Technische Universität München
TUM School of Computation, Information and Technology

Development and Analysis of Algorithms for Ptychographical Imaging

Patricia M. Römer

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Caroline Lasser

Prüfende der Dissertation:

1. Prof. Dr. Felix Krahmer
2. Prof. Dr. Palina Salanevich
3. Prof. Dr. Gerlind Plonka-Hoch

Die Dissertation wurde am 17.02.2025 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 07.07.2025 angenommen.

Abstract

This thesis investigates algorithmic approaches for image reconstruction from ptychographic data that are limited by distortions arising in real-world experiments. Ptychography is an imaging technique that has to deal with intensity measurements, caused by detectors that are not able to sense the phase of the signal comprising information on the specimen of interest. This gives rise to solving a phaseless inverse problem, known as phase retrieval, and demands for the use of computational algorithms to recover an image of the measured object. Imaging experiments are prone to measurement corruptions, which motivates us to develop and analyze algorithms that take potential perturbations into account. Specifically, we consider phaseless reconstruction from measurements with Poisson noise and from measurements distorted by a background signal since these two effects are known as the main sources of contamination in ptychographic experiments. For measurements following a Poisson distribution, we contribute theoretical guarantees for a spectral method and for gradient descent algorithms. We design novel approaches to the problem of phase retrieval with a particular emphasis on low-count Poisson measurements, driven by the practical limitations caused by radiation damage. To address ptychographic measurements with background distortion, we introduce algorithms that are capable of restoring lost data and establish a uniqueness of reconstruction guarantee for non-absorbing objects.

Zusammenfassung

Diese Arbeit untersucht algorithmische Ansätze zur Bildrekonstruktion von ptychographischen Daten, welche beeinträchtigt sind durch Störungen, die in Experimenten unter realen Bedingungen auftreten. Ptychographie ist eine Bildgebungsmethode mit Intensitätsmessungen, da die Detektoren die Phase des Signals, das Informationen über die Probe enthält, nicht messen können. Folglich muss ein phasenloses inverses Problem, bekannt unter dem Namen Phasenrekonstruktion, gelöst werden. Das erfordert den Einsatz von Rechenalgorithmen zur Rekonstruktion eines Bildes des gemessenen Objektes. Bildgebungsexperimente sind anfällig für Messverfälschungen, weshalb wir Algorithmen entwickeln und analysieren, welche mögliche Störungen berücksichtigen. Insbesondere betrachten wir das Problem der Phasenrekonstruktion von Messungen mit Poisson-Rauschen und von Messungen, die verzerrt sind durch ein Hintergrundsignal, da diese beiden Effekte als die Hauptstörungsquellen in ptychographischen Experimenten gelten. Für Poisson-Messungen liefern wir eine Rekonstruktionsgarantie für eine Spektralmethode sowie Konvergenzgarantien für Gradientenabstiegsalgorithmen. Wir entwickeln neue Algorithmen für das Problem der Phasenrekonstruktion mit besonderem Schwerpunkt auf Messungen mit niedriger Zählrate, motiviert durch die Einschränkungen, die durch Strahlungsschäden am gemessenen Objekt entstehen. Für ptychographische Messungen mit Hintergrundrauschen führen wir Algorithmen ein, die durch die Störung verlorene Datenpunkte rekonstruieren können, und zeigen, dass nicht-absorbierende Objekte eindeutig rekonstruierbar sind.

Acknowledgments

First and foremost, I would like to thank my supervisor, Felix Krahmer. Thank you for your support and patience throughout the past years. I appreciate the effort you invested in helping me grow both as a researcher and as a person.

I am equally thankful to my supervisor at Helmholtz Munich, Frank Filbir. Thank you for always providing me with a different perspective on things, and especially for encouraging me to explore interdisciplinary research questions, which opened doors to new perspectives and opportunities.

I would also like to acknowledge my appreciation to the rest of the Helmholtz gang, Benedikt, Oleh and Paul. I could learn a lot from all of you, in so many ways. Our Phase Retrieval Journal Club was always a great opportunity for expanding my knowledge and a source of inspiration.

I want to express my deepest gratitude to the M15 group, especially Pascal, Apostolos, Tim, Jona, Jenni, Felipe, Cristina, the Alessandros, and my office buddy Marco. Whether it was bouldering or beer activities, drawing all attention to us during lunch breaks with our bursts of laughter or any of our many social activities and spontaneous dinners - I always enjoyed spending time with you.

And while this list could reveal the lack of great women in mathematics, here it is because two of them deserve special mention:

Hannah, you are one of the reasons I am glad my PhD took a bit longer. Thank you for the constant exchange of thoughts, and for all our not-so-serious conversations. I am beyond grateful to have had a colleague, but more importantly, a friend like you.

Anna, over the last years, you have been the best friend I could have wished for. Thank you for always being there for me. I could always rely on you to turn both good and difficult days into much better ones.

I also want to thank Palina Salanevich and Sjoerd Dirksen for the great collaboration and the insightful discussions we have had. Moreover, my appreciation goes to Knut Müller-Caspary and Max Leo Leidl for providing me with invaluable insights into the physics of the ptychographic problem.

Moreover, I am very grateful to Gerlind Plonka-Hoch and Palina Salanevich for dedicating their time to reviewing this thesis.

And last, but most importantly, thanks to my family, Mama, Papa, Nini & Philip ♡

Contents

1. Introduction

1.1. Motivation

Ptychography is a computational imaging technique, applied in X-ray or electron microscopy to reconstruct small-scale objects such as proteins or viruses from diffraction data. The considered data set comprises multiple diffraction patterns that are obtained by localized illuminations of the specimen, which is shifted in each stage of the experiment to finally cover the full specimen. A detector placed behind the object quantizes the incoming wave that carries information about the object. Essentially, it is counted how many radiation particles arrive per pixel of the detector. An image of the measured object is recovered from the set of diffraction patterns via a computational algorithm. Detectors commonly employed in such imaging experiments cannot capture the phase of the incoming wave, but only measure intensities. Consequently, ptychographic experiments give rise to solving a phase retrieval problem, a quadratic inverse problem with phaseless measurements.

In real-world scenarios, the observed data contains contamination of different sources. The primary causes of distortion are the particle counting procedure, resulting in Poisson noise, and scattering of the illumination beam by components of the experimental setup around the object of interest, effecting additional particle counts. Potential damage of the object through the radiation, moreover, requires to restrict the radiation dose, leading to data with low signal-to-noise ratio which limits the attainable reconstruction quality. To avoid artifacts dominating the reconstructed images, these measurement corruptions have to be taken into account in the computational algorithms.

With this thesis, we contribute to the understanding of how algorithms are limited by faulty measurement and present novel techniques, based on state-of-the-art approaches, that face the perturbed data. With the results discussed in this manuscript, we provide a step towards theoretical foundations for ptychographic imaging under real-world conditions. The fundamental questions to be asked are how measurement distortions can be taken into account in algorithmic designs and what can theoretically be guaranteed for the developed methods.

1.2. Outline of the thesis

We give a short overview of the structure of this thesis.

In Chapter 2, we introduce the problem of ptychography - from an experimental as well as from a mathematical perspective. The ptychographic inverse problem is a special case of the phase retrieval problem, which we will also state in its general formulation.

Experimental limitations, such as experimental noise, radiation damage and other types of measurement distortion, arise in ptychographic imaging experiments. The two main sources of corruption, Poisson noise and background distortion, will be discussed in Chapter 3.

Chapter 4 covers first insights on algorithms that are well-known in the phase retrieval community and will be the framework for the methods developed and analyzed in the sequel. We give an overview of the spectral method, Wirtinger gradient descent algorithms, the PhaseLift approach and the Wigner Distribution Deconvolution method.

Based on the spectral method, we present in Chapter 5 a first global recovery guarantee for phase retrieval in a random measurement setup with Poisson noise that is applicable also in case of very low-count measurements.

Chapter 6 focuses on a measurement scenario with low particle counts. We propose several algorithmic approaches to the Poisson phase retrieval problem and establish a new parameter rule for preprocessing low-count Poisson data. The suggested algorithms are supported by convergence guarantees and tested in extensive numerical experiments.

Chapter 7 considers the extreme scenario of one-bit measurements. We provide a recovery guarantee for the spectral method in the case of Bernoulli data involving a random measurement system in Section 7.1 and design novel algorithmic approaches for one-bit measurements in Section 7.2. Empirically, we demonstrate that these methods can improve upon state-of-the-art algorithms in case of extremely low-count data.

Chapter 8 deals with the distortion of ptychographic data by a background signal. Based on the Wigner Distribution Deconvolution method, we develop reconstruction algorithms that remove the background signal and allow for exact reconstruction. We contribute a uniqueness of reconstruction guarantee for so-called phase objects. The empirical performance of the proposed algorithms is showcased with numerical experiments.

In the Appendix, we provide technical lemmas and their proofs that support the main results and were shifted only to improve the readability of the main part. Furthermore, we report some basic results from the literature required for the analyses presented in this thesis.

1.3. Publications

Parts of the results presented in this thesis have been published in the following journals and conference proceedings:

- A summary of the results presented in Chapter 5 and Section 7.1 with the coauthors Sjoerd Dirksen, Felix Krahmer and Palina Salanevich appears under the title *Recovery guarantees for low-count Poisson and Bernoulli phase retrieval* in the proceedings of the *15th International Conference on Sampling Theory and Applications (SampTA 2025)*; see [36].
- The results of Chapter 6 with the coauthors Benedikt Diederichs and Frank Filbir have been published in *Inverse Problems* under the title *Wirtinger gradient descent methods for low-dose Poisson phase retrieval*; see [33].
- The results of Section 7.2 with the coauthor Felix Krahmer have been published under the title *A one-bit quantization approach for low-dose Poisson phase retrieval* in the proceedings of the *International Workshop on the Theory of Computational Sensing and its Applications to Radar, Multimodal Sensing and Imaging (CoSeRa 2024)*; see [113].
- The results of Chapter 8 with the coauthor Oleh Melnyk have been published under the title *Background Removal for Ptychography via Wigner Distribution Deconvolution* in the *SIAM Journal on Imaging Sciences*; see [87].

1.4. Fundamentals and notation

In this section, we introduce the fundamental definitions and notation used in this thesis.

We work with index sets $[d] := \{1, \dots, d\}$ and $[d]_0 := \{0, \dots, d-1\}$ for $d \in \mathbb{N}$. It is only for convenience of notation that we start indexing at 0 when a Fourier transform is involved and at 1 in all other instances. Note that in Section 2.4, Section 4.4 and Chapter 8 the entries of vectors $x \in \mathbb{C}^d$ are enumerated from x_0 to x_{d-1} and all indices are considered modulo d, but we forgo the notation using $\bmod d$.

Throughout this manuscript, we use the following generic notation:

(i) $x \in \mathbb{K}^d$, with $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, is the ground-truth object that we aim to reconstruct.

(ii) d is the dimension of the ground-truth object.

(iii) m denotes the sample complexity.

(iv) With the letter y, typically in the form $y = (y_i)_{i \in [m]}$, we denote the measurements. In the context of ptychographic measurements, it is more convenient to summarize the measurements in a matrix $Y = (Y_{k,r})_{k,r \in [d]_0}$. The capital letter Y will, hence, be the matrix comprising all ptychographic measurements, but in some instances Y will also denote a random variable. It will be clear from the context which concept is meant.

The $d \times d$ identity matrix is denoted by I_d. We work with the indicator function

$$\mathbb{1}_E := \begin{cases} 1, & \text{if event } E \text{ holds true,} \\ 0, & \text{otherwise.} \end{cases}$$

Definition 1.4.1 (Complex numbers)**.** *Any $z \in \mathbb{C}$ is composed as $z = |z|e^{i\phi}$ with $|z| \in \mathbb{R}_{\geq 0}$ and $\phi \in (0, 2\pi]$. We call $|z|$ the magnitude, and $\phi =: \arg(z)$ the argument of z. For $z \in \mathbb{C} \setminus \{0\}$, $\operatorname{sgn}(z) := \frac{z}{|z|}$ denotes the phase of z. For $z = 0$, it is $\operatorname{sgn}(0) := 0$. For $v \in \mathbb{C}^d$, the operations $|v|$, $\arg(v)$, $\operatorname{sgn}(v)$ are applied entry-wise.*

Definition 1.4.2 (Vector and matrix operations)**.**

(i) The complex conjugate of $v \in \mathbb{C}^d$ is denoted by $\overline{v}$ and the conjugate transpose by v^*, while transposing a vector without conjugation is denoted by v^{T}.

(ii) For two vectors $u, v \in \mathbb{C}^d$, the Hadamard product $u \circ v$ is defined by the entry-wise products $(u \circ v)_t := u_t v_t, \ t \in [d]$.

(iii) The trace of a matrix $A \in \mathbb{C}^{d \times d}$ is $\operatorname{tr}(A) := \sum_{j=1}^d A_{jj}$.

Definition 1.4.3. *For a Hermitian matrix $A \in \mathbb{C}^{d \times d}$, we write $A \succeq 0$ to say that A is positive semi-definite, that is, A satisfies $v^* A v \geq 0$ for all $v \in \mathbb{C}^d$.*

Definition 1.4.4 (Singular value decomposition)**.** *For $A \in \mathbb{C}^{d_1 \times d_2}$, there exist unitary matrices $U \in \mathbb{C}^{d_1 \times d_1}$, $V \in \mathbb{C}^{d_2 \times d_2}$, and uniquely defined non-negative numbers $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min\{d_1, d_2\}} \geq 0$, the singular values of A, such that $A = U \Sigma V^{\mathrm{T}}$ with $\Sigma = diag(\sigma_1, \ldots, \sigma_{\min\{d_1,d_2\}}) \in \mathbb{R}^{d_1 \times d_2}$. Here, we denote by diag(v) the diagonal matrix with the vector v determining the entries on the main diagonal. Furthermore, we use the notation $\sigma_{\min}$ and $\sigma_{\max}$ for the smallest and the largest singular value, respectively.*

Definition 1.4.5 (Matrix and vector norms)**.**

(i) The Euclidean norm of a vector $v \in \mathbb{C}^d$ is $\|v\|_2 := \sqrt{\sum_{j=1}^d |v_j|^2}$.

(ii) The spectral norm of a matrix is defined as $\|A\| := \max_{v \in \mathbb{S}^{d-1}} \|Av\|_2$, where $\mathbb{S}^{d-1} := \left\{v \in \mathbb{C}^d : \|v\|_2 = 1\right\}$ denotes the unit sphere of $\mathbb{C}^d$.

(iii) $\langle A, B \rangle_{HS} := \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} A_{ij} \overline{B}_{ij}$ denotes the Hilbert-Schmidt inner product of matrices $A, B \in \mathbb{C}^{d_1 \times d_2}$, $d_1, d_2 \in \mathbb{N}$. The Hilbert-Schmidt inner product is also known as the Frobenius inner product. It induces the Frobenius norm of a matrix $A \in \mathbb{C}^{d_1 \times d_2}$, which is given as $\|A\|_F := \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |A_{ij}|^2$.

(iv) The nuclear norm of a matrix $A \in \mathbb{C}^{d_1 \times d_2}$ *is defined as* $\|A\|_* := \sum_{j=1}^{\min\{d_1,d_2\}} \sigma_j$.

Definition 1.4.6 (Phaseless distance). *If* $u, v \in \mathbb{C}^d$*, the phaseless distance metric is defined via*

$$dist(u,v) := \min_{\theta \in (0,2\pi]} \left\| u - e^{i\theta} v \right\|_2 .$$

For $u, v \in \mathbb{R}^d$*, we instead work with*

$$dist(u,v) := \min_{\alpha \in \{-1,1\}} \| u - \alpha v \|_2 .$$

Over the whole manuscript, we will measure the quality of a reconstruction $\tilde{x}$ obtained by a numerical algorithm in comparison to the ground-truth solution x via the relative reconstruction error $dist(x, \tilde{x}) / \|x\|_2$.

Definition 1.4.7 (Probability distributions). *A random variable* Y

(i) has Gaussian distribution with mean $\nu \in \mathbb{R}$ *and variance* $\sigma^2 \in \mathbb{R}_{\geq 0}$*, denoted by* $Y \sim \mathcal{N}(\nu, \sigma^2)$*, if it has a continuous distribution with probability density function*

$$f(y; \nu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\nu)^2}{2\sigma^2} \right)$$

for $y \in \mathbb{R}$*. A standard Gaussian random variable has mean zero and variance one. When talking about standard Gaussian vectors, we mean random vectors with independent entries which are standard Gaussian distributed, denoted by* $\mathcal{N}(0, I_d)$*.*

(ii) follows the Poisson distribution with parameter $\lambda \in \mathbb{R}_{\geq 0}$*, denoted by* $Y \sim Poisson(\lambda)$*, if its probability function is*

$$\mathbb{P}(Y = k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

(iii) is Bernoulli distributed with parameter $p \in [0,1]$*, written as* $Y \sim Bernoulli(p)$*, if it takes only the two values* 0 *and* 1 *with success probability* $\mathbb{P}(Y = 1) = p$*.*

(iv) follows the binomial distribution with parameters $n \in \mathbb{N}$ *and* $p \in [0,1]$ *if*

$$\mathbb{P}(Y = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k} =: b(k; n, p),$$

meaning that exactly k *out of* n *independent Bernoulli experiments with probability parameter* p *were successful.*

Definition 1.4.8 (Fourier transform). *The discrete Fourier transform of a vector $v \in \mathbb{C}^d$ and its inverse are denoted by*

$$(Fv)_k := \sum_{t=0}^{d-1} e^{-\frac{2\pi i t k}{d}} v_t, \quad \textit{and} \quad (F^{-1}v)_t := \frac{1}{d} \sum_{k=0}^{d-1} e^{\frac{2\pi i k t}{d}} v_k, \quad k, t \in [d]_0,$$

involving the Fourier matrix $F \in \mathbb{C}^{d \times d}$ with entries $(F)_{t,k} = e^{-\frac{2\pi i t k}{d}}$.

Definition 1.4.9. *The circular shift operator with parameter $r \in \mathbb{Z}$, the modulation operator with parameter $k \in \mathbb{Z}$, and the reflection operator on $\mathbb{C}^d$ are defined as*

$$(S_r v)_t := v_{t-r}, \quad (M_k v)_t := e^{\frac{2\pi i t k}{d}} v_t, \quad (Rv)_t := v_{-t}, \quad t \in [d]_0,$$

respectively.

Definition 1.4.10. *The discrete circular convolution of $u, v \in \mathbb{C}^d$ is defined as*

$$(u * v)_n := \sum_{t=0}^{d-1} u_{n-t}\, v_t, \quad n \in [d]_0.$$

2. Ptychography

Ptychography is a computational imaging technique based on diffraction patterns. As it does not directly produce an image of the object but only diffraction data, it requires a numerical algorithm that recovers the object.

In this chapter, we start with describing the experimental setup, followed by a mathematical description of the problem. Throughout this thesis, we discuss various reconstruction algorithms for the discrete version of the ptychographic problem. Therefore, we introduce a discretization and a generalization of the problem. In addition, we provide a short summary of fundamental properties of the phaseless inverse problem regarding the uniqueness and stability of reconstruction.

2.1. Experimental setup

A ptychographic experiment collects multiple diffraction patterns, each of them comprising information on different but overlapping small parts of the object of interest.

We begin with a description of the experimental setup to record one such diffraction pattern, which includes the interplay of three components: a radiation source, the specimen of interest, and a detector. A diagram of the experimental setup is illustrated in Figure 2.1.

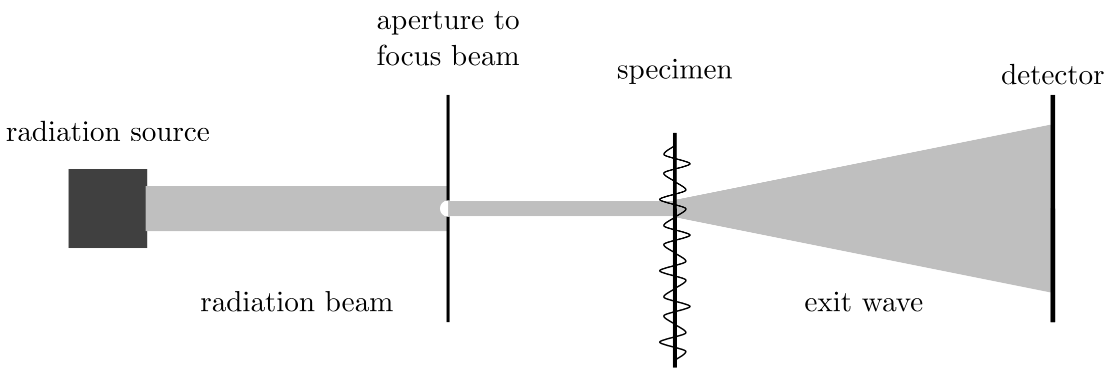


Figure 2.1.: Schematic of the setup of a ptychographic imaging experiment.

The radiation source generates a beam that is focused by a pinhole onto a small area of the object which shall be imaged. The resulting localized wave is typically referred to as probe.

When the beam passes through the object, it encodes information about the object. The properties of main interest are the absorption and the scattering of the beam particles caused by the specimen. These two quantities characterize the wave exiting the object.

Subsequently, this wave propagates towards a detector placed behind the object, where the intensity of the incoming wave is captured. Typically, the measurements are taken by pixel array detectors, for example charge-coupled device (CCD) cameras, which count how many radiation particles arrive in each detector bin. The observed data set is commonly called a diffraction pattern as it comprises information on the diffraction of the radiation by the object.

The field strength observed at the detector corresponds to the Fourier transform of the exit wave if the detector is in the far-field region. Far-field means that the distance z of the detector plane to the object plane satisfies the Fraunhofer condition $z > 2D^2/\lambda$, where D denotes the diameter of the aperture and λ is the radiation wavelength [52].

The measurements are only intensity measurements as the phase information of the incoming wave is lost due to the high-frequency oscillation of the radiation type used in high-resolution experiments that cannot be measured by any detector. Consequently, the phase has to be retrieved from the intensity data.

In a ptychographic experiment, the object is then displaced, and the next part is scanned through the localized illumination. The two adjacent illuminated regions of the object have to overlap such that multiple measurements contain information on the same area of the specimen. The resulting redundancy in the data ensures that the phase problem can be solved, which requires a sufficiently overdetermined system. Empirically, an overlap of approximately 60% between adjacent scan positions was found to yield satisfactory reconstructions [18].

