note that

$$u_t^{j,r} = e^{-\frac{2\pi ir[s-\delta+j]}{d}}\sum_{p\in[d/s]} d^j(U)_{t+(p+1)s-\delta+j}e^{-\frac{2\pi irp}{d/s}} = e^{-\frac{2\pi ir[s-\delta+j]}{d}}F_{d/s}[d^j(U)_{t+s-\delta+j+\cdot s}]_r,$$

and, thus, the non-zero entries of diagonal $d^j(U)$ are recovered via the inverse Fourier transform

$$d^j(U)_{t+s-\delta+j+ps} = \frac{s}{d}\sum_{r\in[d/s]} e^{\frac{2\pi ir[s-\delta+j]}{d}}u_t^{j,r}e^{\frac{2\pi irp}{d/s}}, \quad t \in [\delta - j],\ p \in [d/s].$$

A.2 Case $\gamma_j > 0$

In this case we continue to transform representation (A.1) of the matrix $M^{j,r}$. We split $t \in \{j, j+1\ldots, \delta-1\}$ into $t = j + t_1s + t_2$, where $t_1 \in [\gamma_j]$ and $t_2 \in [s]$. This leads to

$$\begin{aligned}
M_{q,k}^{j,r} &= e^{-j\alpha_q}e^{\frac{2\pi i(r-kd/s)j}{d}}\sum_{t\in[\delta-j]} e^{-2t\alpha_q}e^{\frac{2\pi i(r-kd/s)t}{d}} \\
&= e^{-j\alpha_q}e^{\frac{2\pi irj}{d}}e^{-\frac{2\pi ikj}{s}}\left[\sum_{t_1\in[\gamma_j]}\sum_{t_2\in[s]} e^{-2(t_1s+t_2)\alpha_q}e^{\frac{2\pi i(r-kd/s)(t_1s+t_2)}{d}}\right. \\
&\qquad \left. + \sum_{t_2\in[\beta_j]} e^{-2(\gamma_js+t_2)\alpha_q}e^{\frac{2\pi i(r-kd/s)(\gamma_js+t_2)}{d}}\right] \\
&= e^{-j\alpha_q}e^{\frac{2\pi irj}{d}}e^{-\frac{2\pi ikj}{s}}\left[\sum_{t_1\in[\gamma_j]} e^{-2t_1s\alpha_q}e^{\frac{2\pi irt_1s}{d}}e^{-2\pi ikt_1}\sum_{t_2\in[s]} e^{-2t_2\alpha_q}e^{\frac{2\pi irt_2}{d}}e^{-\frac{2\pi ikt_2}{s}} + \right. \\
&\qquad \left. + e^{-2\gamma_js\alpha_q}e^{\frac{2\pi ir\gamma_js}{d}}e^{-2\pi ik\gamma_j}\sum_{t_2\in[\beta_j]} e^{-2t_2\alpha_q}e^{\frac{2\pi irt_2}{d}}e^{-\frac{2\pi ikt_2}{s}}\right].
\end{aligned} \tag{A.3}$$

Note that the exponents $e^{-2\pi i k t_1}$ and $e^{-2\pi i k \gamma_j}$ are equal to one and vanish. With the vector

$$d_q^{2,j,r} := \frac{\sum_{t\in[\gamma_j]} e^{-2ts\alpha_q} e^{\frac{2\pi i r t s}{d}}}{e^{-2\gamma_j s\alpha_q} e^{\frac{2\pi i r \gamma_j s}{d}}}, \quad d^{2,j,r} \in \mathbb{C}^s$$

we rewrite the entries of $M^{j,r}$ as

$$M_{q,k}^{j,r} = d_q^{j,r}(M_{-j})_{k,k}\left[d_q^{2,j,r}[F_s \operatorname{diag}(\overline{f^r})c^{q,j,s}]_k + [F_s \operatorname{diag}(\overline{f^r})P_{\beta_j}^* c^{q,j,\beta_j}]_k\right],$$

or the full matrix as

$$M^{j,r} = \operatorname{diag}(d_q^{j,r})\left[\operatorname{diag}(d^{2,j,r})C^s + C^{\beta_j}P_{\beta_j}\right]\operatorname{diag}(\overline{f^r})F_s M_{-j}.$$

The matrices $\operatorname{diag}(d_q^{j,r})$ and $\operatorname{diag}(\overline{f^r})$ are diagonal with non-zero entries and, therefore, invertible. The matrices F_s and M_{-j} are also invertible and, thus, $M^{j,r}$ is invertible if and only if the matrix

$$G^{j,r} := \operatorname{diag}(d^{2,j,r})C^s + C^{\beta_j}P_{\beta_j}$$

is invertible. In order to obtain the invertibility of $G^{j,r}$, we show that its singular values are positive by employing the following inequalities for singular values.