This experiment is repeated with exposures for further shifts of the specimen until measurements were taken for all parts. Finally, an image of the specimen is numerically reconstructed from the diffraction data via a computational algorithm.

Ptychography is applied by the X-ray microscopy community [98], for example at the PETRA III beamline at DESY [118], as well as in electron microscopy [92]. It achieves high-resolution images at nanometer scale of biological specimen [50] or other nanoscale materials [119].

One experimental benefit of ptychography is that it obviates the use of expensive high-quality lenses to obtain high-resolution images. Furthermore, ptychography does not rely on exact knowledge of the optical setup including the probe function. With sufficient redundancy obtained by measuring overlapping parts of the object, it is possible to retrieve both the complex object transmission function and the complex illuminating wave field simultaneously from a set of diffraction patterns. The respective inverse problem is known as blind ptychography.

Alongside this, however, the ptychographic approach builds on a well-performing computational reconstruction algorithm. To advance the study of this crucial aspect, this thesis investigates different types of computational algorithms and provides newly developed variants and extensions of well-known methods, accompanied with theoretical guarantees.

Notes and references. *The expression 'ptychography' was proposed by Reiner Hegerl and Walter Hoppe in [56]. It is derived from the Greek word* $\pi\tau\acute{\upsilon}\xi\eta$ *(in English: fold) due to the convolution (in German: Faltung (in English: folding)) of Fourier transforms of the illumination function and of the object transmission function in Fourier space. Details about this mathematical interpretation will be explained later in Section 4.4.*

In this thesis, we only consider far-field ptychography. A somewhat distinct experimental approach following the same idea of creating redundancy is near-field ptychography, where the illumination is less localized so that illuminated areas are larger and less scans are produced. Moreover, the detector is placed closer to the object than in usual ptychographic experiments, resulting in a Fresnel diffraction problem as opposed to Fourier diffraction. Studies on different aspects of this technique were published in [30, 61, 136].

Furthermore, we focus on the case that the properties of the radiation beam are known, so that only the object needs to be recovered from the measurements. The 'blind' case when also the illumination function needs to be reconstructed from the ptychographic data has been investigated for example in [11, 23, 41, 44]. However, in many applications the illumination function can be formulated as a parametric model [34] which can be retrieved beforehand and then introduced in the computational reconstruction algorithm as a known component.

A thorough introduction to ptychography can be found in [107, 110].

2.2. The ptychographic problem in mathematical terms

In this section, we formulate the mathematical model of a ptychographic experiment. In the following, we focus our consideration to the one-dimensional problem.

The object is described by its object transmission function $f : \mathbb{R} \to \mathbb{C}$. Since f models a physical object, it is reasonable to assume that the physically relevant information of the object is contained in a compact set and that the function characterizing the object is continuous. Without loss of generality, we assume that $f \in L_2(\mathbb{R}) \cap C_c(\mathbb{R})$ and $\operatorname{supp}(f) = [0, 1]$. The complex object transmission function f represents two physical properties of the object. Its magnitude quantifies the percentage of radiation that is not absorbed by the object. A zero magnitude corresponds to the case that no radiation passes through the object. On the other hand, the value is set to one (after normalization) if, in the respective part of the object, no radiation is absorbed. The phase of the object transmission function amounts to the scattering of the radiation by the object. In practice, many objects are so thin that they only scatter the radiation and do not absorb any of it. The transmission function of these objects has then a constant magnitude of value one. Such non-absorbing objects are known as phase objects [55, 115].

If the illumination is localized by a circular pinhole that is placed sufficiently far from the object, the window function can be described by an Airy function with negligible tails, which is commonly approximated by a rapidly decaying Gaussian function. Therefore, it is reasonable to model the probe by a continuous, compactly supported window function $g : \mathbb{R} \to \mathbb{C}$, that is, $g \in L_2(\mathbb{R}) \cap C_c(\mathbb{R})$. Without loss of generality, we presume that $\operatorname{supp}(g) = [0, b]$ with $b < 1$.

The exit wave, resulting from the radiation passing through the object, is represented as the pointwise product of the illumination function and the object transmission function,

$$f(y)g(y-s), \quad y \in \mathbb{R},$$

for any displacement $s \in \mathbb{R}$. This multiplication approximation is accurate for sufficiently thin objects [76, 109].

After propagation to the detector placed in the far-field, the observed field strength equals the Fourier transform of the exit wave,

$$V_g f(s, \xi) = \int_{\mathbb{R}} f(y)g(y-s)e^{-2\pi i \xi \cdot y} dy, \tag{2.1}$$

with frequency variable ξ and shift s. This is commonly referred to as the short-time Fourier transform (STFT) of the object transmission function f with respect to the window g.

Since the detector is only capable to measure the intensities of the incoming waves, the recorded data is given by

$$I(s, \xi) = \left| \int_{\mathbb{R}} f(y)g(y-s)e^{-2\pi i \xi \cdot y} dy \right|^2, \tag{2.2}$$

also known as the spectrogram of f with respect to g.

The task is then to recover the object transmission function f from $|V_g f|^2$, assuming that g is known.

We want to approach this problem from a computational perspective. That means, we aim to reconstruct (a discretization of) the object from samples of the spectrogram, $I(s_r, \xi_k)$, $r \in [R]_0$, $k \in [K]_0$. Note that sampling of $I(s, \cdot)$ for a fixed $s \in \mathbb{R}$ is indeed possible since $V_g f \in C_0(\mathbb{R})$ due to $f, g \in L_2(\mathbb{R}) \cap C_c(\mathbb{R})$. Throughout this thesis, we will study a discrete version of the phaseless inverse problem as we focus on discrete reconstruction algorithms, motivated by real-world imaging applications.

2.3. The discrete phase retrieval problem

Reconstruction from intensity measurements requires to solve an inverse problem with phaseless data, known as 'phase retrieval'. Before we consider the ptychographic problem in its discrete form, we introduce a general formulation of the phase retrieval problem.

In its mathematically abstract form, phase retrieval means the inverse problem of recovering the vector $x \in \mathbb{K}^d$ from phaseless measurements

$$y_i = |\langle a_i, x\rangle|^2, \quad i \in [m],$$

given a frame $\{a_i\}_{i\in[m]} \subset \mathbb{K}^d$. Here, $\mathbb{K}$ is chosen to be $\mathbb{C}$ or $\mathbb{R}$, specified in the respective parts of the manuscript.

A selection of exemplary frames studied in the literature is given in the following.

Example 2.3.1. *(i) A popular frame considered in the phase retrieval community is the Gaussian frame. This means that the vectors a_i, $i \in [m]$, are drawn independently at random from a standard Gaussian distribution, either $a_i \sim \mathcal{N}(0, I_d)$ in the real-valued case or $a_i \sim \mathcal{N}(0, \frac{1}{2}I_d) + i\mathcal{N}(0, \frac{1}{2}I_d)$ in the complex case. While this measurement setup is not of great real-world usefulness, it allows for theoretical analyses and is often considered to provide first guarantees for key research questions. In line with this are our contributions in Chapter 5 and Section 7.1.*

(ii) Another frame of particular interest is the time-frequency frame, also known as Gabor frame. It involves vectors $M_k S_r w$, $(k, r) \in \Lambda \subset [d]_0 \times [d]_0$, that is, time-frequency shifts of a window $w \in \mathbb{C}^d$. We will revisit this example in more detail in the next section, being of specific interest to us and in real-world applications as it corresponds to a discrete model of ptychography.

(iii) A third example that received attention in recent years is phase retrieval with coded diffraction patterns, this means that the measurements originate from frame elements $f_k^* D_r$, $k \in [d]_0$, $r \in [R]$, for some $R \in \mathbb{N}$, where the vectors f_k^* denote the rows of the Fourier matrix F and D_r are diagonal matrices with random entries, together modeling a random masked Fourier transform. It can be understood as an intermediate solution between the fully random Gaussian frame and the structured time-frequency frame. This measurement setup will not be further considered in this thesis.

One of the fundamental questions of phase retrieval concerns the number of frame elements required to allow for unique reconstruction.

First, note that

$$|\langle a_i, \alpha x\rangle|^2 = |\langle a_i, x\rangle|^2$$

for all $\alpha \in \mathbb{T} := \{\beta \in \mathbb{K} : |\beta| = 1\}$. This means that phaseless measurements are the same for objects which differ only up to a global phase factor. Consequently, unique recovery of the ground-truth is always only possible up to a global phase.

We define the equivalence relation

$$x \sim z \Leftrightarrow x = \alpha z \text{ for some } \alpha \in \mathbb{T},$$

and consider a solution to a phase retrieval problem as unique if it is unique up to this global phase equivalence. This also motivates to consider reconstruction errors always in terms of the metric $\min_{\alpha\in\mathbb{T}} \|x - \alpha\hat{x}\|_2$ for $x, \hat{x} \in \mathbb{K}^d$.

To study uniqueness beyond the global phase ambiguity, it has to be investigated under which conditions one can guarantee injectivity of the measurement operator

$$\mathcal{M} : \mathbb{K}^d/\mathbb{T} \to \mathbb{R}^m_{\geq 0},\ x \mapsto \left(|\langle a_i, x\rangle|^2\right)_{i\in[m]},$$

restricting the problem to the set of equivalence classes

$$\mathbb{K}^d/\mathbb{T} = \left\{\{\alpha x : \alpha \in \mathbb{T}\} : x \in \mathbb{K}^d\right\}.$$

In [5] it was shown that for $\mathbb{K} = \mathbb{R}$ and a generic frame $\{a_i\}_{i\in[m]}$, an oversampling with $m \geq 2d - 1$ is sufficient for unique reconstruction. For $\mathbb{K} = \mathbb{C}$ and the same assumption on the measurement system, it was proven in [31] that $m \geq 4d - 4$ measurements are sufficient to guarantee injectivity of $\mathcal{M}$. Here, a frame $\{a_i\}_{i\in[m]}$ is called generic if it is an element of an open dense subset, with respect to the Zariski topology, of the set of all frames with m elements in $\mathbb{K}^d$.

In real-world applications, we have typically only access to measurements

$$y_i = |\langle x, a_i\rangle|^2 + n_i, \quad i \in [m],$$

with unknown perturbation $(n_i)_{i\in[m]}$. This thesis explores phase retrieval under different noise models. The concrete models under consideration will be made more precise in Chapter 3.

Especially in regard to measurements with noise corruption, it is important to investigate whether a stable reconstruction is possible. This means asking the question how much $\hat{x} \in \mathbb{K}^d$ can differ from $x \in \mathbb{K}^d$ if $\mathcal{M}(x)$ and $\mathcal{M}(\hat{x})$ are close. More precisely, $\mathcal{M}$ is called stable if there exists a constant $C > 0$ so that for every $x, \hat{x} \in \mathbb{K}^d$ it holds that

$$\|\mathcal{M}(x) - \mathcal{M}(\hat{x})\|_1 \geq C \min_{\alpha\in\mathbb{T}} \|x - \alpha\hat{x}\|_2.$$

In comparison to the injectivity condition, this is a quantitative statement, making it a stronger property. Furthermore, stability implies uniqueness since it ensures that $\hat{x} \in \{\alpha x : \alpha \in \mathbb{T}\}$ if $\mathcal{M}(x) = \mathcal{M}(\hat{x})$. Vice versa, [54] prove that the finite-dimensional discrete phase retrieval problem is always stable if unique reconstruction is guaranteed.

The stability of reconstruction for phase retrieval with different frames has been intensively studied in the literature; see [6, 39, 114]. From an algorithmic point of view, it was shown that optimization methods as Wirtinger flow [19] (discussed later in Section 4.2) or a PhaseLift approach [20, 21] (introduced later in Section 4.3) provide stable recovery under suitable assumptions. All these results are based on random measurement systems.

2.4. The discrete ptychographic problem

While we will work with the general problem formulation introduced in the preceding section across large parts of the following chapters, some contributions of this thesis, especially the results in Chapter 8, are specifically dedicated to ptychographic data and hence the phase retrieval problem with discrete short-time Fourier transform measurements.

Let us translate the ptychographic problem (2.2) to its discrete analog. We start by approximating the integral in the short-time Fourier transform (2.1) by a trapezoidal rule on a grid with equidistant points. Recall that we assume $\operatorname{supp}(g) \subset \operatorname{supp}(f) = [0,1]$, and further $f, g \in C(\mathbb{R})$. Fix $\xi, s \in \mathbb{R}$ and let $d \in \mathbb{N}$. If we have given samples of f and $g(\cdot - s)$ on the uniform grid $\left\{\frac{t}{d} : t \in [d]_0\right\}$, we obtain

$$\begin{aligned} V_g f(s,\xi) &\approx \frac{1}{2d} \sum_{t=0}^{d-1} \left[f\left(\tfrac{t}{d}\right) g\left(\tfrac{t}{d} - s\right) e^{-2\pi i \xi \cdot \frac{t}{d}} + f\left(\tfrac{t+1}{d}\right) g\left(\tfrac{t+1}{d} - s\right) e^{-2\pi i \xi \cdot \frac{t+1}{d}} \right] \\ &= \frac{1}{d} \sum_{t=0}^{d-1} f\left(\tfrac{t}{d}\right) g\left(\tfrac{t}{d} - s\right) e^{-2\pi i \xi \cdot \frac{t}{d}}, \end{aligned}$$

using that $f(0) = f(1) = 0$. Then, we evaluate ξ on the grid $[K]_0$ for $K \in [d]$, to get

$$V_g f(s,k) \approx \frac{1}{d} \sum_{t=0}^{d-1} f\left(\tfrac{t}{d}\right) g\left(\tfrac{t}{d} - s\right) e^{-\frac{2\pi i k t}{d}},$$

and we evaluate s on the grid $\left\{\frac{rL}{d} : r \in [R]_0\right\}$ for an amount of shifts $R \in [d]$ and shift length $L \in [d]_0 \setminus \{0\}$, to find

$$V_g f\left(\tfrac{rL}{d}, k\right) \approx \frac{1}{d} \sum_{t=0}^{d-1} f\left(\tfrac{t}{d}\right) g\left(\tfrac{t-rL}{d}\right) e^{-\frac{2\pi i k t}{d}}.$$

Let $x \in \mathbb{C}^d$ be the vector comprising the samples of the object transmission function f, that is, $x_t = f\left(\frac{t}{d}\right)$ for all $t \in [d]_0$. Furthermore, let $w \in \mathbb{C}^d$ be defined as the discrete localized window with entries $w_t = g\left(\frac{t}{d}\right)$ for all $t \in [d]_0$. It is $\operatorname{supp}(w) = [\delta]_0$ for $\delta < d$ determined by $\delta = \max\left\{t \in [d]_0 : \frac{t}{d} \le b\right\}$. In summary, ignoring the scaling by d, the discrete short-time Fourier transform of $x \in \mathbb{C}^d$ with respect to $w \in \mathbb{C}^d$ reads as

$$\sum_{t=0}^{d-1} e^{-\frac{2\pi i k t}{d}} x_t w_{t-rL} = \langle x, M_k S_{rL} \overline{w} \rangle,$$

with circular shift operator S_r and modulation operator M_k. It is common to work with circular shifts, assuming a periodic extension of the object.

Consequently, the discrete ptychographic problem is given as the phase retrieval problem with equations

$$Y_{k,r} = |\langle x, \varphi_{k,r}\rangle|^2, \quad k \in [K]_0,\ r \in [R]_0,$$

where $\{\varphi_{k,r}\}_{k\in[K]_0, r\in[R]_0}$ with $\varphi_{k,r} := M_k S_{rL}\overline{w}$ denotes the discrete time-frequency frame with window w.

Again, a fundamental question is whether unique recovery from such phaseless measurements is possible. It is well known that if the frame elements in the general phase retrieval problem represent a discrete Fourier transform, more ambiguities appear than the unavoidable phase ambiguity. This involves trivial ambiguities arising from shifts or reflected conjugation, as well as non-trivial ambiguities characterized in [10]. The redundancy introduced by the ptychographic approach can avert these ambiguities, apart from the global phase ambiguity. This means that unique reconstruction from phaseless discrete short-time Fourier transform measurements can be possible, under suitable assumptions on the object x, the window w, and the parameters K, L and R [16, 40, 65]. We postpone the presentation of one of these results to a later section where it is more appropriately addressed; see Theorem 4.4.5.

Furthermore, the authors of [11] showed that $4d - (d-\delta)/\gcd(L,d)$ phaseless STFT measurements can be sufficient to recover the complex ground-truth object uniquely. Here, $\gcd(L,d)$ denotes the greatest common divisor of L and d. For blind ptychography, that means if w is unknown as well, ambiguities were investigated by [12, 41], and unique recovery, up to trivial ambiguities, was shown to be achievable with $4d + 2\delta - (d - \delta)/\gcd(L,d) - 3$ measurements [11]. However, these results for STFT phase retrieval require the window w and the object x to be generic, which is not necessarily satisfied in applications.

Besides these results on uniqueness of reconstruction, stability estimates for deterministic STFT phase retrieval, and thereby also for ptychography, were presented in [1].

We conclude this section with fixing the notation of the discrete ptychographic problem that we will use throughout the rest of this thesis. So far, we considered a shift of the window since this is more common for the definition of the short-time Fourier transform. However, in the experimental design, we describe a displacement of the object. Note that these two problem formulations are equivalent since

$$\left|\sum_{t=0}^{d-1} e^{-\frac{2\pi i k t}{d}} x_t w_{t-rL}\right|^2 = \left|e^{-\frac{2\pi i k r L}{d}} \sum_{t=0}^{d-1} e^{-\frac{2\pi i k t}{d}} x_{t+rL} w_t\right|^2 = \left|\sum_{t=0}^{d-1} e^{-\frac{2\pi i k t}{d}} x_{t+rL} w_t\right|^2,$$

for all $k \in [K]_0, r \in [R]_0$. From here on, we work with the latter model, using shifts of the object. Moreover, the theoretical investigation of the problem presented in Section 4.4 and Chapter 8 requires full ptychographic data, so that we henceforth assume $K = R = d$ and $L = 1$.

In the rest of the manuscript, we adopt the following notation for the discrete short-time Fourier transform of x with respect to the window w,

$$(F\left[S_{-r}x \circ w\right])_k = \sum_{t=0}^{d-1} e^{-\frac{2\pi i t k}{d}} x_{t+r} w_t, \quad k, r \in [d]_0,$$

so that the corresponding discrete ptychographic measurements are given by

$$Y_{k,r} = \left|(F\left[S_{-r}x \circ w\right])_k\right|^2, \quad k, r \in [d]_0.$$

Naturally, real-world ptychographic measurements are perturbed by noise or other sorts of distortions. In Section 3.3, we introduce a specific type of measurement corruption that is relevant in ptychographic experiments. Recovery algorithms for the resulting problem are provided in Chapter 8, and we will show that such distorted measurements still allow for unique reconstruction.

Notes and references. *For ease of notation, we restrict ourselves in all theoretical considerations to the one-dimensional problem. However, all later discussed algorithms can be extended to a 2D version, and parts of the numerical experiments reported in this thesis were conducted in the two-dimensional setting.*

3. Experimental noise, radiation damage and other measurement distortions

The data recorded in an imaging experiment under real-world conditions is commonly contaminated. This can be caused by experimental noise or other types of distortion, or due to a damage of the specimen by the radiation.

Taking the main sources of perturbation known to appear in ptychographic experiments into account, we are faced with measurements

$$Y_{k,r} \sim Poisson\left(|\langle x, M_k S_{rL} \overline{w}\rangle + \beta_{k,r}|^2 + b_k\right) \; + \; n_{k,r}, \quad k, r \in [d]_0,$$

with

(i) potentially unknown window w,

(ii) an unknown background signal $\beta_{k,r},\ k, r \in [d]_0$, caused by coherent scattering,

(iii) an unknown shift-invariant background signal $b_k,\ k \in [d]_0$, caused by incoherent parasitic scattering of the radiation in the experimental setup,

(iv) Poisson noise as a result of the quantizing particle counting procedure,

(v) and additional random read-out noise $n_{k,r},\ k, r \in [d]_0$, due to the analog-to-digital conversion, typically modeled as Gaussian noise.

A more comprehensive list of potential data corruption sources is given in [24], albeit with the indication that other types of perturbation can usually be determined in the measuring process. Hence, the above specified distortions are of key interest. As mentioned earlier, it is fair to assume that the window w is known. Moreover, the read-out noise can often be minimized by adjusting the hardware, and the background due to coherent scattering is negligible in comparison to the underlying ground-truth signal. Consequently, the incoherent background scattering and the Poisson noise are the two main sources of perturbation which we address in the course of this manuscript.

In this chapter, we present and discuss these primary types of measurement distortions that arise in ptychographic experiments, along with their underlying causes. In the remainder of the manuscript, we will concentrate on either one of the perturbation causes at a time to facilitate mathematical analysis of the respective problem setting.

3.1. Poisson noise

In imaging experiments such as electron or X-ray microscopy, the measurements correspond to counts of detected electrons or photons, respectively, which are observed per pixel of the camera. This amounts to a quantization of the incoming radiation wave by the detector.

In this section, we aim to describe the statistical model of particle count data. That means, we introduce the Poisson noise model, which is most suitable for many types of imaging experiments, including the one considered in this manuscript. To derive this model, we follow the discussion in [42] on the Poisson approximation to the binomial distribution. For simplicity, we regard here only the counting procedure for one single detector pixel, and assume the pixel to be a domain with area 1.

We aim to describe the probability that exactly k particle arrivals are counted in the considered pixel. The model will be based on the following two fundamental physical assumptions:

(i) The probability of counting more than one arrival in a domain of small area is small compared to the size of the considered domain.

(ii) The number of particles arriving in one subdomain is stochastically independent from the number of particles arriving in another non-overlapping subdomain.

Let the pixel be fragmented into n subdomains of equal size. Assume that for all these subdomains the probability of at least one particle arriving is the same and equals p_n. Hence, we are interested in whether the subdomain remains empty or not, and consequently model this as a Bernoulli experiment with success probability p_n. Based on the independence assumption (ii) for non-overlapping subdomains, we can then understand the probability that k subdomains are hit by (at least) one particle as given by the Binomial probability mass function

$$b(k; n, p_n) = \binom{n}{k} p_n^k (1 - p_n)^{n-k}.$$

The considered model is of discrete nature and to make it more precise we need to pass from the finite approximation to the limit $n \to \infty$ by refining the partition of the pixel.

The probability that no particles are counted in the entire pixel equals $b(0; n, p_n) = (1 - p_n)^n$. Studying the Taylor expansion

$$\log(b(0; n, p_n)) = n \log(1 - p_n) = n \left(-p_n + \mathcal{O}(p_n^2)\right),$$

we find that the product np_n must have a finite limit, say $np_n \to \lambda$ for $n \to \infty$ with $\lambda \in \mathbb{R}_{\geq 0}$, to ensure that the probability $b(0; n, p_n)$ is finite even in the limit. With

$np_n \approx \lambda$, we approximate

$$b(0; n, p_n) \approx b\left(0; n, \frac{\lambda}{n}\right),$$

and, with

$$\log\left(b\left(0; n, \frac{\lambda}{n}\right)\right) = n \log\left(1 - \frac{\lambda}{n}\right) = -\lambda + \mathcal{O}\left(\frac{\lambda^2}{n}\right),$$

conclude that in the limit $n \to \infty$ we can approximate

$$b\left(0; n, \frac{\lambda}{n}\right) \approx e^{-\lambda}.$$

Starting from this approximation for $k = 0$, and using the approximate relation

$$\frac{b\left(k; n, \frac{\lambda}{n}\right)}{b\left(k-1; n, \frac{\lambda}{n}\right)} = \frac{\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k}}{\binom{n}{k-1}\left(\frac{\lambda}{n}\right)^{k-1}\left(1 - \frac{\lambda}{n}\right)^{n-k+1}} = \frac{(n-k+1)\frac{\lambda}{n}}{k\left(1 - \frac{\lambda}{n}\right)} \approx \frac{\lambda}{k},$$

which holds in the limit $n \to \infty$, we determine via induction that

$$b\left(k; n, \frac{\lambda}{n}\right) \approx \frac{\lambda^k}{k!}e^{-\lambda}.$$

Under assumption (i), we conclude that, in the limit, the number of subdomains with particle counts equals the amount of particles counted in the whole pixel. In summary, we found that the distribution with probability mass function $\frac{\lambda^k}{k!}e^{-\lambda}$ describes the particle count in one pixel, where λ, the limit of np_n for $n \to \infty$, is to be understood as the underlying ground-truth particle count.

More abstractly, we say from now on that a random variable Y follows a Poisson distribution with parameter $\lambda \in \mathbb{R}_{\geq 0}$, denoted by $Y \sim \mathit{Poisson}(\lambda)$, if

$$\mathbb{P}(Y = k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k \in \mathbb{N}.$$

The expected value of a Poisson random variable Y is

$$\mathbb{E}(Y) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda} \cdot \left(0 + \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}\right) = e^{-\lambda} \cdot \lambda \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda.$$

In terms of a particle counting experiment, this means that each ground-truth intensity measurement equals the expected number of particles arriving at the respective detector pixel. A popular approach to recover the parameter λ from Poisson data is the maximum likelihood method. We defer the discussion on this to Section 6.1.

The variance of a Poisson random variable with parameter λ is also equal to λ. This can be obtained via

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \lambda,$$

using that

$$\begin{aligned}\mathbb{E}(Y^2) &= \sum_{k=0}^{\infty} k^2 \cdot \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \cdot \left(0 + \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!}\right) = e^{-\lambda} \cdot \lambda \cdot \sum_{k=0}^{\infty} (k+1) \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \cdot \lambda \cdot \left(\lambda e^{\lambda} + e^{\lambda}\right) = \lambda^2 + \lambda.\end{aligned}$$

Consequently, the variance increases if the distribution mean increases.

The signal-to-noise ratio of a Poisson distributed random variable is equal to

$$\mathrm{SNR} = \frac{\mathbb{E}\left[Y\right]}{\sqrt{\mathbb{V}\left[Y\right]}} = \frac{\lambda}{\sqrt{\lambda}} = \sqrt{\lambda}.$$

Therefore, if λ is sufficiently large, the SNR is also large and the Poisson noise not of severe concern.

When λ is sufficiently large, the Poisson distribution with parameter λ can furthermore be approximated by the Gaussian distribution $\mathcal{N}(\lambda, \lambda)$. This is a result of the central limit theorem, which guarantees that

$$\sqrt{\lambda}\left(\frac{1}{\lambda}\sum_{i=1}^{\lambda} Y_i - 1\right) \text{ converges in distribution to } \mathcal{N}(0,1) \text{ for } \lambda \to \infty$$

if we assume that $Y_i,\ i \in [m]$, are independent and identically distributed random variables with mean and variance equal to 1. Since we can write $Y \sim \mathit{Poisson}(\lambda)$ as a sum of independent and identically distributed random variables $Y_i \sim \mathit{Poisson}(1)$, we find that

$$Y = \sqrt{\lambda}\left[\sqrt{\lambda}\left(\frac{1}{\lambda}\sum_{i=1}^{\lambda} Y_i - 1\right)\right] + \lambda \text{ converges in distribution to } \mathcal{N}(\lambda,\lambda) \text{ for } \lambda \to \infty.$$

Accordingly, in scenarios with high expected particle counts, it is popular to process Poisson distributed data using methods designed for Gaussian data, although often disregarding the signal-dependent variance. Variance stabilizing transformations that address this issue are discussed later in Section 6.2.

In the limit $\lambda \to 0$, the signal-to-noise ratio converges to zero. For $\lambda < 1$, the standard deviation $\sqrt{\lambda}$ even gets comparably larger in relation to the expected value, causing the noise in the corresponding data to dominate significantly. In the next subsection, we explain why the Poisson distribution with small distribution parameter is of particular interest in the context of ptychographic imaging.

How to deal with low-count measurements, that is, Poisson data with small distribution parameters, is the key focus of Chapter 6.

In terms of the phase retrieval problem, we work with the following measurement model:

For all $i \in [m]$, let y_i be a realization of a random variable $Y_i \sim \mathit{Poisson}\left(|\langle a_i, x\rangle|^2\right)$.

For notational convenience, we will later treat the y_i, $i \in [m]$, themselves as random variables instead of realizations of a random variable. If an observation is assumed to be drawn from a Poisson distribution, we also say that it is perturbed by Poisson noise. Phase retrieval with Poisson noise is the key focus of Chapter 5 to 7.

Notes and references. *The Poisson distribution is attributed to Siméon D. Poisson [101]. An alternative derivation of the Poisson model based on different hypotheses describing the physical model of a particle counting detector can be found in [53, Chapter 3.7.2].*

A comprehensive overview of inverse problems involving Poisson measurements is given in [57].

3.2. Dose restriction for damage prevention

In imaging experiments, the radiation can severely damage the specimen [51, 121], limiting the quality of the experimental data. Radiation sensitivity is particularly pronounced for biological objects, especially organic materials like proteins, a factor that complicates the imaging of, for example, viruses. This problem arises both in X-ray microscopy [116] and in electron microscopy [4], limiting the ptychographical approach in these experiments since it entails multiple exposures of the same specimen.

It was discovered that frozen specimens can cope with higher radiation dose before being destroyed [9], explaining the strong interest in methods like cryo-electron microscopy. However, the damage is only diminished and not averted completely.

As an additional measure, the damage is typically mitigated by reducing the dose of radiation, which, on the downside, results in poor resolution images [58]. Using a low dosage of radiation produces low particle counts per pixel, which amounts to Poisson data with small distribution parameter and therefore measurements that are highly dominated by noise.

The low dosage restriction also makes ptychographic experiments especially sensitive to the dwell time of the utilized detector. The longer the measurement at each scanning position takes, the more prone the data set is to artifacts. A hardware solution that improves low-dose imaging using a detector with shorter dwell time was suggested for example by [66].

Furthermore, the data quality can be limited by the dead time of a detector. This is the period of time after one particle is detected at an individual detector pixel during which this pixel is unable to record further particles. Throughout the dead time, the pixel is blind to all arriving particles, resulting in lost particle counts. Hence, we face a trade-off between dwell time and dead time of the detector. If the dead time is so high in comparison to the scanning time of one diffraction pattern so that no more than one particle can be counted per pixel, this corresponds to extremely low-count Poisson data, and the data might rather be mathematically modeled as truncated Poisson data or simply as Bernoulli measurements, interpreting the experiment only in a binary 'count/no count' framework.

In this regard, we are interested in the following measurement model:

Let $y_i,\ i \in [m]$, be realizations of random variables $Y_i \sim$ Bernoulli $\left(1 - e^{-|\langle a_i, x\rangle|^2}\right)$.

The distribution parameter is derived from the fact that the probability of observing a zero in a Poisson experiment equals

$$\mathbb{P}(Y_i = 0) = \frac{\left(|\langle a_i, x\rangle|^2\right)^0 e^{-|\langle a_i, x\rangle|^2}}{0!} = e^{-|\langle a_i, x\rangle|^2},$$

and every count larger or equal to 1 is simply truncated to 1, so that

$$\mathbb{P}(Y_i = 1) = 1 - \mathbb{P}(Y_i = 0) = 1 - e^{-|\langle a_i, x\rangle|^2}.$$

To sum up, ptychographic imaging experiments with high exposure time demand low radiation dosage, and processing low-count Poisson data depends on suitable algorithmic approaches that take these measurement features into account. Algorithmic designs that can deal with moderately low-count data are investigated in this thesis in Chapter 6 and the extreme case with mainly 0 and 1 counts is addressed in Chapter 7.

Notes and references. *First algorithmic attempts to low-count Poisson ptychography were discussed in [77, 93, 94]. Further insights into low-dose ptychography from a rather applied perspective can be found in [72, 74].*

We contribute novel algorithmic approaches for the extreme case of merely 0 *or* 1 *counts; see Section 7.2. To the best of our knowledge, we were the first to provide theoretical guarantees for Bernoulli phase retrieval. These results are presented in Section 7.1.*

3.3. Background scattering

Aside from the counting noise that impedes reconstruction from real ptychographic data, the second main source of measurement distortion in diffraction experiments is parasitic scattering [104]. It stems from scattering of the radiation beam with air molecules, components of the experimental configuration, and the detector itself. These effects cause a random change in the wave, and hence are classified as so-called incoherent scattering. In the ptychographic measurements, this incoherent scattering amounts to additional particle counts observed by the detector together with the intensity Fourier measurements of the scattering signal corresponding to the specimen. The respective offset is commonly referred to as background signal or simply background.

Due to the incoherence of the background scattering, the background signal is presumed to be approximately the same for each diffraction pattern [15, 24, 91]. Consequently, in a ptychographic experiment, the background is modeled to be independent of the shift position of the sample.

In this manuscript, the model of ptychographic measurements distorted by a background signal is given by

$$\tilde{Y}_{k,r} = Y_{k,r} + b_k, \quad k, r \in [d]_0,$$

with unknown shift-invariant background $b_k \in \mathbb{R}$, $k \in [d]_0$, and phaseless short-time Fourier measurements $Y_{k,r} = \left|(F\left[S_{-r}x \circ w\right])_k\right|^2$, $k, r \in [d]_0$.

Due to the shift-invariance of the background, an approximation of it can be obtained by measuring an additional diffraction pattern of an empty beam, meaning that no object is placed in the experimental setup. From this extra diffraction pattern, the intensity of the forward model applied to the multiplication of the probe function with a constant one function (modeling the empty setup) can be subtracted, resulting in a background estimate which can then be removed from all the diffraction patterns. However, this simple subtraction approach is described in the literature to not be efficient enough [130, 133]. Alternatively, practitioners propose experimental designs to suppress the background scattering [105, 133]. Such approaches enhancing the data quality are required if the employed reconstruction algorithm does not take the background into account since the distortion introduces artifacts in the reconstructions and deteriorates the reconstruction quality. From computational perspective, it was suggested to preprocess the distorted data to reduce the background signal in the diffraction data [130], or to incorporate the background as an additional unknown into an optimization approach for solving the phaseless inverse problem [24, 83].

In contrast to these works, we developed a method which entirely removes the background distortion from the measurements. Our approach enabled additionally a uniqueness of reconstruction guarantee for the ptychographic problem with background. These findings are presented in Chapter 8.

Notes and references. *The literature seems divided over the question whether a background should be seen as noise or an additional signal component. Correspondingly, we decided to name it a background distortion rather than background noise.*

An insightful overview of the physics behind the background scattering is given in [104].

4. An overview on algorithmic approaches

In recent years, various types of algorithms for solving the phase retrieval problem in general or the ptychographic problem in specific have been proposed in the literature. In this chapter, we introduce state-of-the-art methods which were the respective foundations of the research projects presented in this thesis. That means, the algorithms discussed in the following in an introductory manner will later be elaborated and analyzed for different measurement scenarios.

4.1. Phase retrieval via a spectral method

As a first step towards recovering an approximation to the ground-truth solution $x \in \mathbb{C}^d$ of the phaseless inverse problem

$$y_i = |\langle a_i, x\rangle|^2, \quad i \in [m], \tag{4.1}$$

the literature proposes to use a spectral method; see Algorithm 1.

Algorithm 1 Spectral method

Input: Measurements $y_i,\ i \in [m]$, and measurement vectors $a_i,\ i \in [m]$.

Consider the matrix

$$S := \frac{1}{m} \sum_{i=1}^{m} y_i a_i a_i^*.$$

Let $\tilde{x}_0$ be the leading eigenvector of S and set

$$x_0 = \frac{\tilde{x}_0}{\|\tilde{x}_0\|_2} \cdot \sqrt{\frac{d \sum_{i=1}^m y_i}{\sum_{i=1}^m \|a_i\|_2^2}}. \tag{4.2}$$

Output: x_0

The spectral method is motivated by the following observation. Assume that the measurement frame consists of independent and identically distributed random vectors $a_i \sim \mathcal{N}(0, I_d)$ and that $y_i,\ i \in [m]$, are given by noise-free measurements (4.1). Then, we find that

$$\mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} y_i a_i a_i^*\right] = 2xx^* + \|x\|_2^2 \cdot I_d,$$

and note that any leading eigenvector of the matrix $2xx^* + \|x\|_2^2 \cdot I_d$ is equal to sx, for some scalar $s \in \mathbb{R} \setminus \{0\}$. Consequently, the leading eigenvector of S can serve as an approximation of the direction of x. The norm approximation via (4.2) is justified by

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \left|\langle a_i, x\rangle\right|^2\right] = \|x\|_2^2 \quad \text{and} \quad \mathbb{E}\left[\|a_i\|_2^2\right] = d.$$

Proofs of these properties are deferred to Appendix A.

Naturally, the sample mean $\frac{1}{m}\sum_{i=1}^{m} y_i a_i a_i^*$ equals $2xx^* + \|x\|_2^2 \cdot I_d$ only in the limit $m \to \infty$, but [89] nevertheless proposed the spectral method as a reasonable recovery strategy by showing that $m = \mathcal{O}(d \log^3(d))$ independent random Gaussian measurements are sufficient to obtain a reconstruction x_0 that is close to the ground-truth x. This was later improved by [19] to $m = \mathcal{O}(d \log(d))$.

Further advancements of the spectral method replace S by

$$\frac{1}{m}\sum_{i=1}^{m} \mathcal{T}(y_i) a_i a_i^*$$

with a preprocessing of the data via a transform $\mathcal{T} : \mathbb{R} \to \mathbb{R}$. This preprocessing can for example be a truncation $\mathcal{T}(y) = y \cdot \mathbb{1}_{y \le \tau}$ for some threshold $\tau > 0$, cutting off dominating terms that yield an output x_0 which is not sufficiently well aligned with x. This is of concern if $\frac{m}{d}$ is not large enough and caused by the heavy tails of $\left|\langle a_i, x\rangle\right|^2 a_i a_i^*$. If the threshold τ is chosen properly, the technique is more robust and allows for theoretical guarantees with a sample complexity of $m = \mathcal{O}(d)$; see [28, 139].

The authors of [28] provide a recovery guarantee for their truncated spectral method also for measurements with Poisson noise, the scenario we are especially interested in. However, they require the assumption $\|x\|_2 \ge \log^{1.5}(m)$, so that the result does not extend to a low-count Poisson setting. We will present this guarantee and take it into perspective when we revisit the spectral method again in Chapter 5 and Section 7.1, where we contribute an analysis of the spectral method (without preprocessing) for phaseless measurements following a Poisson and a Bernoulli distribution, respectively. We will show that we can go beyond the limiting minimal norm condition $\|x\|_2 \ge \log^{1.5}(m)$ in Theorem 5.2.1 at the expense of an increased sample complexity.

Notes and references. *Besides the recovery guarantees for the spectral method proven in [19, 28, 89], an asymptotic characterization of the spectral method with preprocessing $\mathcal{T}$ is provided in [79]. Their results apply to a broad variety of measurement models, including Poisson data. Furthermore, an optimal choice for the preprocessing transform $\mathcal{T}$ in case of Poisson data is given in [81], albeit also in an asymptotic sense. We will investigate the effect of this preprocessing numerically in Section 7.2.5.*

4.2. Phase retrieval via an optimization approach

Typically, the spectral initialization is not seen as a self-sufficient reconstruction and one aims at refining the result by local optimization methods, for example gradient descent algorithms. With some iterations of an iterative search algorithm, one hopes to get closer to a global optimum. If the objective function in the considered optimization problem is non-convex, iterative approaches, on the other hand, heavily rely on a good initialization, which is obtained by the spectral method. In this regard, combining optimization approaches with the spectral method is very popular in the phase retrieval community. In this section, we will give an introductory overview of optimization techniques for phaseless reconstruction, before we will later go in more detail, proposing variants of the problem suitable for Poisson data and establishing convergence guarantees for multiple versions of gradient descent algorithms; see Chapter 6.

A common approach to phase retrieval is to formulate an optimization problem

$$\min_{z} \ \mathcal{L}(z),$$

with a suitable loss function $\mathcal{L} : \mathbb{C}^d \to \mathbb{R}$, and then solve this by any optimization algorithm to determine a good approximation to the ground-truth solution of the phaseless inverse problem.

In general, when studying a phase retrieval problem, the loss function is of the form $\mathcal{L}(z) = \ell \circ \varphi(z)$, with a function $\ell : \mathbb{R}^m \to \mathbb{R}$ and $\varphi : \mathbb{C}^d \to \mathbb{R}^m$, $z \mapsto (|\langle a_i, z\rangle|^2)_{i\in[m]}$. Commonly, the function ℓ is motivated by a maximum likelihood approach adapted to prior knowledge on the distribution of the measurements. Various suitable loss functions are discussed in more detail in Section 6.1 and Section 6.3. Unfortunately, many adequate functions ℓ yield a non-convex composition $\ell \circ \varphi$, leading to a non-convex optimization problem. That means one has to deal with an optimization problem that can have multiple local minima, making it hard to find a global minimum.

In the phase retrieval community, the two mainly used optimization methods are (stochastic) gradient descent algorithms, see for example [19, 86], and the alternating direction method of multipliers [23–25].

In this thesis, more specifically in Chapter 6, we study gradient descent algorithms as in Algorithm 2 for different types of loss functions $\mathcal{L}$.

Note that the notation $\nabla_z \mathcal{L}$ in (4.3) means a gradient based on Wirtinger derivatives. In the context of phase retrieval where real functions in complex variables are considered, it is common to work with such Wirtinger gradients since the definition of Wirtinger derivatives allows to differentiate non-holomorphic functions; see Appendix B for an overview of Wirtinger calculus.

Methods like Algorithm 2 were firstly introduced by Candès, Li and Soltanolkotabi in

Algorithm 2 Wirtinger gradient descent algorithm

Input: Measurements $y_i,\ i \in [m]$, measurement vectors $a_i,\ i \in [m]$, learning rate $\mu_k > 0$, stopping rule K, initialization $z_0 \in \mathbb{C}^d$ (e.g., $z_0 = x_0$ with output x_0 of Algorithm 1).

for $k = 0$ **to** $K - 1$ **do**

$$z_{k+1} = z_k - \mu_k \nabla_z \mathcal{L}(z_k) \tag{4.3}$$

end

Output: z_K

[19] with their so-called 'Wirtinger flow' (WF) algorithm. The WF update is

$$z_{k+1} = z_k - \mu_k \cdot \frac{1}{m} \sum_{i=1}^{m} (|\langle a_i, z \rangle|^2 - y_i) a_i a_i^* z, \quad k \geq 0,$$

which is the gradient update (4.3) corresponding to the loss function

$$\mathcal{L}(z) = \frac{1}{2m} \sum_{i=1}^{m} \left(|\langle a_i, z \rangle|^2 - y_i\right)^2. \tag{4.4}$$

This approach initiated a long list of versions of gradient descent algorithms using Wirtinger derivatives for various types of loss functions. While (4.4) is commonly referred to as intensity-based empirical loss, one can also work with an amplitude-based loss

$$\mathcal{L}(z) = \frac{1}{2m} \sum_{i=1}^{m} (|\langle a_i, z \rangle| - \sqrt{y_i})^2. \tag{4.5}$$

In comparison to (4.4), the amplitude-based loss is of lower order, which comes along with computational advantages. The Wirtinger gradient descent type algorithm corresponding to (4.5) is commonly named 'amplitude flow'.

First versions of Wirtinger gradient descent algorithms for a loss adjusted to Poisson measurements were proposed and analyzed in [28, 69]. We will follow up on this in Chapter 6.

Notes and references. *Among practitioners in the field of ptychography, a version of stochastic gradient descent gained particular popularity. In the community, it is known under the name PIE, short for ptychographic iterative engine [109]. Its convergence was recently studied in [84].*

Gradient descent methods involving the loss function (4.5) *had been intensively investigated over the past years; see for example [49, 80, 131, 132, 139]. A special case of a stochastic version for the amplitude flow method that also generated attention is a phase retrieval variant of the Kaczmarz algorithm [59, 122]. The convergence of this algorithm in case of bounded additive noise had been analyzed in the Master's thesis of the author* *of this thesis; see the published version [112].*

Most of these so far listed papers study the global convergence of the respective algorithms in case of a random measurement frame. Other than that, convergence of amplitude flow to stationary points of (4.5) *for any type of measurement vectors, random or deterministic, was analyzed in [135] and [44]. In line with these works, we provide convergence guarantees for Wirtinger gradient descent algorithms for a Poisson maximum likelihood loss function and a generalization of the amplitude loss in Section 6.4.*

4.3. Phase retrieval via low-rank matrix recovery

A different algorithmic approach to the phase retrieval problem is the PhaseLift technique. We will revisit this algorithm in Section 7.2.2, providing an advanced version of the method in case of extremely low-count Poisson measurements. In the following, we give an introduction to the method.