Theorem A.2.1 (Weil's inequality for singular values, [48, Problem III.6.5])**.** *Consider $A, B \in \mathbb{C}^{p\times n}$ for some $1 \le p \le n$. Then, for all $j \in [p]$ we have*

$$|\sigma_j(A+B) - \sigma_j(A)| \le \sigma_1(B),$$

which is equivalent to

$$\sigma_j(A) - \sigma_1(B) \le \sigma_j(A+B) \le \sigma_j(A) + \sigma_1(B).$$

In addition, we require a bound for singular values of a product of two matrices.

Theorem A.2.2 (Multiplication bounds [48, Problem III.6.2])**.** *Consider $A \in \mathbb{C}^{p\times n}$ and $B \in \mathbb{C}^{n\times r}$ for some $1 \le p \le n$ and $1 \le r$. Then, for all $j \in [p]$ we have*

$$\sigma_j(AB) \le \sigma_j(A)\sigma_1(B).$$

In particular, if one of the matrices is invertible, this leads to the following corollary.

Corollary A.2.3. *Consider $A \in \mathbb{C}^{n\times p}$ and invertible $B \in \mathbb{C}^{n\times n}$ for some $1 \le p \le n$. Then, for all $j \in [p]$ we have*

$$\sigma_j(A)\sigma_n(B) \le \sigma_j(BA) \le \sigma_j(A)\sigma_1(B),$$

Proof. For the upper bound we apply the previous theorem to $(BA)^T$. That is, for $j \in [p]$ we have

$$\sigma_j(BA) = \sigma_j((BA)^T) = \sigma_j(A^T B^T) \le \sigma_j(A^T)\sigma_1(B^T) = \sigma_j(A)\sigma_1(B).$$

For the lower bound we apply the established inequality to $\tilde{A} = BA$ and $\tilde{B} = B^{-1}$, which gives

$$\sigma_j(A) = \sigma_j(B^{-1}BA) = \sigma_j(\tilde{B}\tilde{A}) \le \sigma_j(\tilde{A})\sigma_1(\tilde{B}) = \sigma_j(BA)\sigma_1(B^{-1}) = \sigma_j(BA)\frac{1}{\sigma_n(B)}.$$

□

Furthermore, we would require the bounds on the singular values for each component of $G^{j,r}$. For $\operatorname{diag}(d^{2,j,r})$ we have the following statement.

Lemma A.2.4. *The minimal singular value of the matrix* $\operatorname{diag}(d^{2,j,r})$ *satisfies*

$$\sigma_s(\operatorname{diag}(d^{2,j,r})) \geq \frac{a_0^{-\gamma_j s} - 1}{1 + a_0^s},$$

with a_0 *given by* (A.2).

Proof. The matrix $\operatorname{diag}(d^{2,j,r})$ is diagonal and, thus, we have

$$\begin{aligned}\sigma_s(\operatorname{diag}(d^{2,j,r})) &= \min_{q\in[s]} |d_q^{2,j,r}| = \min_{q\in[s]} \left| \frac{\sum_{t\in[\gamma_j]} e^{-2ts\alpha_q} e^{\frac{2\pi irts}{d}}}{e^{-2\gamma_j s\alpha_q} e^{\frac{2\pi ir\gamma_j s}{d}}} \right| \\ &= \min_{q\in[s]} \left| \frac{1 - e^{-2\gamma_j s\alpha_q} e^{\frac{2\pi ir\gamma_j s}{d}}}{e^{-2\gamma_j s\alpha_q} \left(1 - e^{-2s\alpha_q} e^{\frac{2\pi irs}{d}}\right)} \right|,\end{aligned}$$

where we used that the sum is a geometric sum with at least one non-zero summand due to $\gamma_j > 0$. Computing the absolute values results in