The key feature of the PhaseLift method is a linearization of the quadratic inverse problem. For this, note that

$$\begin{aligned}
|\langle a_i, x\rangle|^2 &= \langle a_i, x\rangle \overline{\langle a_i, x\rangle} = a_i^* x x^* a_i = \operatorname{tr}(a_i^* x x^* a_i) = \operatorname{tr}(a_i a_i^* x x^*) \\
&= \sum_{j_1=1}^{d} \left(\sum_{j_2=1}^{d} (a_i a_i^*)_{j_1 j_2} (x x^*)_{j_2 j_1} \right) = \sum_{j_1, j_2=1}^{d} (a_i a_i^*)_{j_1 j_2} \overline{(x x^*)}_{j_1 j_2} \\
&= \langle a_i a_i^*, x x^* \rangle_{HS} \,, \qquad (4.6)
\end{aligned}$$

where the fourth equality follows from the cyclic property of the trace which says that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ for any $A \in \mathbb{C}^{d_1 \times d_2}, B \in \mathbb{C}^{d_2 \times d_1}$. With this trick, we can replace the vector recovery problem corresponding to intensities $|\langle a_i, x\rangle|^2$ by a matrix recovery problem corresponding to Hilbert-Schmidt inner products $\langle a_i a_i^*, x x^* \rangle_{HS}$ with positive semi-definite rank-one matrix xx^*. This reinterpretation of the phase retrieval problem was proposed in [21, 22] and is referred to as PhaseLift since the phase retrieval problem is lifted via (4.6) to a higher-dimensional matrix recovery problem. The phase retrieval problem can then be posed as the matrix recovery problem

$$\begin{aligned}
&\text{minimize} && \operatorname{rank}(X) \qquad (4.7)\\
&\text{subject to} && \mathcal{A}(X) = y, \\
& && X \succeq 0,
\end{aligned}$$

where we denote by $y = (y_i)_{i \in [m]}$ the vector containing all measurements, and $\mathcal{A}$ is the linear operator

$$\mathcal{A} :\ \mathbb{C}^{d \times d} \to \mathbb{R}^m, \quad X \mapsto (\langle a_i a_i^*, X \rangle_{HS})_{i \in [m]} \,.$$

Since the non-convex rank minimization problem (4.7) is NP-hard, it is typically surrogated by the convex problem of minimizing the nuclear norm, that is,

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & \mathcal{A}(X) = y, \\ & X \succeq 0. \end{aligned} \tag{4.8}$$

This semi-definite program is understood as a convex relaxation of the rank minimization problem (4.7) since instead of minimizing the amount of non-zero singular values one minimizes the sum of all singular values. It was shown in [20] that, in case of a Gaussian random measurement frame with a sample complexity $\mathcal{O}(d)$, with high probability, solving the convex program (4.8) results in exact recovery of xx^*.

After successfully reconstructing the rank-one matrix xx^*, it remains to recover the vector x. The solution x corresponds to the leading eigenvector of xx^*, normalized with the square root of the leading eigenvalue of xx^*.

When the measurements y are contaminated, solving (4.8) corresponds to fitting the noise. A more suitable optimization problem that takes noise into account is

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & \|\mathcal{A}(X) - y\|_2^2 \leq \varepsilon, \\ & X \succeq 0, \end{aligned} \tag{4.9}$$

with some threshold $\varepsilon \geq 0$ chosen according to prior knowledge on the noise. For example, if the measurements are perturbed by additive noise $n_i \in \mathbb{R}$ so that $y_i = |\langle a_i, x \rangle|^2 + n_i$ for all $i \in [m]$, and the noise is known to have bounded norm $\|n\|_2^2 \leq \varepsilon$, this program has been shown by [21] to yield stable recovery. In case of a low signal-to-noise ratio, however, this approach relies on a good choice of ε to avoid extreme overfitting of the noisy data. This issue will be addressed later in Section 7.2.2.

The solver $\tilde{X}$ of the matrix recovery problem (4.9) is not necessarily of rank 1. Alternatively, it is common to use the best rank-one approximation $\tilde{x}\tilde{x}^*$ with $\tilde{x} = \sqrt{\sigma_1} u_1$, where σ_1 denotes the leading singular value of $\tilde{X}$ and u_1 its first left singular vector, and consider $\tilde{x}$ a suitable approximation to the ground-truth x; see [60].

To conclude, we emphasize again that the PhaseLift approach allows to replace computationally demanding non-convex optimization problems as discussed in Section 4.2 by a convex program. However, as the problem is lifted to a higher-dimensional problem, essentially from dimension d to d^2, the approach gets computationally impractical up to impossible if the object dimension d increases.

4.4. Wigner Distribution Deconvolution

When regarding ptychographic measurements in specific, as opposed to the general phase retrieval problem considered in the previous subsections, a similar linearization trick can be applied, which motivated the proposal of a direct solver for the ptychographic inverse problem by [26, 108]. A discrete version of the same algorithmic approach was independently developed in [106] and in [32, 62, 95]. The method is known as Wigner Distribution Deconvolution (WDD).

Before presenting in Chapter 8 our contribution building upon this approach for ptychographic reconstruction, we present the WDD algorithm in this section to provide the reader with a self-contained introduction. The WDD approach is based on the following property. For the sake of completeness, we provide a proof of the result, also to point out the origin of the method's name.

Theorem 4.4.1. *Let $Y \in \mathbb{R}^{d\times d}$ be the matrix with entries*

$$Y_{k,r} = \left|(F\left[S_{-r}x \circ w\right])_k\right|^2, \quad k, r \in [d]_0. \tag{4.10}$$

Then, for all $j, k \in [d]_0$,

$$\left(F^{-1}YF\right)_{j,k} = (F\left[x \circ S_j\overline{x}\right])_k \cdot \overline{(F\left[\overline{w} \circ S_j w\right])_k}. \tag{4.11}$$

Proof. We follow the proof of [96, Lemma 7]. First, note that

$$\left|(F\left[x \circ S_r w\right])_k\right| = \left|(F\left[S_r\left(S_{-r}x \circ w\right)\right])_k\right| = \left|(M_{-r}F\left[(S_{-r}x \circ w)\right])_k\right| = \left|(F\left[(S_{-r}x \circ w)\right])_k\right|$$

for all $k, r \in [d]_0$, where the second equality is due to Lemma C.1 (i). Using Lemma C.3 (i) and (ii), we obtain

$$\begin{aligned}\left|(F\left[x \circ S_r w\right])_k\right|^2 &= \left(F\left[(x \circ S_r w) * \left(\overline{Rx} \circ \overline{RS_r w}\right)\right]\right)_k \\ &= \left(F\left[(x \circ S_r w) * \left(\overline{Rx} \circ S_{-r}\overline{Rw}\right)\right]\right)_k.\end{aligned}$$

Taking the inverse Fourier transform with respect to k results in

$$\begin{aligned}\left(F^{-1}Y\right)_{j,r} &= \left((x \circ S_r w) * \left(\overline{Rx} \circ S_{-r}\overline{Rw}\right)\right)_j \\ &= \left((x \circ S_j\overline{x}) * \left(Rw \circ S_{-j}\overline{Rw}\right)\right)_r.\end{aligned}$$

The second equality follows from Lemma C.3 (iii). Taking the Fourier transform with respect to r, the convolution theorem, Theorem C.4 (ii), yields

$$\left(F^{-1}YF\right)_{j,k} = (F\left[x \circ S_j\overline{x}\right])_k \cdot \left(F\left[Rw \circ S_{-j}\overline{Rw}\right]\right)_k. \tag{4.12}$$

With

$$Rw \circ S_{-j}\overline{Rw} = Rw \circ \overline{RS_j w} = \overline{R\left(\overline{w} \circ S_j w\right)},$$

and

$$F\left[\overline{R\left(\overline{w} \circ S_j w\right)}\right] = R\overline{F\left[R\left(\overline{w} \circ S_j w\right)\right]} = \overline{F\left[\left(\overline{w} \circ S_j w\right)\right]},$$

using again Lemma C.3 (ii), Lemma C.1 (ii) and (iii), we obtain the statement of the theorem. □

Remark 4.4.2. *The continuous equivalent to $F\left[x \circ S_j\overline{x}\right]$ is known as the Wigner distribution function. This, together with the deconvolution step in* (4.12)*, motivated the name of Wigner Distribution Deconvolution.*

Theorem 4.4.1 separates the object and the window information and, therefore, provides a reconstruction strategy for ptychographic data. In the following, we delve into the algorithmic procedure.

The first step is to recover the Fourier coefficients $\left(F\left[x \circ S_j\overline{x}\right]\right)_k$. This can be performed via the relation (4.11) if the following assumption applies.

Assumption 1. *For $0 < \gamma \leq d$, let the window w satisfy*

$$\left(F\left[\overline{w} \circ S_j w\right]\right)_k \neq 0 \quad \textit{for all } -\gamma < j < \gamma,\ k \in [d]_0.$$

The question raises which windows $w \in \mathbb{C}^d$ satisfy this assumption for γ large enough. We find that caution is advised in ptychographic applications due to the following observations.

Remark 4.4.3. *(i) We assume in the context of ptychography that* $\operatorname{supp}(w) = [\delta]_0$ *with $\delta < d$. If $\delta \leq \frac{d}{2}$, Assumption 1 cannot be satisfied for $\gamma \geq \delta$. However, ptychographic experiments commonly require $\delta \leq \frac{d}{2}$.*

(ii) If d is even and $w_t = \overline{w}_{\delta-1-t}$ for all $t \in [\delta_0]$, [46] observed that Assumption 1 is not satisfied for all $0 < \gamma \leq \delta$. This condition applies for example for Gaussian windows w with $\operatorname{supp}(w) = [\delta]_0$ *and $w_t = \exp\left(-(t - \frac{\delta-1}{2})(2\sigma^2(\delta-1)^2)^{-1}\right)$, $t \in [\delta]_0$, for some $\sigma > 0$, which can be seen as the discretization of window functions typically used in ptychography [120].*

Positive examples that satisfy Assumption 1 are discussed in [16].

Assuming that Assumption 1 is satisfied for some $0 < \gamma \leq d$, we obtain the following recovery formula from Theorem 4.4.1.

Corollary 4.4.4. *If Assumption 1 holds for some $0 < \gamma \leq d$, the Fourier coefficients $\left(F\left[x \circ S_j\overline{x}\right]\right)_k$ for all $-\gamma < j < \gamma$, and all $k \in [d]_0$, can be recovered via*

$$\left(F\left[x \circ S_j\overline{x}\right]\right)_k = \left(F^{-1}YF\right)_{j,k} \ / \ \overline{\left(F\left[\overline{w} \circ S_j w\right]\right)_k}.$$

The next step of the algorithm is to use the Fourier coefficients obtained by Corollary 4.4.4 to recover the products $(x \circ S_j \overline{x})_t,\ t \in [d]_0$, for all $-\gamma < j < \gamma$. The vector $x \circ S_j \overline{x}$ corresponds to the jth lower off-diagonal of the rank-one matrix xx^*, and we will refer to these vectors in this manuscript as 'diagonals'.

If $\gamma < d$, only some diagonals of the matrix xx^* can be recovered. More precisely, if Assumption 1 is satisfied for $\gamma \leq d$, we can reconstruct the matrix $X = (X_{j_1,j_2})_{j_1,j_2 \in [d]_0}$ with entries

$$X_{j_1,j_2} := \begin{cases} (x \circ S_{j_1-j_2}\overline{x})_{j_1} = x_{j_1}\overline{x}_{j_2}, & \text{if } \min\{|j_1-j_2|, d-|j_1-j_2|\} < \gamma, \\ 0, & \text{otherwise.} \end{cases} \tag{4.13}$$

This matrix is symmetric, which means that the lower off-diagonals equal the upper off-diagonals. Consequently, we can neglect the cases $-\gamma < j < 0$.

The last algorithm stage is the recovery of x from the matrix X. From the main diagonal of X, that is, the vector $x \circ S_0\overline{x} = |x|^2$, the magnitudes of x can be found. Using $x \circ S_1\overline{x}$, the first off-diagonal of X, we can furthermore compute the argument differences $\arg(x_t) - \arg(x_{t-1})$ for all $t \in [d]_0$. Clearly, the recovery of x can never circumvent the global phase ambiguity, so that $\arg(x_0)$ can always be chosen arbitrarily. Then, the arguments of $x_t,\ t > 0$, are reconstructed via the recursive argument relations. Note that phase recovery from the first off-diagonal alone is only possible if $\min_{t\in[d]_0} |x_t| > 0$. If this is not satisfied, further off-diagonals have to be taken into account to reconstruct all required argument differences.

A more robust alternative to the phase reconstruction uses an eigenvector-based approach. More precisely, [63, 129] propose to define the matrix of phase differences

$$(\operatorname{sgn}(X))_{j_1,j_2} := \begin{cases} \operatorname{sgn}(x_{j_1})\operatorname{sgn}(\overline{x}_{j_2}), & X_{j_1,j_2} \neq 0, \\ 0, & \text{otherwise,} \end{cases}$$

since the top eigenvector of the matrix $\operatorname{sgn}(X)$ corresponds to the vector of phases of x, that is, $\operatorname{sgn}(x)$. This technique is commonly known as angular synchronization.

We provide a summary of the presented main steps of the recovery algorithm for ptychographic data based on Wigner Distribution Deconvolution in Algorithm 3.

The WDD algorithm can be understood as a constructive proof for the following uniqueness of reconstruction guarantee for the noise-free ptychographic problem.

Theorem 4.4.5 ([16, Theorem 2.2 and 2.4]). *(i) If Assumption 1 holds with $\gamma = d$, any $x \in \mathbb{C}^d$ is uniquely determined by measurements* (4.10).

(ii) Let Assumption 1 hold with $\gamma \geq 2$. Then, any $x \in \mathbb{C}^d$ that satisfies $\min_{t\in[d]_0} |x_t| > 0$ *is uniquely determined by measurements* (4.10).

Theorem 4.4.5 (ii) follows from the fact that, in case of $\min_{t\in[d]_0} |x_t| > 0$, the first off-

Algorithm 3 Wigner Distribution Deconvolution (WDD)

Input: Ptychographic measurements $Y \in \mathbb{R}^{d \times d}$ as in (4.10) with $w \in \mathbb{C}^d$ satisfying $\operatorname{supp}(w) = [\delta]_0$ for $\delta < d$ and Assumption 1 for $0 < \gamma \leq d$.

Step 1: Use Corollary 4.4.4 to compute $(F[x \circ S_j \overline{x}])_k$ for all $j \in [\gamma]_0,\ k \in [d]_0$.

Step 2: Via inverse Fourier transforms, recover $(x \circ S_j \overline{x})_t$ for all $j \in [\gamma]_0,\ t \in [d]_0$, to obtain X with entries (4.13).

Step 3: Compute the magnitudes $|\tilde{x}_t| = \sqrt{X_{t,t}},\ t \in [d]_0$.

Step 4: Compute the top eigenvector $\tilde{z}$ of $\operatorname{sgn}(X)$ and set $\operatorname{sgn}(\tilde{x}) = \operatorname{sgn}(\tilde{z})$.

Output: $\tilde{x} = |\tilde{x}| \cdot \operatorname{sgn}(\tilde{x}) \in \mathbb{C}^d$, with $\tilde{x} \sim x$ if γ is sufficiently large

diagonal is sufficient to recover all argument differences, as explained above. Consequently, the mitigated assumption on the window, Assumption 1 with $\gamma = 2$, is sufficient for the class of non-vanishing objects, this means all objects $x \in \mathbb{C}^d$ with $\min_{t \in [d]_0} |x_t| > 0$, allowing for unique reconstruction even if $\delta \leq \frac{d}{2}$. One of the main contributions of this thesis, presented in Section 8.4, sets this result into perspective with a uniqueness of reconstruction guarantee in case of ptychographic data with background distortion.

If the measurements are perturbed by any bounded additive noise, the WDD algorithm was also shown to satisfy the following recovery guarantee.

Theorem 4.4.6 ([102, Theorem 1])**.** *Let $\delta > 2$ and $d \geq 4\delta$. If Algorithm 3 is applied for corrupted ptychographic measurements $|(F[S_{-r}x \circ w])_k|^2 + N_{k,r},\ k, r \in [d]_0$, with noise $N = (N_{k,r})_{k,r \in [d]_0} \in \mathbb{R}^{d \times d}$, the output $\tilde{x} \in \mathbb{C}^d$ satisfies*

$$dist(x, \tilde{x}) \leq 24 \frac{\|x\|_\infty}{\min_{t \in [d]_0} |x_t|^2} \cdot \frac{d^2}{\delta^{\frac{5}{2}}} \cdot \|N\|_F \, \sigma_{min}^{-1} + d^{\frac{1}{4}} \sqrt{\|N\|_F \, \sigma_{min}^{-1}}\,,$$

where σ_{min} denotes the smallest singular value of the linear operator corresponding to the ptychographic measurements.

Remark 4.4.7. *In [85, p. 74, eq. (3.38)] it was proven that*

$$\sigma_{min} = \sqrt{d} \min_{j \in [\delta]_0, k \in [d]_0} |F[\overline{w} \circ S_j w]_k|.$$

Hence, if Assumption 1 holds with $\gamma = \delta$, the quantity σ_{min} in Theorem 4.4.6 is guaranteed to be nonzero.

Theorem 4.4.6 together with Remark 4.4.7 shows that, for non-vanishing objects, the WDD algorithm is robust to additive noise. However, stable reconstruction requires $\min_{j \in [\delta]_0, k \in [d]_0} |F[\overline{w} \circ S_j w]_k|$ and $\min_{t \in [d]_0} |x_t|^2$ to be sufficiently large.

In this thesis, we advance the uniqueness of reconstruction further for ptychographic measurements distorted by a background signal. This contribution is based on an extension for the WDD algorithm which is equipped to remove the background distortion. The respective results will be presented in Chapter 8.

Notes and references. *The relation of the Wigner Distribution Deconvolution algorithm to the lifting approach discussed in Section 4.3 is more intuitively understood when it is contrasted with the equivalent 'BlockPR' algorithm [62, 63], which lifts the quadratic ptychographic measurements as in* (4.6) *and solves the problem via an inversion of the resulting linear problem. The equivalence of these two methods is analyzed in [85, Theorem 3.6.4].*

Further advancements of the WDD algorithm include more stable approaches for estimating the magnitude [85, 103] and other versions for angular synchronization [43, 75, 102]. Angular synchronization was also used in other approaches for phase retrieval [2, 82, 97].

Variants of the WDD algorithm were developed for near-field ptychography [61] and for live processing of ptychographic reconstruction [7].

So far, we only discussed ptychographic measurements which are created by shifting the object pixel by pixel, that is, with a shift of size $L = 1$. *In real-world experiments, it might be favorable to work with larger shifts. Extensions that allow for* $L > 1$ *can be found in [96] for bandlimited objects and in [85] for piecewise constant objects. The results presented in Chapter 8 require shifts of size 1. Nevertheless, we perform numerical experiments using the algorithmic approach for larger step sizes of [85] as a heuristic. This is discussed in more detail in Section 8.5.*

5. Poisson phase retrieval

In many imaging applications, such as ptychography, the dominant source of noise is caused by the particle counting procedure. We derived in Section 3.1 that this motivates a model with measurements which follow a Poisson distribution. In this chapter, we analyze the spectral method, outlined in Section 4.1, for measurements

$$y_i \sim \mathit{Poisson}\left(|\langle a_i, x\rangle|^2\right), \quad i \in [m].$$

In the following, we work with the phase retrieval problem for $\mathbb{K} = \mathbb{R}$ and with a random Gaussian measurement frame $\{a_i\}_{i\in[m]}$. We derive a recovery guarantee that reveals the trade-off between the radiation dose and the sample complexity.

The results presented in this chapter are based on joint work with Sjoerd Dirksen, Felix Krahmer and Palina Salanevich; see also [36, 37].

5.1. The spectral method for Poisson phase retrieval

As described in Section 3.1, the signal-to-noise ratio of Poisson data highly depends on the size of the expected signal, which in turn is determined by the radiation dose used in the experiment. Driven by this rationale, our goal is to describe the quality of a reconstruction obtained from phaseless Poisson measurements in relation to the dose. While the experimental interpretation would be that the radiation dose is a parameter of the measurement setup, we fix the normalization of the measurement vectors $a_i,\ i \in [m]$, and incorporate the dose in the norm $\|x\|_2^2$ of the object x for notational convenience. That means, we work in this chapter under the assumption that the ground-truth object $x \in \mathbb{R}^d$ is known to be renormalized such that $\|x\|_2^2 = \kappa$ with dose parameter $\kappa > 0$, so that we consider measurements

$$y_i \sim \mathit{Poisson}\left(\kappa \left|\left\langle a_i, \frac{x}{\|x\|_2}\right\rangle\right|^2\right), \quad i \in [m]. \tag{5.1}$$

We aim to analyze the spectral method for Poisson measurements. Let $x_0 \in \mathbb{R}^d$ be the output of Algorithm 1, that means, x_0 is the eigenvector corresponding to the largest eigenvalue of

$$S := \frac{1}{m}\sum_{i=1}^{m} y_i a_i a_i^{\mathrm{T}},$$

with a normalization $\|x_0\|_2^2 = \kappa$.

We evaluate how well x_0 approximates the ground-truth object x in case of phaseless Poisson measurements and obtain the following recovery guarantee. This result is based on two technical lemmas which were deferred to the appendix to improve readability; see Lemma A.10 and Lemma A.12.

Theorem 5.1.1. *Fix $x \in \mathbb{R}^d$ and let $m \geq d$. If $y_i \sim$ Poisson$\left(|\langle a_i, x\rangle|^2\right)$, $i \in [m]$, with independent random vectors $\mathbb{R}^d \ni a_i \sim \mathcal{N}(0, I_d)$, $i \in [m]$, there exist absolute constants $C, c > 0$, such that, with probability at least*

$$1 - 24m \exp\left(-cd\right) - \frac{6}{d} - \frac{18}{m},$$

the vector x_0 satisfies

$$\frac{dist(x, x_0)^2}{\|x\|_2^2} \leq 2 - 2\sqrt{1 - C \cdot \left(1 + \frac{1}{\kappa}\right) \cdot \sqrt{\log(d)}\log(m)\sqrt{\frac{d}{m}}}\ .$$

Proof. We aim to bound the phaseless distance

$$\begin{aligned} dist(x_0, x)^2 &= \min_{\alpha \in \{-1,1\}} \|x_0 - \alpha x\|_2^2 \\ &= \min_{\alpha \in \{-1,1\}} \left(\|x_0\|_2^2 - 2\alpha \langle x_0, x\rangle + \|x\|_2^2\right) \\ &= \|x_0\|_2^2 - 2\left|\langle x_0, x\rangle\right| + \|x\|_2^2 \\ &= 2\kappa - 2\left|\langle x_0, x\rangle\right|. \end{aligned}$$

Thus, we need to establish a lower bound for $|\langle x_0, x\rangle|$. We proceed similarly to the proof of [19, Theorem 3.3], using that x_0 is the eigenvector corresponding to the largest eigenvalue λ_0 of S normalized so that $\|x_0\|_2^2 = \kappa$. We have $\kappa\lambda_0 = x_0^{\mathrm{T}} S x_0$ and hence

$$\begin{aligned} \left|\kappa\lambda_0 - 2\left|\langle x_0, x\rangle\right|^2 - \kappa\|x_0\|_2^2\right| &= \left|x_0^{\mathrm{T}} S x_0 - x_0^{\mathrm{T}}\left(2xx^{\mathrm{T}} + \|x\|_2^2 \cdot I_d\right)x_0\right| \\ &= \left|x_0^{\mathrm{T}}\left(S - \mathbb{E}\left[S\right]\right)x_0\right| \\ &\leq \|x_0\|_2^2 \left\|S - \mathbb{E}\left[S\right]\right\|, \end{aligned}$$

using that

$$\mathbb{E}\left[S\right] = 2xx^{\mathrm{T}} + \|x\|_2^2 \cdot I_d$$

as proven in Lemma A.10.

Setting $\rho := \|S - \mathbb{E}[S]\|$, the inequality above yields

$$|\langle x_0, x\rangle|^2 \geq \frac{1}{2}\left(\kappa\lambda_0 - \kappa^2 - \kappa\rho\right).$$

Furthermore, the largest eigenvalue of S satisfies

$$\begin{aligned}
\lambda_0 &\geq \frac{1}{\|x\|_2^2} \cdot x^{\mathrm{T}} S x = \frac{1}{\|x\|_2^2} \cdot x^{\mathrm{T}}\left(S - \mathbb{E}[S] + \mathbb{E}[S]\right) x \\
&\geq - \sup_{\|z\|_2 = 1} \left| z^{\mathrm{T}}\left(S - \mathbb{E}[S]\right) z\right| + \frac{1}{\|x\|_2^2} \cdot x^{\mathrm{T}} \mathbb{E}[S]\, x \\
&= -\rho + \frac{1}{\|x\|_2^2} \cdot x^{\mathrm{T}} \mathbb{E}[S]\, x = -\rho + 3\kappa,
\end{aligned}$$

so that

$$|\langle x_0, x\rangle|^2 \geq \frac{1}{2}\left(2\kappa^2 - 2\kappa\rho\right).$$

We conclude that

$$\frac{dist(x, x_0)^2}{\kappa} \leq 2 - 2\sqrt{1 - \frac{1}{\kappa} \cdot \rho}\ . \tag{5.2}$$

Lemma A.12 yields that there exist absolute constants $C, c > 0$, such that, with probability at least

$$1 - 24m \exp\left(-cd\right) - \frac{6}{d} - \frac{18}{m},$$

it holds that

$$\rho \leq C\lceil\kappa\rceil \sqrt{\log(d)} \log(m) \sqrt{\tfrac{d}{m}}.$$

Inserting this bound for ρ into (5.2) and using $\lceil\kappa\rceil < 1 + \kappa$ yields the statement. $\square$

From Theorem 5.1.1 we deduce the following conclusion about the required sample complexity.

Corollary 5.1.2. *From the bound in Theorem 5.1.1, we conclude that*

$$\frac{m}{\log(m)^2} = \mathcal{O}\left(d \log(d) \cdot \left(1 + \frac{1}{\kappa}\right)^2\right) \tag{5.3}$$

measurements are required to recover the ground-truth object up to a small constant relative error from measurements (5.1) *by solving Algorithm 1 with appropriate normalization.*

This theoretical guarantee suggests that m may need to be extremely large if the radiation dose is extremely small. Typically, if just a certain amount of measurements can be produced, the spectral method is only used as a first rough approximation, with controlled error, and then adopted as an initialization for an iterative reconstruction algorithm that refines the estimate by converging closer to the ground-truth solution. In this fashion, we analyze optimization algorithms for Poisson phase retrieval in Chapter 6, especially aiming for a method that works well for low-count Poisson measurements.

Notes and references. *A future project is to extend the result presented in this section to* $\mathbb{K} = \mathbb{C}$. *Furthermore, it would be interesting to investigate whether one can also obtain a statement that holds uniformly for all* $x \in \mathbb{K}^d$.

5.2. Advancing beyond known results

Our recovery guarantee in Theorem 5.1.1 is not the first theoretical result for Poisson phase retrieval. Based on a truncated version of the spectral method, [28] proved the following error bound.

Theorem 5.2.1 ([28, Proposition 3])**.** *Fix* $\epsilon > 0$ *and* $x \in \mathbb{R}^d$ *satisfying* $\|x\|_2 \geq \log^{1.5}(m)$. *Let* $y_i \sim$ *Poisson*$(|\langle a_i, x\rangle|^2)$, $i \in [m]$, *with independent random vectors* $a_i \sim \mathcal{N}(0, I_d)$, $i \in [m]$. *If* $m = \mathcal{O}(d)$, *the solution* x_0 *to a version of Algorithm 1 with a truncation rule* $\mathcal{T}(y) = y \cdot \mathbb{1}_{y \leq \tau}$, *for some* $\tau > 0$, *applied to the measurements* y_i, *satisfies*

$$\frac{dist(x, x_0)^2}{\|x\|_2^2} \leq \epsilon$$

with high probability.

In comparison to our result, this only requires $m = \mathcal{O}(d)$, which is achieved by the appropriate truncation. However, the condition $\|x\|_2 \geq \log^{1.5}(m)$ together with a standard Gaussian vector a_i asks for

$$\mathbb{E}\left[y_i\right] = \mathbb{E}_{a_i}\left[\mathbb{E}_{y_i}\left[y_i\right]\right] = \mathbb{E}_{a_i}\left[|\langle a_i, x\rangle|^2\right] = \|x\|_2^2 \geq \log^3(m),$$

where the third equality is due to (A.1). We conclude that this requirement clearly limits the applicability of the result to real-world scenarios, excluding experiments that require a low dose of radiation due to reasons discussed in Section 3.2.

However, [28] mention that this minimal norm condition is crucial if one aims for a sample complexity $m = \mathcal{O}(d)$. In this regard, we found in Theorem 5.1.1 that we can go beyond the limiting dose restriction with a sample complexity only increased by logarithmic factors.

6. Low-dose Poisson phase retrieval

As described in Section 3.2, many real-world experiments require to restrict the radiation dose to mitigate damage of the specimen, resulting in low-count Poisson data. In a low-dose scenario, state-of-the-art algorithms often exhibit poor reconstruction quality. Driven by this practical concern, we aim to improve standard phase retrieval techniques in the low-count regime. In particular, we strive for enhancing the optimization algorithms as described in Section 4.2 for low-count Poisson measurements.

Based on the maximum likelihood method, we derive different loss functions to design an optimization problem that suits the practical model. We develop an improved version of a variance-stabilizing transformation for the Poisson distribution and employ this as a decision rule for a regularization parameter used in an averaged amplitude-based loss. Finally, we establish step size rules that guarantee descent of the respective loss functions.

Throughout this chapter, we can assume $\mathbb{K} = \mathbb{C}$ and the measurement vectors $a_i,\ i \in [m]$, are of arbitrary choice, including the ptychographic model as one potential application. While we are motivated by the inverse problem of recovering the ground-truth $x \in \mathbb{C}^d$ from measurements

$$y_i \sim \mathit{Poisson}\left(|\langle a_i, x\rangle|^2\right), \quad i \in [m],$$

the theoretical guarantees are not restricted to this noise model.

The approaches and results discussed in the following are based on joint work with Benedikt Diederichs and Frank Filbir and were published in [33].

6.1. Maximum log-likelihood loss functions

Working with data drawn from a probability distribution which can be stated in terms of a parametric model, a common strategy is to use the method of maximum likelihood estimation if one aims at determining the respective distribution parameters. In this section, we briefly introduce the method before we derive the maximum likelihood function for relevant distributions, which we will then use to formulate optimization problems for phase retrieval.

Abstractly speaking, the idea of the maximum likelihood method is as follows. Assuming that $Y_i,\ i \in [m]$, are independent random variables with probability (density) functions $f(y_i; \theta_i)$, given distribution parameters θ_i the joint probability (density) function equals $\prod_{i=1}^m f(y_i; \theta_i)$.

Definition 6.1.1. *The likelihood function is defined as the joint probability (density) function* $\prod_{i=1}^m f(y_i;\theta_i)$ *considered as a function of the parameters* θ_i.

The idea of maximum likelihood estimation is to find the set of parameters that maximizes the likelihood function given observed data $y_i,\ i \in [m]$.

Definition 6.1.2. *The maximum likelihood estimate of the parameter* $\theta = (\theta_1, \ldots, \theta_m)$ *corresponding to observations* $Y_i = y_i,\ i \in [m]$, *is any*

$$\hat{\theta} \in \underset{\theta=(\theta_1,\ldots,\theta_m)}{\arg\max} \prod_{i=1}^m f(y_i;\theta_i). \tag{6.1}$$

Typically, it is used that

$$\begin{aligned}\underset{\theta=(\theta_1,\ldots,\theta_m)}{\arg\max} \prod_{i=1}^m f(y_i;\theta_i) &= \underset{\theta=(\theta_1,\ldots,\theta_m)}{\arg\max} \ \log\left(\prod_{i=1}^m f(y_i;\theta_i)\right) \\ &= \arg\max_{\theta} \sum_{i=1}^m \log\left(f(y_i;\theta_i)\right),\end{aligned}$$

which holds since the logarithm is a strictly monotone increasing function. Accordingly, the estimate (6.1) can equivalently be obtained via the often more convenient approach

$$\hat{\theta} \in \underset{\theta=(\theta_1,\ldots,\theta_m)}{\arg\max} \sum_{i=1}^m \log\left(f(y_i;\theta_i)\right).$$

This is commonly referred to as the maximum log-likelihood method.

We start with examining the maximum likelihood estimation for the Gaussian distribution, the prime example of a broadly studied distribution. In the optimization community in general, and also in the literature on discrete phase retrieval problems approached via optimization methods, this manifests itself in the popularity of considering loss functions of type (4.4), which we will find are derived via the maximum likelihood method for Gaussian random variables.

If $Y_i,\ i \in [m]$, are independent Gaussian random variables with mean ν_i and variance σ_i^2, their probability density functions read as

$$f(y_i;\nu_i,\sigma_i^2) = \tfrac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\tfrac{(y_i-\nu_i)^2}{2\sigma_i^2}\right).$$

The maximum likelihood estimate of the distribution mean given observed data $y_i,\ i \in [m]$, equals

$$\hat{\nu} \in \underset{\nu=(\nu_1,\ldots,\nu_m)}{\arg\max} \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i-\nu_i)^2}{2\sigma_i^2}\right),$$

under the assumption that the variances σ_i^2 are known. Clearly, it is preferable to consider the equivalent log-likelihood version, that is

$$\begin{aligned}\hat{\nu} \in & \underset{\nu=(\nu_1,\ldots,\nu_m)}{\arg\max} \sum_{i=1}^{m} \log\left(\tfrac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \tfrac{1}{2\sigma_i^2}\left(y_i - \nu_i\right)^2 \\ = & \underset{\nu=(\nu_1,\ldots,\nu_m)}{\arg\min} \sum_{i=1}^{m} \tfrac{1}{2\sigma_i^2}\left(y_i - \nu_i\right)^2 .\end{aligned}$$

In the context of a phase retrieval problem with measurements

$$y_i \approx |\langle a_i, x\rangle|^2, \quad i \in [m],$$

this technique can be adopted. Assuming that the measurements $y_i,\ i \in [m]$, are random variables following the distributions $\mathcal{N}(|\langle a_i, x\rangle|^2, \sigma_i^2),\ i \in [m]$, one aims to find

$$\hat{z} \in \underset{z \in \mathbb{C}^d}{\arg\min} \sum_{i=1}^{m} \tfrac{1}{2\sigma_i^2}\left(y_i - |\langle a_i, z\rangle|^2\right)^2 .$$

Accordingly, we formulate a minimization problem with loss

$$\mathcal{L}(z) = \sum_{i=1}^{m} \tfrac{1}{2\sigma_i^2}\left(y_i - |\langle a_i, z\rangle|^2\right)^2 , \tag{6.2}$$

which resembles (4.4).

Returning back to our original motivation, we seek an approach suitable for phaseless data with Poisson noise. In Section 3.1, we have seen that an approximation of the distribution $Poisson(\lambda)$ by a Gaussian distribution $\mathcal{N}(\lambda, \lambda)$ might be justified if the distribution parameter λ is sufficiently large. Using this approximation would suggest to work with a loss function

$$\mathcal{L}(z) = \sum_{i=1}^{m} \frac{1}{2\,|\langle a_i, z\rangle|^2}\left(y_i - |\langle a_i, z\rangle|^2\right)^2 .$$

This function, however, is not well-defined everywhere, and hence working with this loss is not recommended.

Comparing (6.2) with (4.4) we note that the latter model, which seems to be more popular in the literature, assumes a fixed variance σ^2 of the underlying distribution. The variance of a Poisson distribution is, however, clearly signal dependent. Nevertheless, loss functions as (4.4) are commonly used in applications where the data follows a Poisson distribution. We want to advice against such approaches and alternatively propose different loss functions in the following.

Naturally, we can also directly work with the log-likelihood function for the Poisson distribution itself. If $Y_i,\ i \in [m]$, are independent Poisson random variables with distribution parameter λ_i, their probability functions read as $\mathbb{P}(Y_i = y_i \mid \lambda_i) = \frac{1}{y_i!}\lambda_i^{y_i}\exp(-\lambda_i)$.

A maximum log-likelihood estimate of the distribution parameter given observed data $y_i,\ i \in [m]$, is

$$\begin{aligned}\hat{\lambda} &\in \underset{\lambda=(\lambda_1,\dots,\lambda_m)}{\arg\max} \sum_{i=1}^{m} y_i \log(\lambda_i) - \lambda_i - \log(y_i!) \\ &= \underset{\lambda=(\lambda_1,\dots,\lambda_m)}{\arg\min} \sum_{i=1}^{m} \lambda_i - y_i \log(\lambda_i).\end{aligned}$$

Using this approach for solving a phase retrieval problem with measurements $y_i \sim Poisson(|\langle a_i, x\rangle|^2),\ i \in [m]$, we aim to determine

$$\hat{z} \in \underset{z\in\mathbb{C}^d}{\arg\min} \sum_{i=1}^{m} |\langle a_i, z\rangle|^2 - y_i\ \log(|\langle a_i, z\rangle|^2).$$

Therefore, we formulate an optimization problem

$$\min_{z\in\mathbb{C}^d}\ \mathcal{L}_P(z)$$

with loss function

$$\mathcal{L}_P(z) = \sum_{i=1}^{m} |\langle a_i, z\rangle|^2 - y_i\ \log(|\langle a_i, z\rangle|^2). \tag{6.3}$$

Note that this function has singularities at points that satisfy $|\langle a_i, z\rangle|^2 = 0$. A simple strategy to deal with them is shifting the logarithmic term by a positive constant $\varepsilon_P > 0$, meaning that we would work with the loss function

$$\mathcal{L}_{P,\varepsilon_P}(z) = \sum_{i=1}^{m} |\langle a_i, z\rangle|^2 - y_i\ \log(|\langle a_i, z\rangle|^2 + \varepsilon_P). \tag{6.4}$$

This problem might be sensitive to the choice of the regularization parameter ε_P. Hence, we strive for a decision rule to determine the choice of ε_P. One option would be to work with a discrepancy principle to select a suitable parameter. That was investigated for inverse problems with Poisson data for example in [13, 14], albeit for a different type of regularization. To avoid a possibly expensive data exploration that is required for the application of the discrepancy principle, we propose to alternatively work with an approximation of the Poisson loss (6.4) with a rather inherent regularization parameter.

A first natural approach to this seems to be the following, as suggested in [123]. For $y_i > 0$, consider the Taylor approximation of the logarithm around $|\langle a_i, z\rangle| = \sqrt{y_i}$. This provides

$$\begin{aligned}
|\langle a_i, z\rangle|^2 - y_i \log\left(|\langle a_i, z\rangle|^2\right) &= |\langle a_i, z\rangle|^2 - 2y_i \log\left(|\langle a_i, z\rangle|\right) \\
&\approx |\langle a_i, z\rangle|^2 - 2y_i \left(\log(\sqrt{y_i}) + \tfrac{|\langle a_i, z\rangle| - \sqrt{y_i}}{\sqrt{y_i}} - \tfrac{\left(|\langle a_i, z\rangle| - \sqrt{y_i}\right)^2}{2y_i}\right) \\
&= 2\left(|\langle a_i, z\rangle| - \sqrt{y_i}\right)^2 - y_i \log(y_i) + y_i,
\end{aligned}$$

suggesting to replace (6.3) by

$$\mathcal{L}(z) = \sum_{i=1}^{m} 2\left(|\langle a_i, z\rangle| - \sqrt{y_i}\right)^2.$$

This matches the amplitude-based loss (4.5) introduced in Section 4.2. Note that this loss is not everywhere differentiable. Commonly, this is counteracted with a regularization of the form

$$\mathcal{L}(z) = \sum_{i=1}^{m} 2\left(\sqrt{|\langle a_i, z\rangle|^2 + \varepsilon_A} - \sqrt{y_i}\right)^2, \tag{6.5}$$

with $\varepsilon_A > 0$. Again, we are required to select a regularization parameter. In this regard, we do not benefit of this way of approximating the Poisson log-likelihood loss, hence we aim for a different approach.

We will find that the same type of loss, albeit with an inherent decision rule for the regularization parameter, is obtained when using the Gaussian log-likelihood for a Poisson random variable after a so-called variance stabilization transformation. This strategy is discussed in more detail in the following two sections, with the additional goal of improving the method in the regime of low-count Poisson data.

Notes and references. *We based the definition of a maximum likelihood estimator on [117], adapting also the respective terminology. When speaking of continuous distributions, we use the term probability density function, while for discrete distributions we talk about the probability function, which is otherwise often known as the probability mass function. However, we adapt the definitions to the case of independent but not-necessarily identically distributed random variables.*

6.2. Variance stabilization

As discussed above, utilizing the loss (4.4) for data with Poisson noise is questionable due to the signal-dependent variance. However, there exists a method that transforms Poisson random variables with signal-dependent variance into random variables of approximately Gaussian nature with fixed variance. This is known as the variance stabilization method.

In this section, we will introduce different adaptions of one type of variance-stabilizing transformations and contribute a variant that improves the stabilization attempt for Poisson data with small distribution parameter. Later, in Section 6.3, we will discuss how these transformations can be used to formulate a version of (6.2) that is more suitable for Poisson data than the popular least-squares minimization with (4.4).

The rationale of the variance stabilization method is as follows. Let $\lambda > 0$ and $Y \sim Poisson(\lambda)$. The mean $\mathbb{E}(Y)$ and the variance $\mathbb{V}(Y)$ of the random variable Y are both equal to λ. We aim to find a transformation $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ that satisfies $\mathbb{V}(f(Y)) \approx \sigma^2$ for some $\sigma \in \mathbb{R}_{\geq 0}$.

Assuming that $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is sufficiently smooth, its first order Taylor approximation around the variance λ is

$$f(t) \approx f(\lambda) + (t - \lambda)f'(\lambda).$$

Then, the variance of the transformed random variable is approximately equal to

$$\mathbb{V}(f(Y)) \approx \mathbb{V}(f(\lambda) + (Y - \lambda)f'(\lambda)) = \mathbb{V}(Y)\ (f'(\lambda))^2 = \lambda(f'(\lambda))^2.$$

In order to obtain an approximately constant variance $\mathbb{V}(f(Y)) \approx \sigma^2$, solve

$$\lambda(f'(\lambda))^2 = \sigma^2.$$

This is satisfied for $f(t) = 2\sigma\sqrt{t}$. Without loss of generality, we choose in the following $\sigma = \frac{1}{2}$ and aim at stabilizing the variance around $\frac{1}{4}$.

The above introduced approach of using a square-root transformation to obtain stabilized variance goes back to [8, 124]. It was also discovered empirically that a shifted square-root transformation $f(t) = \sqrt{t + c}$ with $c > 0$ might improve the variance stabilization. In [3], a fifth order Taylor approximation was used to determine a good choice of the parameter c. Following the same arguments as above, one obtains

$$\begin{aligned}
\mathbb{V}\left(\sqrt{Y+c}\,\right) &\approx (\lambda + c) \cdot \mathbb{V}\Bigg[1 + \frac{1}{2} \cdot \frac{Y-\lambda}{\sqrt{\lambda+c}} - \frac{1}{8} \cdot \left(\frac{Y-\lambda}{\sqrt{\lambda+c}}\right)^2 + \frac{1}{16} \cdot \left(\frac{Y-\lambda}{\sqrt{\lambda+c}}\right)^3 \\
&\qquad - \frac{5}{128} \cdot \left(\frac{Y-\lambda}{\sqrt{\lambda+c}}\right)^4 + \frac{7}{256} \cdot \left(\frac{Y-\lambda}{\sqrt{\lambda+c}}\right)^5\Bigg] \\
&\approx \frac{1}{4} \cdot \left(1 + \frac{\frac{3}{8} - c}{\lambda} + \frac{32c^2 - 52c + 17}{32\lambda^2}\right),
\end{aligned}$$

which suggests to choose $c = \frac{3}{8}$ to achieve $\mathbb{V}(Y) \approx \frac{1}{4}$. The function $f(t) = \sqrt{t + \frac{3}{8}}$ is commonly known as Anscombe transform, named after the author of [3].