$$\begin{aligned}\sigma_s(\operatorname{diag}(d^{2,j,r})) &= \min_{q\in[s]} \frac{\sqrt{1 + e^{-4\gamma_j s\alpha_q} - 2e^{-2\gamma_j s\alpha_q} \cos\left(\frac{2\pi r\gamma_j s}{d}\right)}}{e^{-2\gamma_j s\alpha_q} \sqrt{1 + e^{-4s\alpha_q} - 2e^{-2s\alpha_q} \cos\left(\frac{2\pi rs}{d}\right)}} \\ &\geq \min_{q\in[s]} \frac{\sqrt{1 + e^{-4\gamma_j s\alpha_q} - 2e^{-2\gamma_j s\alpha_q}}}{e^{-2\gamma_j s\alpha_q} \sqrt{1 + e^{-4s\alpha_q} + 2e^{-2s\alpha_q}}} \\ &= \min_{q\in[s]} \frac{1 - e^{-2\gamma_j s\alpha_q}}{e^{-2\gamma_j s\alpha_q}(1 + e^{-2s\alpha_q})} = \min_{q\in[s]} \frac{e^{2(\gamma_j+1)s\alpha_q} - e^{2s\alpha_q}}{1 + e^{2s\alpha_q}}.\end{aligned}$$

Let us consider the function

$$f(x) = \frac{x^{\gamma_j+1} - x}{1 + x},$$

for $x > 1$. Its derivative is given by

$$f'(x) = \frac{((\gamma_j + 1)x^{\gamma_j} - 1)(1 + x) - (x^{\gamma_j+1} - x)}{(1 + x)^2} = \frac{\gamma_j x^{\gamma_j+1} + (\gamma_j + 1)x^{\gamma_j} - 1}{(1 + x)^2}.$$

Since $x > 1$ and $\gamma_j > 0$, we have

$$f'(x) > \frac{0 + (0 + 1) \cdot 1 - 1}{(1 + x)^2} = 0,$$

and, thus, $f(x)$ is increasing for $x > 1$. Recall that by the assumption of Lemma 3.6.48 the parameters α_q, $q \in [s]$, satisfy $0 < \alpha_0 < \alpha_1 < \ldots < \alpha_{s-1}$. Hence, by the monotonicity of the exponent

$$1 < e^{2s\alpha_0} < e^{2s\alpha_1} < \ldots < e^{2s\alpha_{s-1}},$$

and by the monotonicity of f, we get

$$\sigma_s(\operatorname{diag}(d^{2,j,r})) \geq \frac{e^{2(\gamma_j+1)s\alpha_0} - e^{2s\alpha_0}}{1 + e^{2s\alpha_0}} = \frac{a_0^{-s\gamma_j} - 1}{1 + a_0^s}.$$

□

For the Vandermonde matrix, the next lemma provides bounds on the maximal and the minimal singular values.

Lemma A.2.5. *Consider $1 \leq n \leq s$. Let V be a $s \times n$ Vandermonde matrix*

$$V = \begin{bmatrix} 1 & a_0 & a_0^2 & \dots & a_0^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & a_{s-1} & a_{s-1}^2 & \dots & a_{s-1}^{n-1} \end{bmatrix},$$

with $1 > a_0 > a_1 > \ldots > a_{s-1} > 0$ given by (A.2). Then,

$$\sigma_1(V) \leq \sqrt{sn}.$$

Furthermore, if $n = s$ we have

$$\sigma_s(V) \geq \frac{\min_{q\in[s-1]}(a_q - a_{q+1})^{s-1}}{\sqrt{s}(1+a_0)^{s-1}}.$$

Proof. For the maximal singular value, we use that $\sigma_1^2(V) = \|V\|_\infty \leq \|V\|_F$ and $a_q \leq 1$, which leads to

$$\sigma_1^2(V) \leq \|V\|_F^2 = \sum_{q\in[s]} \sum_{j\in[n]} |a_q^j|^2 \leq sn.$$

Now, set n equal to s, so that V is a square matrix. For the bound on the minimal singular value of the inverse matrix, we apply the results of [199] (see also [200, pp. 46-47 and Lemma 2.1.8]), which states that

$$\left\|V^{-1}\right\|_\infty \leq \sqrt{s} \max_{q\in[s]} \prod_{\substack{p\in[s] \\ p\neq q}} \frac{1+a_p}{|a_q - a_p|}$$