While asymptotically a variance stabilization with Anscombe's transform or also the square-root transform with $c = 0$ works well, for $\lambda \in [0, 3]$ neither of these transformations performs satisfactorily; see Figure 6.1.

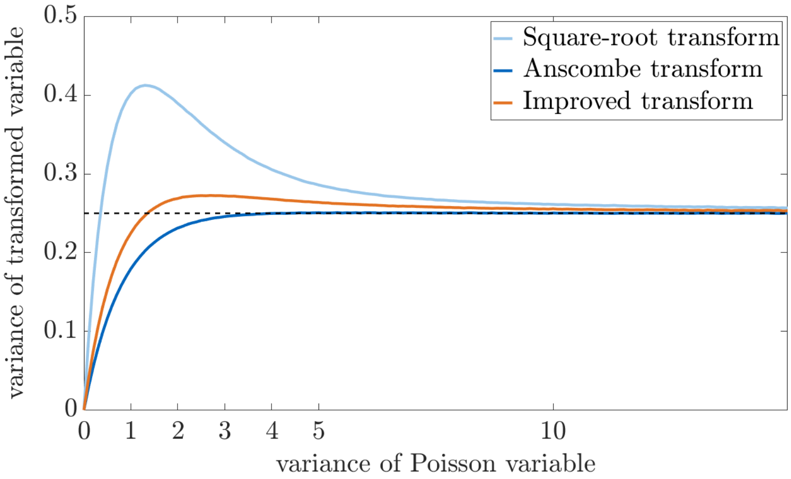


Figure 6.1.: Variance-stabilizing transformations.

Aiming for a better suited variance-stabilizing transformation of type $f(t) = \sqrt{t+c}$ for specific values of λ, we determine the parameter $c > 0$ such that $\mathbb{V}(\sqrt{Y+c}\,) = \frac{1}{4}$ by considering

$$\begin{aligned}\mathbb{V}\left(\sqrt{Y+c}\,\right) &= \mathbb{E}(Y) + c - \left[\mathbb{E}\left(\sqrt{Y+c}\,\right)\right]^2 \\ &= \lambda + c - \left(\sum_{k=0}^{\infty} \sqrt{k+c} \cdot \frac{\exp(-\lambda)\lambda^k}{k!}\right)^2 . \qquad (6.6)\end{aligned}$$

We find an approximately optimal value for c by setting this expression equal to $\frac{1}{4}$ and numerically solving the equation, using that the infinite sum decays quickly. For example, for $\lambda = 1$ we obtain $c \approx 0.12$ and for $\lambda = 2$ we get $c \approx 0.27$.

If we would like to cover more than only one specific value of λ, we can consider an averaging transformation

$$f(t) = \tfrac{1}{2}\left(\sqrt{t+c_1} + \sqrt{t+c_2}\,\right) \qquad (6.7)$$

for some parameters $c_1, c_2 \geq 0$. A popular example for such an averaging variance stabilization transformation is the Tukey-Freeman transform [48],

$$f(t) = \tfrac{1}{2}\left(\sqrt{t} + \sqrt{t+1}\,\right). \qquad (6.8)$$

Obviously, there is no choice of c (or c_1 and c_2) that provides the best variance stabilization universally for all distribution parameters λ. For different parameters λ, our calculations using (6.6) suggest different parameters. However, we can approach the transformation formulation heuristically from the experimental perspective. As an example, we would propose to select the parameters $c_1 = 0.12$ and $c_2 = 0.27$ if the considered

Poisson data are dominated by 1 and 2 counts. This choice is justified since for Poisson measurements of this size it is very likely that the underlying ground-truth is close to 1 or 2, respectively. We call in the following

$$f(t) = \frac{1}{2}\left(\sqrt{t+0.12} + \sqrt{t+0.27}\right) \tag{6.9}$$

an improved variance-stabilizing transform for low-count data.

The performance of this transformation in comparison to the square-root and the Anscombe transform is shown in Figure 6.1. We observe that for distribution parameters around 1 the proposed transform results in a variance closer to $\frac{1}{4}$ than the outcome of the compared transformations. Note that this is the regime we are interested in, and hence the proposed transform meets our demands. For larger values, the proposed transform performs similarly to Anscombe's transform, and for a sufficiently large distribution parameter all methods work well.

For λ close to 0 we note that the square-root transformation $f(t) = \sqrt{t}$ might be preferred over the other variance-stabilizing transformations. This is also confirmed by the method using (6.6), which suggests to choose $c = 0$ for $\lambda \to 0$.

In Section 6.5 we will show empirically that also in the context of phase retrieval with Poisson data the application of (6.9) exhibits an advancement compared to the simple square-root transformation or Anscombe's transform.

Notes and references. *A different approach for optimizing the variance stabilization transformation was proposed in [45]. However, this method does not provide a parametric description of the transformation and is hence not applicable for our purpose.*

The need for a better variance stabilizing transformation for low-count Poisson data than for example the Anscombe transform was also emphasized by [138]. This paper addresses the problem with an improved rule for the choice of c for an application involving filtered discrete Poisson processes. In the problem setting considered in this chapter, this approach reduces albeit again to the choice $c = \frac{3}{8}$.

The choice of transformation discussed here is driven rather by the variance stabilization objective than by achieving approximate normality with the resulting distribution. The trade-off between these two ambitions was discussed in detail in [38]. For example, other transformations would perform better in terms of reducing the skewness of the Poisson distribution to obtain something approximately Gaussian, a distribution with zero skewness. However, applying square-root type transformations as discussed in this section to approximate normality is reasonable since also these transformations reduce the distribution skewness. In terms of this perspective, [137] also discovered that averaging transformations (6.7) *are advantageous over simple square-root transforms in case of low distribution parameters* λ*, justifying our choice of transformation in the sequel.*

6.3. Approximations of the Poisson log-likelihood loss

Being equipped with a method to transform Poisson data into (approximately) Gaussian distributed data with constant variance, we return now to the discussion on suitable loss functions for Poisson data.

Assume that we consider $y_i \sim \mathit{Poisson}(|\langle a_i, x\rangle|^2)$, $i \in [m]$. Applying any of the previously presented variance-stabilizing transformations $f(t) = \sqrt{t+c}$, $c \geq 0$, to these random variables, we interpret the transformed variables $f(y_i)$ to be random variables following a Gaussian distribution $\mathcal{N}\big(\mathbb{E}\left(f(y_i)\right), \frac{1}{4}\big)$. If, for some random variable Y, the parameter $c \geq 0$ is chosen such that $\mathbb{V}\left(\sqrt{Y+c}\right) \approx \frac{1}{4}$, the mean of the transformed random variable is approximately equal to

$$\mathbb{E}\big(\sqrt{Y+c}\,\big) = \sqrt{\,\mathbb{E}\left(Y+c\right) - \mathbb{V}\big(\sqrt{Y+c}\,\big)\,} \approx \sqrt{\,\mathbb{E}(Y) + c - \tfrac{1}{4}\,}\,. \tag{6.10}$$

Hence, after transformation, we work with data that follows (approximately) the distribution $\mathcal{N}\big(\sqrt{|\langle a_i, x\rangle|^2 + c - \frac{1}{4}}\,, \frac{1}{4}\big)$. For this type of data, we would like to work with the loss (6.2), that is, the loss derived from the log-likelihood function for a Gaussian distribution. For the transformed data, the loss reads as

$$\mathcal{L}(z) = \sum_{i=1}^{m} 2\left(\sqrt{|\langle a_i, z\rangle|^2 + c - \tfrac{1}{4}} - \sqrt{y_i + c}\right)^2. \tag{6.11}$$

A variant of this type of loss function is the amplitude loss

$$\mathcal{L}(z) = \sum_{i=1}^{m} 2\left(\sqrt{|\langle a_i, z\rangle|^2 + \varepsilon_A} - \sqrt{y_i}\right)^2, \tag{6.12}$$

with $\varepsilon_A > 0$. This arises from the Gaussian model when using a square-root transform with $c = 0$ but disregarding the relation (6.10) and instead using the coarse approximation $\mathbb{E}\big(\sqrt{Y}\,\big) \approx \sqrt{\,\mathbb{E}(Y)\,}$. We had already seen this loss function earlier in (6.5) derived via a Taylor approximation of the Poisson log-likelihood function. Therefore, we conclude that the amplitude loss can be regarded as a reasonable alternative to the original Poisson log-likelihood loss from two approximation perspectives.

Note that functions (6.11) are not well-defined for z with $|\langle a_i, z\rangle|^2 \in \big[0, \frac{1}{4} - c\big)$, and not differentiable at $|\langle a_i, z\rangle|^2 = \frac{1}{4} - c$. Thus, applying a gradient descent method to such loss functions is problematic for choices $c \leq \frac{1}{4}$ and we need to regularize again to counteract this issue. Thereto, we decide to also adapt the coarse approximation $\mathbb{E}\left(\sqrt{Y+c}\,\right) \approx \sqrt{\mathbb{E}(Y) + c}$ and work with

$$\mathcal{L}(z) = \sum_{i=1}^{m} 2\left(\sqrt{|\langle a_i, z\rangle|^2 + c} - \sqrt{y_i + c}\right)^2. \tag{6.13}$$

Compared to (6.11), this loss function can be considered as unbiased since it is minimal for z satisfying $|\langle a_i, z\rangle|^2 = y_i,\ i \in [m]$, and hence, in case of noise-free measurements, the ground-truth solution would be a minimizer of this loss.

More generally, variance stabilization can also be obtained via an averaging transformation (6.7) with values $c_1, c_2 \geq 0$. Due to the same reasons as before, we coarsely approximate

$$\mathbb{E}\left(\sqrt{Y+c_1} + \sqrt{Y+c_2}\,\right) \approx \sqrt{\mathbb{E}(Y)+c_1} + \sqrt{\mathbb{E}(Y)+c_2}\,,$$

and consider loss functions of the form

$$\mathcal{L}(z) = \sum_{i=1}^{m} \frac{1}{2}\Big(\sqrt{|\langle a_i, z\rangle|^2 + c_1} + \sqrt{|\langle a_i, z\rangle|^2 + c_2} - \sqrt{y_i + c_1} - \sqrt{y_i + c_2}\,\Big)^2.$$

For distribution parameters λ close to 0, all discussed variance-stabilizing transformations result in a variance rather close to 0 than close to $\frac{1}{4}$. For this reason, in case of ground-truth values equal or close to 0, we might better abstain from using the loss functions resulting from the variance stabilization approach. In experiments, the ground-truth values are obviously not known - otherwise there would be no need of taking the noise distribution into account. However, if a value $y_i = 0$ is measured, it is likely that the corresponding ground-truth is close to 0. Since we wish to avoid variance-stabilizing transformations in these instances, we aim for a loss adaption for zero measurements.

Note that for $y_i = 0$ the loss function (6.4) reduces to $|\langle a_i, z\rangle|^2$, and hence no regularization is required in the exact Poisson log-likelihood model. Consequently, we propose to work with

$$\begin{aligned}\mathcal{L}_0(z) = \sum_{i=1}^{m} & \mathbb{1}_{y_i=0} \cdot |\langle a_i, z\rangle|^2 \\ & + \mathbb{1}_{y_i>0} \cdot \frac{1}{2}\left(\sqrt{|\langle a_i, z\rangle|^2 + c_1} + \sqrt{|\langle a_i, z\rangle|^2 + c_2} - \sqrt{y_i + c_1} - \sqrt{y_i + c_2}\,\right)^2,\end{aligned} \tag{6.14}$$

with $c_2 \geq c_1 > 0,\ i \in [m]$, in applications with low-dose data dominated by zero measurements. Combining this loss function with the parameters c_1, c_2 derived in Section 6.2, we can formulate an optimization problem for phase retrieval with particular emphasis on low-count data.

Notes and references. *Loss functions based on the maximum log-likelihood model for a Gaussian distribution using the Anscombe transform were considered before in [70]. A similar approach was also suggested for mixed Poisson-Gaussian noise in [140]. Both of these works also used the approximation* $\mathbb{E}\big(\sqrt{Y+\frac{3}{8}}\,\big) \approx \sqrt{\mathbb{E}(Y)+\frac{3}{8}}$ *instead of* (6.10). *Loss functions of type* (6.13) *were also investigated in [49] under the name perturbed amplitude-based model, with the intention to smoothen the amplitude-based loss function*

(6.12). *Furthermore, it is mentioned that the shift by* $c > 0$ *effects a truncation of too large contributions to the gradient of the loss. In our approach, on the other hand, the choice* $c > 0$ *is motivated by the improved variance stabilization. We conclude that multiple arguments speak in favor of working with loss functions of type* (6.13) *or* (6.14)*, with the latter being an advancement of the other functions in respect of low-count data dominated by* 0 *measurements.*

6.4. Convergence guarantees

In this section, we derive convergence guarantees for Wirtinger gradient descent algorithms used for solving optimization problems involving the loss functions discussed in Section 6.1 and Section 6.3. The results apply for any measurement frame, including also the example of ptychographic measurements. Moreover, note that the guarantees hold for all types of noise, but ensure descent in the loss, that means fitting of the measurements, as opposed to global convergence to the ground-truth.

As mentioned in Section 4.2, an optimization problem $\min_z \mathcal{L}(z)$, involving a loss function $\mathcal{L} : \mathbb{C}^d \to \mathbb{R}$, can be solved with a gradient descent algorithm using Wirtinger derivatives. An overview about Wirtinger derivatives is provided in Appendix B, including a general convergence guarantee (Proposition B.1) that we will apply in the following to examine the convergence behavior of the loss functions of our interest. All loss functions considered in this chapter are of the form $\mathcal{L}(z) = \ell \circ \varphi(z)$, with a function $\ell : \mathbb{R}^m \to \mathbb{R}$ and $\varphi : \mathbb{C}^d \to \mathbb{R}^m$, $z \mapsto (|\langle a_i, z\rangle|^2)_{i\in[m]}$, and we specified the requirements for Proposition B.1 accordingly; see Lemma B.2 and Lemma B.3. With these auxiliary results, we can prove convergence guarantees for Wirtinger gradient descent algorithms for various loss functions.

We start by analyzing the Wirtinger gradient descent algorithm, Algorithm 2, corresponding to $\mathcal{L}_{P,\varepsilon_P}$, the regularized Poisson log-likelihood loss defined in (6.4). Since

$$\nabla_z \mathcal{L}_{P,\varepsilon_P}(z) = \sum_{i=1}^{m} \left(1 - \frac{y_i}{|\langle a_i, z_k\rangle|^2 + \varepsilon_P}\right) \langle a_i, z_k\rangle\, a_i,$$

the respective gradient descent update rule with constant step size μ reads as

$$z_{k+1} = z_k - \mu \cdot \sum_{i=1}^{m} \left(1 - \frac{y_i}{|\langle a_i, z_k\rangle|^2 + \varepsilon_P}\right) \langle a_i, z_k\rangle\, a_i. \tag{6.15}$$

We find a bound on the step size μ that guarantees descent of the loss (6.4).

Theorem 6.4.1. *Let* $\mathcal{L}_{P,\varepsilon_P}$ *as given in* (6.4) *with* $\varepsilon_P > 0$*. Let* M *be the matrix with rows* $\sqrt{1 + y_i/(8\varepsilon_P)} \cdot a_i^*$, $i \in [m]$*, and choose* $\mu \le \|M\|^{-2}$*. Then, the sequence* $(z_k)_{k\ge 0} \subset \mathbb{C}^d$ *defined by* (6.15)*, with an arbitrary initialization* $z_0 \in \mathbb{C}^d$*, converges to a stationary point of* $\mathcal{L}_{P,\varepsilon_P}$*.*

Proof. The considered loss is of the form $\mathcal{L}_{P,\varepsilon_P}(z) = \sum_{i=1}^m \ell_i \circ \varphi(z)$ with $\ell_i : [0,\infty) \to \mathbb{R}$, $t \mapsto t - y_i \log(t + \varepsilon_P)$ and $\varphi(z) = |\langle a_i, z\rangle|^2$. We compute

$$2 \cdot \ell_i''(t) \cdot t + \ell_i'(t) = 2 \cdot \frac{y_i}{(t+\varepsilon_P)^2} \cdot t + 1 - \frac{y_i}{t+\varepsilon_P} = 1 + \frac{y_i(t-\varepsilon_P)}{(t+\varepsilon_P)^2} \leq 1 + \frac{y_i}{8\varepsilon_P},$$

using that $\frac{s-\varepsilon_P}{(s+\varepsilon_P)^2} \leq \frac{1}{8\varepsilon_P}$ for every $s \in \mathbb{R}$ and any fixed $\varepsilon_P > 0$. Lemma B.2 together with this bound yields

$$\begin{aligned}\begin{pmatrix} v \\ \overline{v} \end{pmatrix}^* \nabla^2 \mathcal{L}_{P,\varepsilon_P}(z) \begin{pmatrix} v \\ \overline{v} \end{pmatrix} &\leq 2 \cdot \sum_{i=1}^m \left| \left\langle \sqrt{1 + \tfrac{y_i}{8\varepsilon_P}} \cdot a_i, v \right\rangle \right|^2 \\ &\leq 2\left\| M \right\|^2 \left\| v \right\|_2^2 = \left\| M \right\|^2 \left\| \begin{pmatrix} v \\ \overline{v} \end{pmatrix} \right\|_2^2\end{aligned}$$

for all $z, v \in \mathbb{C}^d$. Due to Proposition B.1, this suggests to choose a step size $\mu \leq \left\| M \right\|^{-2}$.

Furthermore, Lemma B.3 and Remark B.4 provide that $\mathcal{L}_{P,\varepsilon_P}$ has compact sublevel sets on the subspace $z_0 + \text{range}(A^*)$, which contains all feasible iterates of the algorithm. Clearly, all functions ℓ_i are continuous and bounded from below, more precisely $\ell_i(t) \geq y_i - y_i \log(y_i)$ for all $t \in [0,\infty)$. Moreover, for all $t \in [0,\infty)$, we have

$$\ell_i(t) = \tfrac{1}{2}t + \tfrac{1}{2}t - y_i \log(t + \varepsilon_P) \geq \tfrac{1}{2}t + y_i - \tfrac{1}{2}\varepsilon_P - y_i \log(2y_i).$$

Consequently, $\ell_i(t) \to \infty$ for $t \to \infty$.

With Proposition B.1 we conclude that the considered gradient descent algorithm converges to a stationary point of $\mathcal{L}_{P,\varepsilon_P}$. □

The key conclusion of Theorem 6.4.1 is that the step size μ must be chosen of the order of the regularization parameter ε_P to guarantee convergence. If the problem shall be manipulated with only a small regularization, this requires a small step size, coming along with slow convergence.

Next, we investigate the convergence of gradient descent algorithms for loss functions discussed in Section 6.3. These losses are all variations of a function

$$\mathcal{L}_{avg}(z) = \sum_{i=1}^m \frac{1}{2} \left(\sqrt{\left|\langle a_i, z\rangle\right|^2 + c_1} + \sqrt{\left|\langle a_i, z\rangle\right|^2 + c_2} - C_i \right)^2, \tag{6.16}$$

with constants $c_2 \geq c_1 > 0$, $C_i \geq 0$, $i \in [m]$. The Wirtinger gradient of $\mathcal{L}_{avg}$ is

$$\nabla_z \mathcal{L}_{avg}(z) = \frac{1}{2} \sum_{i=1}^{m} \left(\sqrt{|\langle a_i, z_k \rangle|^2 + c_1} + \sqrt{|\langle a_i, z_k \rangle|^2 + c_2} - C_i \right) \cdot \left(\frac{1}{\sqrt{|\langle a_i, z_k \rangle|^2 + c_1}} + \frac{1}{\sqrt{|\langle a_i, z_k \rangle|^2 + c_2}} \right) \langle a_i, z_k \rangle \, a_i.$$

For any initialization $z_0 \in \mathbb{C}^d$, define a sequence $(z_k)_{k \geq 1}$ by

$$z_{k+1} = z_k - \mu \cdot \frac{1}{2} \sum_{i=1}^{m} \left(\sqrt{|\langle a_i, z_k \rangle|^2 + c_1} + \sqrt{|\langle a_i, z_k \rangle|^2 + c_2} - C_i \right) \cdot \left(\frac{1}{\sqrt{|\langle a_i, z_k \rangle|^2 + c_1}} + \frac{1}{\sqrt{|\langle a_i, z_k \rangle|^2 + c_2}} \right) \langle a_i, z_k \rangle \, a_i, \tag{6.17}$$

with step size $\mu > 0$. The Wirtinger gradient descent algorithm with update (6.17) satisfies the following convergence guarantee.

Theorem 6.4.2. *Let A denote the matrix with rows a_i^*, $i \in [m]$. The sequence $(z_k)_{k \geq 0} \subset \mathbb{C}^d$ defined by (6.17), for any initialization $z_0 \in \mathbb{C}^d$, converges to a stationary point of the loss function $\mathcal{L}_{avg}$ defined in (6.16) provided $\mu \leq 2 \left(\left(3 + \sqrt{c_2/c_1}\right) \|A\|^2 \right)^{-1}$.*

Proof. We proceed again as in the proof of Theorem 6.4.1. For loss functions $\mathcal{L}_{avg}$, a bound for the Hessian is obtained from Lemma B.2 with $\ell_i : [0, \infty) \to [0, \infty)$, $t \mapsto \frac{1}{2} \left(\sqrt{t + c_1} + \sqrt{t + c_2} - C_i \right)^2$. The first derivative is

$$\begin{aligned} \ell_i'(t) &= \frac{1}{2} \left(\sqrt{t + c_1} + \sqrt{t + c_2} - C_i \right) \cdot \left(\frac{1}{\sqrt{t + c_1}} + \frac{1}{\sqrt{t + c_2}} \right) \\ &= \frac{1}{2} \left[2 + \frac{\sqrt{t + c_1}}{\sqrt{t + c_2}} + \frac{\sqrt{t + c_2}}{\sqrt{t + c_1}} - C_i \cdot \left(\frac{1}{\sqrt{t + c_1}} + \frac{1}{\sqrt{t + c_2}} \right) \right], \end{aligned}$$

and the second derivative satisfies

$$\begin{aligned} \ell_i''(t) &= \frac{1}{4} \left(\frac{1}{\sqrt{t + c_1}} + \frac{1}{\sqrt{t + c_2}} \right)^2 - \frac{1}{4} \left(\sqrt{t + c_1} + \sqrt{t + c_2} - C_i \right) \left(\frac{1}{(t + c_1)^{\frac{3}{2}}} + \frac{1}{(t + c_2)^{\frac{3}{2}}} \right) \\ &= \frac{1}{4} \left[\frac{2}{\sqrt{t + c_1}\sqrt{t + c_2}} - \frac{\sqrt{t + c_1}}{(t + c_2)^{\frac{3}{2}}} - \frac{\sqrt{t + c_2}}{(t + c_1)^{\frac{3}{2}}} + C_i \cdot \left(\frac{1}{(t + c_1)^{\frac{3}{2}}} + \frac{1}{(t + c_2)^{\frac{3}{2}}} \right) \right] \\ &\leq \frac{1}{4} C_i \cdot \left(\frac{1}{(t + c_1)^{\frac{3}{2}}} + \frac{1}{(t + c_2)^{\frac{3}{2}}} \right). \end{aligned}$$

For the last bound we use that $\frac{2}{\alpha\beta} \leq \frac{\alpha}{\beta^3} + \frac{\beta}{\alpha^3}$ for $\alpha, \beta \in \mathbb{R} \setminus \{0\}$. Together, we find

$$\begin{aligned}
2\,t\,\ell_i''(t) + \ell_i'(t) &\leq \frac{1}{2} C_i \cdot \left(\frac{t}{(t+c_1)^{\frac{3}{2}}} + \frac{t}{(t+c_2)^{\frac{3}{2}}} \right) \\
&\quad + \frac{1}{2} \left[2 + \frac{\sqrt{t+c_1}}{\sqrt{t+c_2}} + \frac{\sqrt{t+c_2}}{\sqrt{t+c_1}} - C_i \cdot \left(\frac{1}{\sqrt{t+c_1}} + \frac{1}{\sqrt{t+c_2}} \right) \right] \\
&= \frac{1}{2} \left[-C_i \cdot \left(\frac{c_1}{\sqrt{t+c_1}} + \frac{c_2}{\sqrt{t+c_2}} \right) + 2 + 1 + \sqrt{\frac{c_2}{c_1}} \right] \\
&\leq \frac{1}{2} \left(3 + \sqrt{\frac{c_2}{c_1}} \right),
\end{aligned}$$

and, consequently, Lemma B.2 provides

$$\begin{pmatrix} v \\ \overline{v} \end{pmatrix}^* \nabla^2 \mathcal{L}_{avg}(z) \begin{pmatrix} v \\ \overline{v} \end{pmatrix} \leq \sum_{i=1}^m \left(3 + \sqrt{\frac{c_2}{c_1}} \right) |\langle a_i, v \rangle|^2 \leq \frac{1}{2} \left(3 + \sqrt{\frac{c_2}{c_1}} \right) \|A\|^2 \left\| \begin{pmatrix} v \\ \overline{v} \end{pmatrix} \right\|_2^2$$

for all $z, v \in \mathbb{C}^d$. Combined with Proposition B.1, this suggests to choose the step size $\mu \leq 2 \left(\left(3 + \sqrt{c_2/c_1}\right) \|A\|^2 \right)^{-1}$.

Again, the function $\mathcal{L}_{avg}$ restricted to $z_0 + \operatorname{range}(A^*)$ has compact level sets according to Lemma B.3 and Remark B.4. All functions ℓ_i are continuous, bounded from below, and satisfy $\ell_i(t) \to \infty$ for $t \to \infty$.

Together, Proposition B.1 guarantees convergence of the gradient descent with update (6.17) to a stationary point of the loss function $\mathcal{L}_{avg}$. $\square$

This result also implicates a rule for the step size when using a gradient descent algorithm for a loss (6.11) or (6.12) as these are special cases of $\mathcal{L}_{avg}$.

Corollary 6.4.3. *A Wirtinger gradient descent algorithm for the loss function* (6.12) *for any* $\varepsilon_A > 0$ *is guaranteed to converge to a stationary point of* (6.12) *if the step size is chosen to satisfy* $\mu \leq \left(2\,\|A\|^2\right)^{-1}$.

The important difference between this result and the convergence guarantee for the Poisson log-likelihood loss, as stated in Theorem 6.4.1, is that the bound on the step size needed for convergence is independent of a potentially small regularization parameter.

On the other hand, if the loss (6.16) is based on a variance-stabilizing transformation as the Tukey-Freeman transform (6.8) with $c_1 = 0$, requiring a regularization, we return to the same problem as for the Poisson log-likelihood loss.

Corollary 6.4.4. *A loss function using the Tukey-Freeman transform* (6.8) *for variance stabilization would read as*

$$\mathcal{L}(z) = \sum_{i=1}^{m} \frac{1}{2} \left(\sqrt{\ |\langle a_i, z \rangle|^2 + \varepsilon\ } + \sqrt{\ |\langle a_i, z \rangle|^2 + 1\ } - \sqrt{y_i} - \sqrt{y_i + 1} \right)^2 ,$$

with $\varepsilon > 0$ to avoid the singularity in the derivative. This is a special case of the loss examined in Theorem 6.4.2 and we conclude that a gradient descent algorithm for an optimization problem corresponding to this loss function requires a step size $\mu \leq 2\left(\left(3 + \sqrt{1/\varepsilon}\right) \|A\|^2\right)^{-1}$ for guaranteed convergence.

Finally, we find that the proposed loss $\mathcal{L}_0$ can be treated with the same fixed step size as in Theorem 6.4.2 to guarantee convergence to a stationary point of $\mathcal{L}_0$.

Corollary 6.4.5. *Convergence to a stationary point of the loss $\mathcal{L}_0$ defined in* (6.14) *is achieved by the corresponding Wirtinger gradient descent algorithm if the step size satisfies $\mu \leq 2\left(\left(3 + \sqrt{c_2/c_1}\right) \|A\|^2\right)^{-1}$.*

Proof. In $\mathcal{L}_0$, the functions ℓ_i for all $i \in [m]$ with $y_i = 0$ can be chosen as the identity $\ell_i : [0, \infty) \to [0, \infty)$, $t \mapsto t$, with its derivatives

$$\ell_i'(t) = 1 \quad \text{and} \quad \ell_i''(t) = 0.$$

Together, we obtain

$$2\, t\, \ell_i''\,(t) + \ell_i'\,(t) = 1,$$

and since $1 \leq \frac{1}{2}\left(3 + \sqrt{c_2/c_1}\right)$ for any $c_2 \geq c_1 > 0$, we can simply upper-bound the Hessian as in the proof of Theorem 6.4.2. Then, the same conclusion can be drawn as in Theorem 6.4.2. □

Remark 6.4.6. *Note that for the intensity-based loss* (6.2) *we cannot obtain a similar convergence guarantee based on Proposition B.1 since there is no uniform upper bound for $2\,\ell_i''\,(t)\ t + \ell_i'\,(t) = 6t - 2y_i$ if $\ell_i(t) = (t - y_i)^2$.*

Notes and references. *Corollary 6.4.3 parallels the result of Theorem A.1 in [135], meaning that our result in Theorem 6.4.2 generalizes the convergence guarantee of [135] for the amplitude loss to the class of loss functions* (6.16).

In [28], a convergence analysis for an algorithm similar to (6.15) *is presented. In contrast to our Theorem 6.4.1 that ensures convergence to stationary points of the loss function, the authors of [28] prove a global convergence guarantee, however only in combination with a good initialization and for Gaussian measurement vectors. Our results hold only locally, but independently of the quality of the initialization and for arbitrary measurement vectors, enhancing their relevance for applications. Moreover, our results are independent of the quality of the measurements, and apply for any potential type of* *perturbation. In this respect, however, we have to note that our convergence guarantees ensure that the measurements are fitted in terms of the considered loss functions. This can result in an overfitting of the noise contributions, depending on the choice of regularization.*

Comparable in scope to our result Theorem 6.4.1, [90] prove a convergence guarantee to a minimizer of the Poisson maximum log-likelihood loss function in a lifted matrix recovery version, for a real-valued ground-truth and measurement frame, tackled by a Frank-Wolfe algorithm.

Note that knowing a bound on the step size as derived in this chapter allows for omitting computationally expensive line-search algorithms. On the other hand, choosing a constant step size according to the given bounds corresponds to the worst-case scenario, and a faster convergence could be obtained by a more flexible step size selection.

6.5. Numerical experiments

We corroborate the theoretical consideration on Wirtinger gradient descent algorithms for low-dose Poisson phase retrieval with numerical experiments.

In the following, we experiment with synthetic data based on a test object $x \in \mathbb{C}^d$ and complex Gaussian measurement vectors $a_i \in \mathbb{C}^d,\ i \in [m]$. Unless specified otherwise, we work with $d = 256$ and $m = 10d$, but conclude the section with a more detailed analysis for different object dimensions d and oversampling ratios $\frac{m}{d}$.

We normalize the measurement vectors such that $\sum_{i=1}^{m} |\langle a_i, x\rangle|^2 = 1$, so that a noiseless value $|\langle a_i, x\rangle|^2$ can be interpreted as the probability that particles arrive at the ith detector pixel. We fix a dose parameter $\mathscr{d} > 0$ and work with measurements

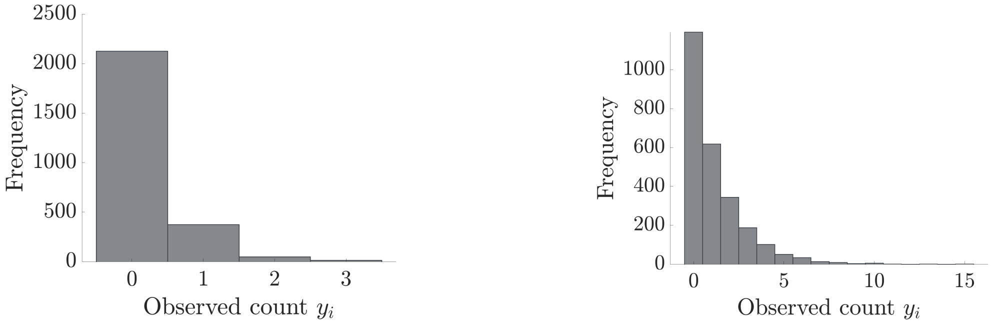
$$y_i \sim \mathit{Poisson}\left(\mathscr{d} \cdot |\langle a_i, x\rangle|^2\right), \quad i \in [m].$$

Except where otherwise stated, we experiment with doses $\mathscr{d} \in \{500, 1000, 1500, \ldots, 3000\}$. The histograms in Figure 6.2 show the distribution of exemplarily realized measurement for a small dose $\mathscr{d} = 500$ and a higher dose $\mathscr{d} = 3000$.

All algorithms discussed in this section were initialized by the output of the spectral method; see Algorithm 1. Furthermore, all algorithms are stopped at the latest after 1000 iterations or as soon as a stopping criterion is met, which requires the respective loss function evaluated at the respective iterate to decrease by less than 10^{-6} within one iteration. Our numerical trials showed that this criterion enabled the algorithms to saturate at a local minimum in the majority of cases.

The performance of the various investigated algorithms is compared in terms of the relative reconstruction error $\mathit{dist}(x, x_K)/\, \|x\|_2$, where x is the ground-truth and x_K is the reconstruction result obtained after K iterations of the respective algorithm.

We start with testing the gradient descent algorithm using the Poisson log-likelihood loss (6.4) for different regularization parameters ε_P. The respective algorithm is named



(a) Dose $\mathscr{d} = 500$. (b) Dose $\mathscr{d} = 3000$.

Figure 6.2.: Distribution of measurements for a low-dose and a higher-dose experiment.

'Poisson flow' (PF). We conclude from Figure 6.3 that the performance of the algorithm clearly depends on the choice of the regularization parameter ε_P. While in higher-dose experiments the algorithm is less sensitive to the parameter selection, a low-dose experiment demands knowledge on a good choice for ε_P.

Note that, alternatively to (6.4), we could also work with the loss

$$\mathcal{L}(z) = \sum_{i=1}^{m} |\langle a_i, z\rangle|^2 - (y_i + \varepsilon_P) \log\left(|\langle a_i, z\rangle|^2 + \varepsilon_P\right). \tag{6.18}$$

Compared to (6.4), this can be interpreted as unbiased since it is minimized by z satisfying $|\langle a_i, z\rangle|^2 = y_i,\ i \in [m]$. Accordingly, this method is less sensitive to the selection of ε_P. However, the numerical experiments shown in Figure 6.4 suggest that the corresponding algorithm performs worse than the algorithm using loss (6.4) with a good parameter ε_P.

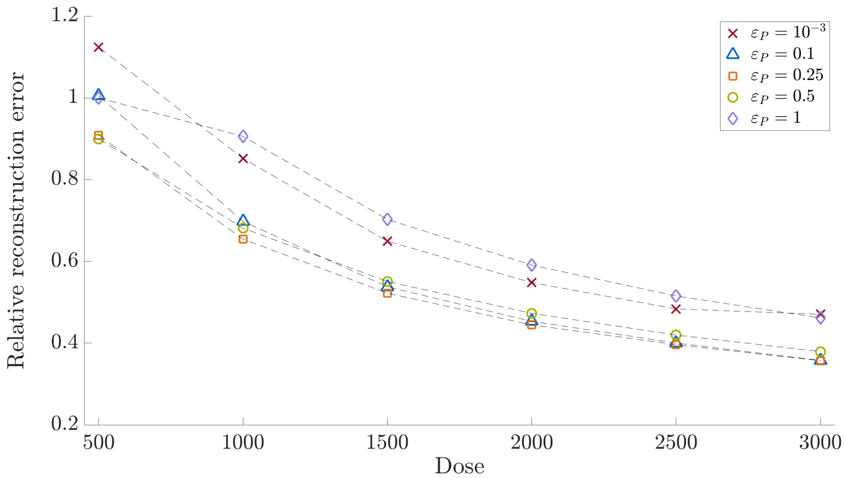


Figure 6.3.: Performance of the Poisson flow with various regularization parameters $\varepsilon_P \in \{10^{-3}, 0.1, 0.25, 0.5, 1\}$. The plot reports averages of 50 trials per choice of dose parameter.

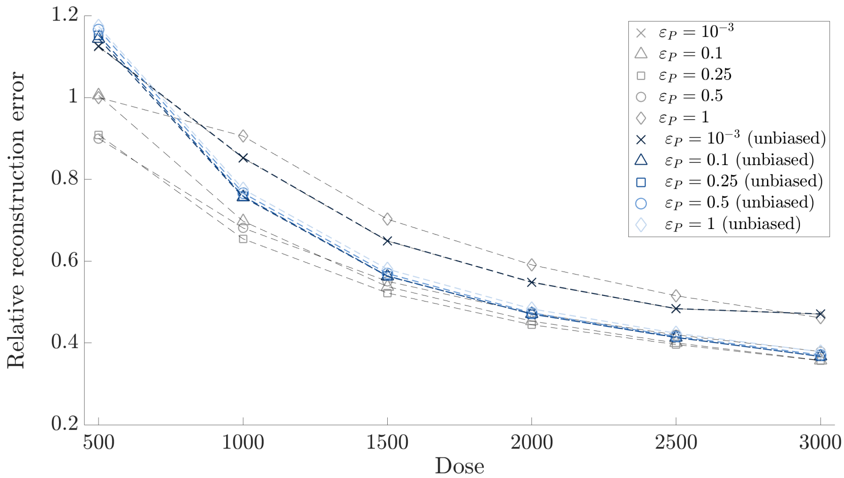


Figure 6.4.: Performance of the Poisson flow for loss (6.4) versus the unbiased loss (6.18) with various regularization parameters $\varepsilon_P \in \{10^{-3}, 0.1, 0.25, 0.5, 1\}$. The plot reports averages of 50 trials per choice of dose parameter.

Moreover, the step size recommended by the convergence guarantee remains parameter dependent.

We continue with a comparison between the Poisson flow and the gradient descent algorithm using the Gaussian log-likelihood loss with the proposed improved variance stabilization transformation. We propose to use the name 'Flow with improved variance stabilization' (FIVS) for the algorithm corresponding to the loss function (6.14) with $c_1 = 0.12$, $c_2 = 0.27$, and $C_i = \sqrt{y_i + c_1} + \sqrt{y_i + c_2}$, $i \in [m]$, and 'Flow with improved variance stabilization without adaption for zero measurements' (FIVS-WA) if (6.16) with

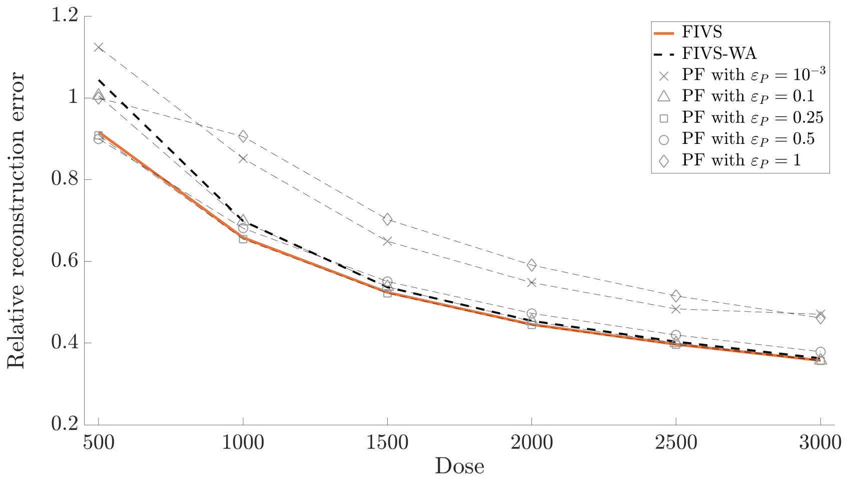


Figure 6.5.: Performance of the gradient descent algorithms for the proposed loss functions based on the improved variance stabilization approach. The gray lines correspond to the Poisson flow as in Figure 6.3. The plot shows averages of 50 trials per choice of dose parameter.

the same parameters is the underlying loss function. Figure 6.5 indicates that the FIVS algorithm performs in all instances comparably to the Poisson flow with the respective close to optimal choice of regularization parameter. Being aware of an optimal parameter selection, the Poisson flow can be the method of choice. However, without a decision rule at hand, FIVS is an advisable alternative. Nevertheless, it seems surprising to us that the biased Poisson flow with non-negligible regularization parameters outputs nearly identical reconstructions as the unbiased FIVS method, also since the unbiased Poisson loss performs noticeably worse in the low-count scenario. Revealing this connection more closely, perhaps by investigating the respective loss landscapes in more detail, is an interesting aspect for future work.

When comparing FIVS and FIVS-WA, we note that the proposed adaption (6.14) of the loss for zero measurements is essential for a low-dose experiment which is dominated by zero measurements. Consequently, we focus solely on FIVS, that means the algorithm with the adaption for zero measurements, in the sequel.

Next, we compare the performance of FIVS with two approaches that are neither built on the variance stabilization idea nor require to select a regularization parameter as in the Poisson flow. As common in the literature, we denote with 'Wirtinger flow' (WF) the algorithm for (6.2) with constant variance $\sigma_i^2 = 1/4,\ i \in [m]$. This was the first proposed variant of a Wirtinger gradient descent algorithm [19]. Moreover, we test the 'truncated Wirtinger flow' (TWF) as defined in [28]. This algorithm is based on a truncation principle for the Poisson maximum log-likelihood loss. In our implementation, we did not tune the required parameters but used the numbers suggested by the authors of [28]. Figure 6.6 shows that neither of these two algorithms can compete with the proposed

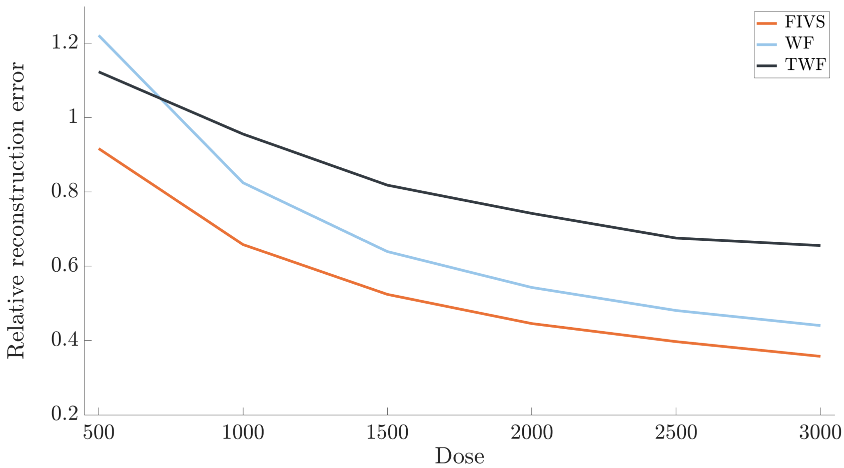


Figure 6.6.: Performance of the gradient descent algorithm for the proposed loss function based on the improved variance stabilization approach (FIVS) versus the Wirtinger flow algorithm [19] and the truncated Wirtinger flow [28]. The plot shows averages of 50 trials per choice of dose parameter.

method (FIVS) in a low-dose setting. Consequently, FIVS beats the TWF method also in performance and not only in the computational cost due to determining the truncation rule in each iteration of TWF. While the poor performance of TWF seems surprising to us, WF was not expected to work well in the low-count regime as it corresponds to the Gaussian loss without variance stabilization.