First, let us bound the denominator from below using the monotonicity of a_q. For all $q \in [s]$, we have

$$\begin{aligned} \prod_{\substack{p\in[s] \\ p\neq q}} |a_q - a_p| &= \prod_{p=0}^{q-1} (a_q - a_p) \prod_{p=q+1}^{s-1} (a_p - a_q) \\ &\geq (a_{q-1} - a_q)^{q-1} (a_q - a_{q+1})^{s-q} \geq \min_{q\in[s-1]} (a_q - a_{q+1})^{s-1}. \end{aligned}$$

Then, we obtain

$$\left\|V^{-1}\right\|_\infty \leq \sqrt{s} \frac{\max_{q\in[s]} \prod_{\substack{p\in[s] \\ p\neq q}} (1+a_p)}{\min_{q\in[s-1]}(a_q - a_{q+1})^{s-1}} \leq \sqrt{s} \frac{(1+a_0)^{s-1}}{\min_{q\in[s-1]}(a_q - a_{q+1})^{s-1}}.$$

The equality $\|V^{-1}\|_\infty = \sigma_s(V)^{-1}$ leads to the desired bound,

$$\sigma_s(V) \geq \frac{\min_{q\in[s-1]}(a_q - a_{q+1})^{s-1}}{\sqrt{s}(1+a_0)^{s-1}}.$$

□

Continuing with the proof of Lemma 3.6.48, we now bound the smallest singular value of the matrix $G^{j,r}$ using the established above bounds. That is, by Theorem A.2.1 and Corollary A.2.3, we obtain

$$\sigma_s(G^{j,r}) \geq \sigma_s(\mathrm{diag}(d^{2,j,r})C^s) - \sigma_1(C^{\beta_j}P_{\beta_j}) \geq \sigma_s(\mathrm{diag}(d^{2,j,r}))\sigma_s(C^s) - \sigma_1(C^{\beta_j}P_{\beta_j}).$$

Since the multiplication with P_{β_j} only appends zero columns to C^{β_j}, the spectral norm of $C^{\beta_j}P_{\beta_j}$ is equal to the spectral norm of C^{β_j}, which is bounded from above by Lemma A.2.5. Consequently, by Lemma A.2.4 and Lemma A.2.5 we have

$$\begin{aligned}\sigma_s(G^{j,r}) &\geq \frac{a_0^{-\gamma_j s}-1}{1+a_0^s}\cdot\frac{\min_{q\in[s-1]}(a_q-a_{q+1})^{s-1}}{\sqrt{s}(1+a_0)^{s-1}} - \sqrt{s\beta_j}\\ &> \frac{a_0^{-s}-1}{1+a_0^s}\cdot\frac{\min_{q\in[s-1]}(a_q-a_{q+1})^{s-1}}{\sqrt{s}(1+a_0)^{s-1}} - s,\end{aligned}$$

where we used that $\gamma_j \geq 1$ and $\beta_j < s$ in the second line. Now, let us apply the assumptions on α_q, $q \in [s]$. That is, the inequality

$$\alpha_{q+1} - \tfrac{1}{2}\log 2 \geq \alpha_q, \quad q \in [s-1],$$

is equivalent to

$$2a_{q+1} \leq a_q \quad \text{or} \quad a_{q+1} \leq a_q - a_{q+1},$$

and, hence,

$$\min_{q\in[s-1]}(a_q-a_{q+1})^{s-1} \geq \min_{q\in[s-1]} a_{q+1}^{s-1} = a_{s-1}^{s-1}.$$

This leads to

$$\sigma_s(G^{j,r}) > \frac{(a_0^{-s}-1)a_{s-1}^{s-1}}{\sqrt{s}(1+a_0^s)(1+a_0)^{s-1}} - s = \frac{(1-a_0^s)a_0^{-s}a_{s-1}^{s-1}}{\sqrt{s}(1+a_0^s)(1+a_0)^{s-1}} - s.$$

The second assumption

$$\alpha_0 \geq \frac{s-1}{s}\alpha_{s-1} + \frac{3}{4s}\log s + \frac{(s+1)}{2s}\log 2$$

is equivalent to

$$a_0^{-s} \geq a_{s-1}^{-(s-1)}s^{3/2}2^{s+1}$$

and implies $a_0^s \leq 1/2 < 1$. Therefore, we obtain

$$\begin{aligned}\sigma_s(G^{j,r}) &> \frac{(1-a_0^s)a_{s-1}^{-(s-1)}s^{3/2}2^{s+1}a_{s-1}^{s-1}}{\sqrt{s}(1+a_0^s)(1+a_0)^{s-1}} - s\\ &= \frac{(1-a_0^s)2^{s+1}s}{(1+a_0^s)(1+a_0)^{s-1}} - s\\ &> \frac{2^{-1}2^{s+1}s}{2\cdot 2^{s-1}} - s = s - s = 0,\end{aligned}$$

so that the matrices $G^{j,r}$ and $M^{j,r}$ are invertible and the proof is concluded.

Finally, we show that there exists a feasible set of parameters α_q, $q \in [s]$. Let us consider the case when the first assumption holds with all equalities,

$$\alpha_{q+1} = \alpha_q + \tfrac{1}{2}\log 2, \quad q \in [s-1],$$

Then, it gives

$$\alpha_{s-1} = \alpha_0 + \tfrac{(s-1)}{2}\log 2,$$

and, consequently, the second assumption reads as

$$\alpha_0 \geq \frac{s-1}{s}\alpha_0 + \tfrac{(s-1)^2}{2s}\log 2 + \frac{3}{4s}\log s + \frac{(s+1)}{2s}\log 2,$$

which is equivalent to

$$\alpha_0 \geq \tfrac{s^2-s+2}{2}\log 2 + \frac{3}{4}\log s > 0.$$

We would like to note that our bounds in this proof are rough and it affects the assumptions on parameters α_q. However, our main goal was to show that there exists a set of windows, which lead to an invertible system (3.122) and it was achieved.

Appendix B

Extra tables

Table B.1: Average relative errors $\|X - Z\|_F / \|X\|_F$ depending on the percentage of truncated singular values during the inversion step of the Block Phase Retrieval algorithm. The bold font highlights the minimum error among the values of q in each column, excluding an adaptive choice of q in the last line.

K	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
$q = 0.00$	414.474	138.269	41.293	12.950	3.842	1.333	0.443	0.133
$q = 0.005$	244.099	79.466	24.888	7.500	2.204	0.765	0.252	0.088
$q = 0.01$	168.320	53.597	16.889	5.154	1.511	0.528	0.173	**0.074**
$q = 0.02$	109.446	35.171	11.077	3.385	1.008	0.356	**0.138**	0.088
$q = 0.03$	87.520	28.060	8.820	2.695	0.803	0.295	**0.138**	0.109
$q = 0.04$	75.099	24.069	7.562	2.316	0.695	0.264	0.140	0.118
$q = 0.05$	66.736	21.487	6.721	2.059	0.627	0.254	0.158	0.146
$q = 0.06$	61.108	19.690	6.173	1.887	0.581	0.252	0.170	0.160
$q = 0.07$	57.088	18.381	5.766	1.763	0.547	**0.248**	0.178	0.169
$q = 0.08$	53.441	17.247	5.413	1.654	0.521	0.253	0.191	0.185
$q = 0.09$	50.515	16.336	5.129	1.563	0.503	0.259	0.204	0.201
$q = 0.10$	48.040	15.530	4.885	1.491	0.485	0.260	0.210	0.208
$q = 0.20$	30.575	9.891	3.122	0.995	**0.421**	0.335	0.323	0.324
$q = 0.30$	20.239	6.547	2.104	0.746	0.440	0.414	0.411	0.411
$q = 0.40$	13.962	4.538	1.499	0.638	0.479	0.476	0.470	0.474
$q = 0.50$	9.767	3.196	1.131	**0.617**	0.538	0.544	0.540	0.543
$q = 0.60$	6.712	2.234	0.910	0.629	0.589	0.600	0.596	0.601
$q = 0.70$	4.301	1.512	0.791	0.663	0.643	0.656	0.652	0.656
$q = 0.80$	2.109	0.972	**0.769**	0.735	0.725	0.740	0.733	0.737
$q = 0.90$	**0.873**	**0.845**	0.848	0.846	0.841	0.844	0.845	0.846
via (6.2)	0.880	0.860	0.763	0.622	0.422	0.252	0.138	0.073

Table B.2: Average relative errors $\| |x| - v \|_2 / \|x\|_2$ depending on the percentage parameter p used for the selection of the width γ for Block Magnitude Estimation. Highlighted are the minimums for each column.

K	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
$p = 0.1$	**0.165**	**0.171**	0.171	0.140	**0.100**	**0.070**	**0.033**	**0.019**
$p = 0.2$	0.179	0.184	0.171	0.140	**0.100**	**0.070**	**0.033**	**0.019**
$p = 0.3$	0.179	0.184	**0.151**	0.140	**0.100**	**0.070**	**0.033**	**0.019**
$p = 0.4$	0.179	0.184	**0.151**	0.140	**0.100**	**0.070**	**0.033**	**0.019**
$p = 0.5$	0.179	0.184	**0.151**	0.140	**0.100**	**0.070**	**0.033**	**0.019**
$p = 0.6$	0.179	0.184	**0.151**	**0.134**	**0.100**	**0.070**	**0.033**	**0.019**
$p = 0.7$	0.179	0.184	**0.151**	**0.134**	**0.100**	**0.070**	**0.033**	**0.019**
$p = 0.8$	0.179	0.184	**0.151**	**0.134**	0.123	**0.070**	**0.033**	**0.019**
$p = 0.9$	0.179	0.184	**0.151**	**0.134**	0.123	**0.070**	**0.033**	**0.019**
$p = 1.0$	0.179	0.184	**0.151**	**0.134**	0.123	0.218	0.173	0.101

Table B.3: Average relative errors $\| |x| - v \|_2 / \|x\|_2$ depending on the choice of percentage parameter p for the selection of diagonals used for Log Magnitude Estimation. Highlighted are the minimum values along the columns.

K	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
$p = 10\%$	0.291	0.434	0.306	0.211	0.121	0.083	**0.041**	0.030
$p = 20\%$	**0.179**	**0.184**	0.329	0.197	0.121	0.083	**0.041**	0.030
$p = 30\%$	**0.179**	**0.184**	0.346	0.245	0.131	0.083	**0.041**	0.030
$p = 40\%$	**0.179**	**0.184**	**0.151**	0.222	0.134	0.083	**0.041**	0.030
$p = 50\%$	**0.179**	**0.184**	**0.151**	0.222	0.134	0.083	**0.041**	0.030
$p = 60\%$	**0.179**	**0.184**	**0.151**	0.225	0.132	0.083	**0.041**	0.030
$p = 70\%$	**0.179**	**0.184**	**0.151**	0.235	0.124	0.082	**0.041**	0.030
$p = 80\%$	**0.179**	**0.184**	**0.151**	**0.134**	**0.115**	0.082	**0.041**	0.030
$p = 90\%$	**0.179**	**0.184**	**0.151**	**0.134**	0.132	**0.081**	**0.041**	0.030
$p = 100\%$	**0.179**	**0.184**	**0.151**	**0.134**	0.123	0.296	0.048	**0.027**

Table B.4: Average relative errors $\text{dist}(x, u)/\sqrt{d}$ depending on the choice of percentage parameter p for the selection of diagonals used for the angular synchronization. Highlighted are the minimum value along the columns.

K	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
Unweighted								
$p = 10\%$	1.255	1.287	0.949	0.389	**0.159**	0.104	**0.051**	**0.029**
$p = 20\%$	1.229	1.201	**0.948**	**0.379**	**0.159**	0.104	**0.051**	**0.029**
$p = 30\%$	**1.224**	1.202	0.955	0.399	**0.159**	0.104	**0.051**	**0.029**
$p = 40\%$	1.229	1.207	1.199	0.396	**0.159**	0.104	**0.051**	**0.029**
$p = 50\%$	1.228	1.204	1.196	0.396	**0.159**	0.104	**0.051**	**0.029**
$p = 60\%$	1.230	**1.196**	1.203	0.419	0.161	0.104	**0.051**	**0.029**
$p = 70\%$	1.234	1.201	1.201	0.486	**0.159**	0.104	**0.051**	**0.029**
$p = 80\%$	1.232	1.219	1.196	1.128	**0.159**	**0.103**	**0.051**	**0.029**
$p = 90\%$	1.225	1.209	1.202	1.109	0.228	0.111	**0.051**	**0.029**
$p = 100\%$	1.227	1.208	1.203	1.098	1.137	0.439	0.074	0.032
Amplitude Weights								
$p = 10\%$	**1.171**	1.218	**0.845**	0.324	**0.137**	0.094	**0.044**	0.027
$p = 20\%$	1.195	**1.198**	0.849	**0.322**	**0.137**	0.094	**0.044**	0.027
$p = 30\%$	1.195	**1.198**	0.927	0.331	**0.137**	0.094	**0.044**	0.027
$p = 40\%$	1.195	**1.198**	1.160	0.349	**0.137**	0.094	**0.044**	0.027
$p = 50\%$	1.195	**1.198**	1.160	0.349	**0.137**	0.094	**0.044**	0.027
$p = 60\%$	1.195	**1.198**	1.160	0.368	0.139	0.094	**0.044**	0.027
$p = 70\%$	1.195	**1.198**	1.160	0.433	**0.137**	0.094	**0.044**	0.027
$p = 80\%$	1.195	**1.198**	1.160	1.089	0.139	**0.093**	**0.044**	0.027
$p = 90\%$	1.195	**1.198**	1.160	1.089	0.199	0.099	**0.044**	0.027
$p = 100\%$	1.195	**1.198**	1.160	1.089	1.146	0.346	0.059	**0.024**
Sq. Amplitude Weights								
$p = 10\%$	**1.185**	**1.171**	**0.859**	0.364	0.170	0.115	**0.047**	**0.033**
$p = 20\%$	1.199	1.183	**0.859**	**0.362**	0.170	0.115	**0.047**	**0.033**
$p = 30\%$	1.199	1.183	0.919	0.391	0.170	0.115	**0.047**	**0.033**
$p = 40\%$	1.198	1.183	1.160	0.411	0.169	0.115	**0.047**	**0.033**
$p = 50\%$	1.199	1.183	1.160	0.411	0.169	0.115	**0.047**	**0.033**
$p = 60\%$	1.198	1.183	1.160	0.417	0.163	0.115	**0.047**	**0.033**
$p = 70\%$	1.197	1.183	1.160	0.449	0.158	0.115	**0.047**	**0.033**
$p = 80\%$	1.198	1.183	1.160	1.082	**0.152**	0.113	**0.047**	**0.033**
$p = 90\%$	1.198	1.183	1.160	1.082	0.197	**0.110**	**0.047**	**0.033**
$p = 100\%$	1.197	1.183	1.160	1.082	1.139	0.424	0.060	0.072

Table B.5: The average relative errors $\|X - Z\|_F / \|X\|_F$ depending on the percentage of truncated singular values during the inversion step of the Block Phase Retrieval algorithm with the shift size $s = 4$. The bold font highlights the minimum error among the values of q in each column, excluding an adaptive choice of q in the last line.

K	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
$q = 0.00$	95.866	30.821	10.663	5.063	4.314	4.170	3.813	4.056
$q = 0.01$	74.596	23.950	8.315	3.486	2.516	2.289	2.268	2.441
$q = 0.01$	66.813	21.298	7.177	3.032	2.214	2.049	2.001	2.148
$q = 0.02$	62.727	20.209	6.817	2.871	2.121	1.963	1.920	2.049
$q = 0.03$	59.531	19.301	6.490	2.712	1.981	1.835	1.825	1.951
$q = 0.04$	23.344	7.452	2.845	1.741	1.608	1.515	1.526	1.682
$q = 0.05$	23.344	7.452	2.845	1.741	1.608	1.515	1.526	1.682
$q = 0.06$	23.344	7.452	2.845	1.741	1.608	1.515	1.526	1.682
$q = 0.07$	23.344	7.452	2.845	1.741	1.608	1.515	1.526	1.682
$q = 0.08$	23.344	7.452	2.845	1.741	1.608	1.515	1.526	1.682
$q = 0.09$	23.344	7.452	2.845	1.741	1.608	1.515	1.526	1.682
$q = 0.10$	23.344	7.452	2.845	1.741	1.608	1.515	1.526	1.682
$q = 0.20$	13.671	4.349	1.706	1.059	0.990	0.922	0.936	0.975
$q = 0.30$	7.210	2.405	1.112	0.862	0.840	0.801	0.808	0.834
$q = 0.40$	3.548	1.344	0.844	**0.771**	**0.764**	**0.747**	**0.744**	**0.764**
$q = 0.50$	2.005	1.000	**0.843**	0.824	0.823	0.810	0.809	0.826
$q = 0.60$	1.150	0.907	0.887	0.880	0.885	0.871	0.875	0.885
$q = 0.70$	0.919	**0.880**	0.886	0.880	0.886	0.876	0.881	0.888
$q = 0.80$	**0.896**	0.886	0.892	0.887	0.892	0.887	0.889	0.895
$q = 0.90$	0.941	0.940	0.944	0.940	0.944	0.940	0.942	0.945
via (6.3)	0.896	0.877	0.869	0.774	0.764	0.747	0.744	0.764