Interpreting the results discussed so far, we have to bear in mind that we experiment in the low-dose regime and hence cannot expect very small reconstruction errors without much redundancy in the data, that means a large amount of measurements. Also, to draw more measurements, the dose per measurement might need to be reduced as more measurements cause more damage. To explore the trade-off between the oversampling ratio (m/d) and the radiation dose used per object pixel $(\mathscr{d}/d)$ in greater detail, we conclude our experiments with an analysis for a larger range of oversampling-dose-pairings; see Figure 6.7. We compare the proposed algorithm, FIVS, with the gradient descent algorithm

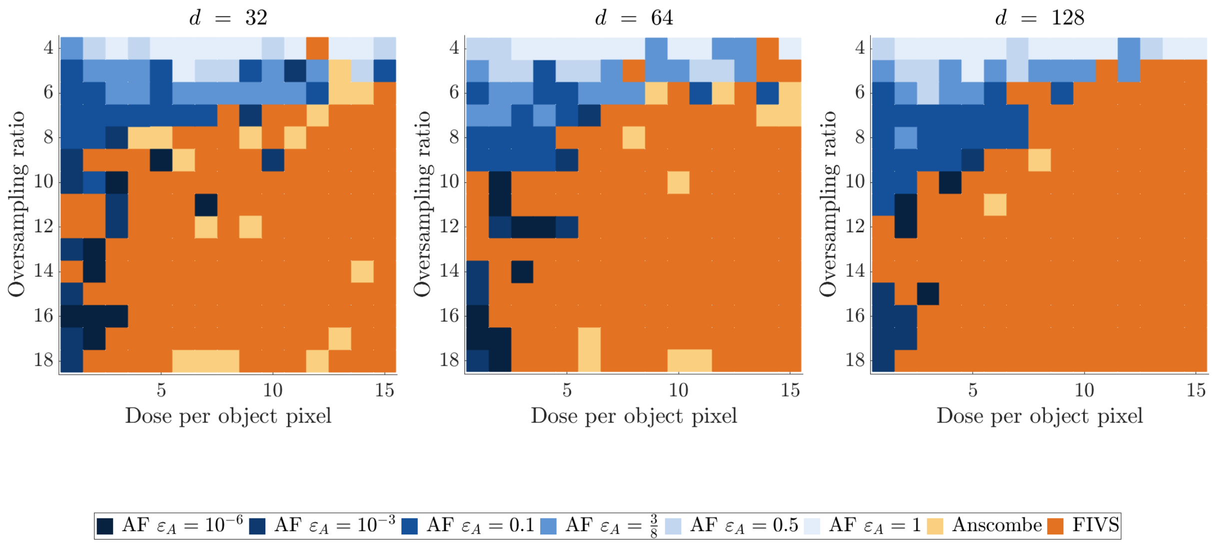


Figure 6.7.: The plot displays the best performing method per oversampling-dose-pairing, represented by the respective assigned color. The performance is measured in terms of relative reconstruction error after the described stopping criterion for the respective loss is met. All results are averaged over 5 trials. For all three object dimensions d, the proposed method FIVS showed the best performance in a majority of combinations of oversampling ratios $\frac{m}{d}$ and dose ratios $\frac{\mathscr{d}}{d}$. In the cases where AF with a certain parameter choice ε_A in (6.12) created better reconstructions, we observe the tendency that a smaller oversampling ratio requires a larger value for ε_A. Nevertheless, no definite rule can be derived for the regularization parameter of AF, whereas FIVS, a method based on a heuristic determining the parameter choice, demonstrates good performance in a wide range of cases. Moreover, the plot shows that FIVS also clearly improves upon a gradient descent for the unbiased loss (6.13) with $c = \frac{3}{8}$, which we denote here as ‘Anscombe’.

for the amplitude-based loss (6.12) for various choices of regularization parameters ε_A. This algorithm is commonly known as 'amplitude flow' (AF). Our proposed method is a variant of amplitude flow with a parameter choice based on the heuristics of the improved variance stabilization and we aim to find out empirically whether this parameter rule is indeed reasonable in a range of applications. From the experiments summarized in Figure 6.7 we derive that for a very low-dose experiment or in case of a small oversampling ratio the AF with a different parameter choice than in FIVS can be beneficial. The numerical trials suggest to choose the parameter ε_A rather large in experiments with a small oversampling ratio. However, no clear decision rule for the regularization parameter can be deduced from this empirical study, and FIVS performs best in a majority of tests.

Finally, we want to explore whether FIVS indeed refines the output of the spectral method discussed in Chapter 5. While we provided a convergence guarantee for FIVS, this covers only descent in the respective loss and it is not ensured that it produces better relative reconstruction errors when compared to the ground-truth solution. As opposed to the results reported so far, we work here with an object $x \in \mathbb{R}^{256}$, measurement vectors $a_i \sim \mathcal{N}(0, I_{256})$, $i \in [m]$, for various choices of $m \in \mathbb{N}$, and normalized the object as described in Chapter 5 to obtain measurements

$$y_i \sim Poisson\left(\kappa \left| \left\langle a_i, \frac{x}{\|x\|_2} \right\rangle \right|^2 \right), \quad i \in [m], \tag{6.19}$$

for a range of dose parameters $\kappa > 0$. The results of the experiment are shown in Figure 6.8. We observe that the reconstruction error of the output of the spectral method decreases with higher dose and oversampling ratio values, as predicted by Theorem 5.1.1.

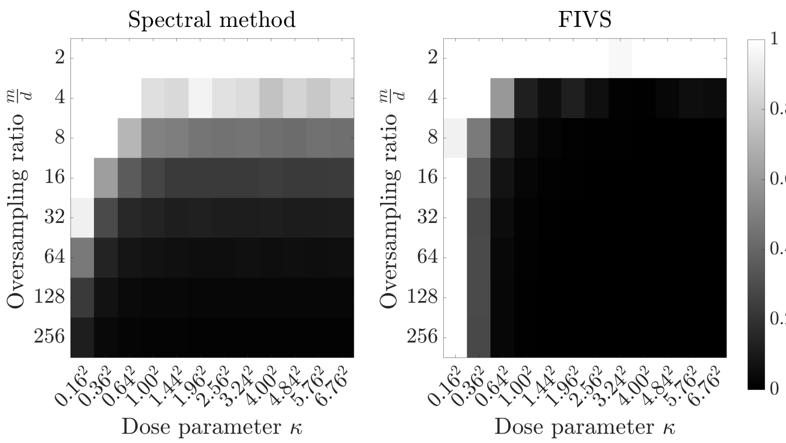


Figure 6.8.: The plot shows the dependence of the relative reconstruction error on the dose and the oversampling ratio for the Poisson measurement model (6.19). Left: the output of the spectral method with renormalization to κ; Right: the spectral method with subsequent 100 iterations of FIVS. The plots show averages over 10 random trials.

Note that the experiment indicates satisfactory reconstruction already for much lower oversampling ratios than suggested by the theoretical guarantee. Nevertheless, running iterations of FIVS initialized with the approximation obtained by the spectral method indeed has the capacity to enhance the reconstruction. We note, however, that if the dose parameter is very small, FIVS deteriorates the reconstruction result in terms of the relative reconstruction error. Recall that FIVS is proven to converge to a stationary point of the non-convex loss (6.14) and hence can also get stuck in a local minimum that is better than the initialization when evaluating the measurement metric (6.14) but not necessarily in the metric $dist(x, x_K)$.

The results in both Figure 6.7 and Figure 6.8 argue in favor of FIVS in moderately low-dose and moderately low oversampling ratio cases. However, for the extreme low-dose and low oversampling cases, we aim to find better methods later in Chapter 7.

We conclude this section with a short discussion of the step size selection for the employed algorithms. For all algorithms based on loss functions that fall within Theorem 6.4.1 or Theorem 6.4.2 we used the step sizes proposed by the theoretical results. For WF we use the step size which was proposed in [29] for loss functions (6.2), and for TWF the step size as suggested by the authors of [28]. As an alternative for the Poisson flow, [77] suggest to use a step size based on an approximation of the Hessian using observed Fisher information as it is larger than the step size involving the inverse of the factor $1 + y_i/(8\varepsilon_P)$ and hence could accelerate the convergence. However, this approach involves a line search in each iteration and hence is computationally expensive. Here, we face a trade-off between a constant step size that guarantees convergence and a step size rule that can be made independent of ε_P by performing a step size optimization in each iteration. Due to the computational advantage, we chose the constant step size. In our numerical trials we found that in case of larger parameters ε_P the step size proposed by [77], in contrary to our step size rule, does not ensure descent of the loss in each iteration. While our step size rule becomes impractical in the limit $\varepsilon_P \to 0$, the adaptively chosen step size becomes more useful as it does not decrease on the order of ε_P. On the other hand, our experiments showed that it is not advisable to choose ε_P extremely small. We conclude that the constant step size in combination with ε_P sufficiently large works well in a large range of applications. Beyond that, we emphasize again that this trade-off problem is avoided by using the Gaussian log-likelihood loss after variance stabilization, such as in the proposed FIVS algorithm.

Notes and references. *All experiments reported in this section were performed in MATLAB on a MacBook Pro equipped with a 2GHz Intel Core i5 and 16 GB of memory. The code for the reported experiments is available at [111].*

Note that in [33] we had used a power method for the spectral initialization, resulting in slightly different but comparable results.

In terms of applying the discussed approaches to real-world data, a suitable regularization term should be incorporated in the optimization problem. This could be for example *of the type discussed in [73] or in [78], depending on the type of object under consideration. It remains future work to find an appropriate problem formulation for objects as, for example, viruses that are typically imaged with low-dose radiation.*

The problem setting considered over the whole chapter was inspired by discussions with Max Leo Leidl and Knut Müller-Caspary. While we focused on analyzing the different loss functions in this thesis, our colleagues contributed a simulation study for scanning transmission electron microscopy data involving these loss functions; see [74].

7. Extremely low-dose Poisson phase retrieval

In some imaging experiments, merely reducing the dose to a moderate level is not sufficient and we have to face data of extremely low signal-to-noise ratio. This is for example the case if the detector pixels have a dead time, barely allowing for measurements larger than one.

This raises the question how the so far studied algorithms perform in such extreme cases. The numerical experiments presented in the preceding chapter demonstrated poor results in the extreme low-dose regime, asking for improvement. In this chapter, we first study how severely the spectral method is affected if the Poisson measurements are truncated to at maximum one counts, and then design new algorithmic approaches that enhance the reconstruction from extremely low-dose data.

7.1. Spectral method for Bernoulli phase retrieval

As explained in Section 3.2, Poisson measurements truncated to $\{0,1\}$ values can be modeled by Bernoulli random variables. Motivated by this, we analyze the spectral method for measurements

$$y_i \sim \mathit{Bernoulli}\left(1 - \exp(-|\langle a_i, x\rangle|^2)\right), \quad i \in [m].$$

We work with the phase retrieval problem for $\mathbb{K} = \mathbb{R}$ and with independent standard Gaussian random vectors a_i, $i \in [m]$. Based on the spectral method, we prove a reconstruction guarantee for Bernoulli phase retrieval and establish the required trade-off between the oversampling ratio and the dose parameter.

The results presented in the following are based on joint work with Sjoerd Dirksen, Felix Krahmer and Palina Salanevich; see also [36, 37].

As in Chapter 5, we presume that the ground-truth object $x \in \mathbb{R}^d$ is known to satisfy $\|x\|_2^2 = \kappa$ with dose parameter $\kappa > 0$, so that we consider measurements

$$y_i \sim \mathit{Bernoulli}\left(1 - \exp\left(-\kappa \left|\left\langle a_i, \frac{x}{\|x\|_2}\right\rangle\right|^2\right)\right), \quad i \in [m]. \tag{7.1}$$

Let $x_0 \in \mathbb{R}^d$ be the output of Algorithm 1, that means, x_0 is the leading eigenvector of

$$S := \frac{1}{m}\sum_{i=1}^{m} y_i a_i a_i^{\mathrm{T}},$$

renormalized such that $\|x_0\|_2^2 = \kappa$.

We show that this vector x_0 approximates the ground-truth object x up to a small reconstruction error provided that m is sufficiently large. This result is based on two technical lemmas that are proven in the Appendix, more precisely in Lemma A.11 and Lemma A.13.

Theorem 7.1.1. *Fix $x \in \mathbb{R}^d$, let $m \geq d\log(d)$ and $y_i \sim$ Bernoulli $\left(1 - \exp\left(-|\langle a_i, x\rangle|^2\right)\right)$, $i \in [m]$, with independent random vectors $a_i \sim \mathcal{N}(0, I_d)$, $i \in [m]$. Then, there exist absolute constants $C, c_1, c_2 > 0$, such that, with probability at least*

$$1 - \frac{2}{d} - 8m\exp\left(-c_1 d\right) - 2\exp\left(-c_2 d\log(d)\right),$$

the vector x_0 satisfies

$$\frac{\mathit{dist}(x, x_0)^2}{\|x\|_2^2} \leq 2 - 2\sqrt{1 - C \cdot \frac{(2\kappa+1)^{\frac{3}{2}}}{\kappa} \cdot \sqrt{\log(d)}\sqrt{\frac{d}{m}}}\ .$$

Proof. According to Lemma A.11, it is

$$\mathbb{E}\left[S\right] = \frac{2}{(2\|x\|_2^2+1)^{\frac{3}{2}}} \cdot xx^{\mathrm{T}} + \left(1 - \frac{1}{(2\|x\|_2^2+1)^{\frac{1}{2}}}\right) \cdot I_d.$$

Let λ_0 be the largest eigenvalue of S. Then, $\kappa\lambda_0 = x_0^{\mathrm{T}} S x_0$ and thus

$$\begin{aligned}
&\left|\kappa\lambda_0 - \frac{2}{(2\|x\|_2^2+1)^{\frac{3}{2}}} \cdot |\langle x, x_0\rangle|^2 - \left(1 - \frac{1}{(2\|x\|_2^2+1)^{\frac{1}{2}}}\right)\|x_0\|_2^2\right| \\
&= \left|x_0^{\mathrm{T}} S x_0 - x_0^{\mathrm{T}}\left(\frac{2}{(2\|x\|_2^2+1)^{\frac{3}{2}}} \cdot xx^{\mathrm{T}} + \left(1 - \frac{1}{(2\|x\|_2^2+1)^{\frac{1}{2}}}\right) \cdot I_d\right) x_0\right| \\
&= \left|x_0^{\mathrm{T}}\left(S - \mathbb{E}\left[S\right]\right) x_0\right| \\
&\leq \|x_0\|_2^2 \left\|S - \mathbb{E}\left[S\right]\right\|.
\end{aligned}$$

Let $\rho := \|S - \mathbb{E}\left[S\right]\|$. Since x_0 is normalized to $\|x_0\|_2^2 = \kappa$, we obtain

$$|\langle x_0, x\rangle|^2 \geq \frac{(2\kappa+1)^{\frac{3}{2}}}{2}\left(\kappa\lambda_0 - \left(1 - \frac{1}{(2\kappa+1)^{\frac{1}{2}}}\right)\kappa - \kappa\rho\right).$$

Furthermore, in Theorem 5.1.1 we have shown that the largest eigenvalue of S satisfies

$$\begin{aligned}\lambda_0 &\geq -\rho + \frac{1}{\|x\|_2^2} \cdot x^{\mathrm{T}} \mathbb{E}\left[S\right] x \\ &= -\rho + \kappa \cdot \frac{2}{(2\kappa+1)^{\frac{3}{2}}} + \left(1 - \frac{1}{(2\kappa+1)^{\frac{1}{2}}}\right),\end{aligned}$$

so that

$$\begin{aligned}\left|\langle x_0, x\rangle\right|^2 &\geq \frac{(2\kappa+1)^{\frac{3}{2}}}{2} \left(\kappa^2 \cdot \frac{2}{(2\kappa+1)^{\frac{3}{2}}} - 2\kappa\rho\right) \\ &= \kappa^2 - \kappa\left(2\kappa+1\right)^{\frac{3}{2}} \rho.\end{aligned}$$

Recall from the proof of Theorem 5.1.1 that

$$\mathit{dist}(x_0, x)^2 = 2\kappa - 2\left|\langle x_0, x\rangle\right|.$$

Hence,

$$\frac{\mathit{dist}(x_0, x)^2}{\kappa} \leq 2 - 2\sqrt{1 - \frac{(2\kappa+1)^{\frac{3}{2}}}{\kappa} \cdot \rho}\ . \tag{7.2}$$

We apply Lemma A.13 to obtain that there exist absolute constants $C, c_1, c_2 > 0$, such that, with probability at least

$$1 - \frac{2}{d} - 8m\exp\left(-c_1 d\right) - 2\exp\left(-c_2 d\log(d)\right),$$

it holds that

$$\rho \leq C\sqrt{\log(d)}\sqrt{\tfrac{d}{m}}.$$

Inserting this into (7.2) completes the proof. □

Based on Theorem 7.1.1, we find a rule on the sample complexity required for successful reconstruction.

Corollary 7.1.2. *From Theorem 7.1.1, we conclude that*

$$m = \mathcal{O}\left(d\log(d) \cdot \frac{(2\kappa+1)^3}{\kappa^2}\right)$$

measurements are required to recover the ground-truth up to a small constant relative error from Bernoulli measurements (7.1) *by solving Algorithm 1 with proper normalization.*

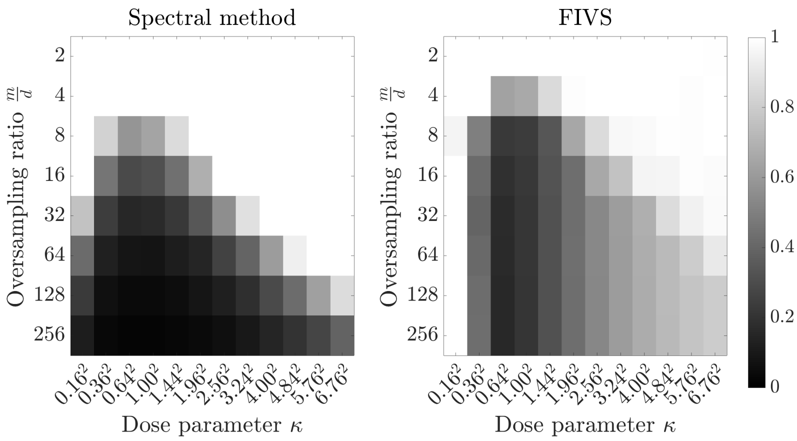


Figure 7.1.: The plot shows the dependence of the relative reconstruction error on the dose and the oversampling ratio for the Bernoulli measurement model (7.1). Left: the output of the spectral method with renormalization to κ; Right: the spectral method with subsequent 100 iterations of FIVS. The plots show averages over 10 random trials.

We note that the factor $\frac{(2\kappa+1)^3}{\kappa^2}$ is minimal for $\kappa = 1$ and its contribution can be interpreted as follows. For very small, as well as for rather large values of κ, we need a lot of oversampling. In the former case, this is due to the small amount of information captured by the measurements. This phenomenon is naturally the same for general Poisson observations, hidden in the factor $\left(1 + \frac{1}{\kappa}\right)^2$ in the sample complexity (5.3). In the latter case, more measurements are required due to a loss of information because of the truncation of observations larger than 1 in the Bernoulli model.

We conclude with a numerical experiment, investigating the dose versus oversampling trade-off of the spectral method for Bernoulli phaseless measurements. Again, we compare the output also with a subsequent application of the FIVS method discussed in Section 6.5. We experiment with an object $x \in \mathbb{R}^{256}$, measurement vectors $a_i \sim \mathcal{N}(0, I_{256}),\ i \in [m]$, and measurements (7.1). The results of the experiment are reported in Figure 7.1. As predicted by Theorem 7.1.1, the reconstruction error decreases with sufficient oversampling and the experiment indicates that reconstruction is possible for already much lower oversampling ratios than theoretically postulated. However, we see that the reconstruction error deteriorates when the dose κ gets large, indicating that too much information is lost due to the truncation to $\{0, 1\}$ measurements. Subsequent iterations of FIVS improve the reconstruction here only in the low oversampling ratio regime and for κ sufficiently large, otherwise the spectral initialization alone shows better performance. As already mentioned in Section 6.5, this can be explained by the fact that FIVS is proven to converge to a stationary point of a non-convex measurement error metric and hence can deteriorate the estimate in terms of the relative reconstruction error if the initialization falls within the basin of attraction of a poor local minimum.

Motivated by the observation that in any case a sample size larger than $4d$ was required to obtain reasonable reconstructions, we develop new techniques for the extreme low-dose regime in the next section and will find that these enhance upon the spectral method and FIVS in experiments with small sample complexity.

Notes and references. *To the best of our knowledge, we were the first to prove a reconstruction guarantee for Bernoulli phase retrieval.*

[27] provided a theoretical guarantee for phase retrieval from one-bit data, also with potential bit flips, that means noise in the quantization. However, their result is restricted to an amount of potential bit flips that does not apply to the more extreme case of Bernoulli random variables.

An extension of our result to $\mathbb{K} = \mathbb{C}$ is yet to be determined.

7.2. A one-bit quantization approach

The empirical tests depicted in Figure 7.1 reveal potential for refinement for phase retrieval in the extreme low-dose scenario and in case of a truncation of the measurements to only binary 'count/no count' data. We aim to develop algorithms that enhance the reconstruction quality in case of a very small dose and a reasonably small oversampling ratio. In the following, we propose some novel approaches to this problem, exploiting the connection to the problem of one-bit quantization.

In the following, we can work with $\mathbb{K} = \mathbb{C}$. Once again, we aim to solve the inverse problem of recovering the ground-truth $x \in \mathbb{C}^d$ from phaseless measurements

$$y_i \sim \mathit{Poisson}\left(|\langle a_i, x\rangle|^2\right), \quad i \in [m].$$

In this section, we focus on the regime that the y_i are in $\{0, 1\}$ with high probability, and replace them by their truncated representation $\min\{y_i, 1\}$. In general, the algorithms proposed over the following subsections can be applied for any measurement frame. In some instances, we provide exemplary elaborations of parts of the algorithmic designs for either Gaussian random measurement vectors or ptychographic measurements.

This section introduces joint work with Felix Krahmer. A first discussion of these results was published in [113].

7.2.1. Understanding low-count Poisson data as one-bit data

If the ground-truth measurements $|\langle a_i, x\rangle|^2$ are so small that $y_i \in \{0, 1\}$ with high probability, we are (almost) in the extreme scenario of one-bit data. Moreover, attributing values in $\mathbb{N}$ to ground-truth measurements $|\langle a_i, x\rangle|^2 \in \mathbb{R}_{\geq 0}$ according to the Poisson distribution can be understood as a quantization of the measurements to values in $\mathbb{N}$. Combining these two observations, we decide to associate low-count Poisson measurements as one-bit quantized data.

Definition 7.2.1. *The one-bit quantization scheme for a value $\alpha \in \mathbb{R}$ according to a deterministic threshold $\tau \in \mathbb{R}$ generates values in $\{0,1\}$ with the following rationale:*

$$q_i = \begin{cases} 0, & \textit{if } \alpha < \tau, \\ 1, & \textit{if } \alpha \geq \tau. \end{cases}$$

Hence, according to the one-bit quantization scheme with threshold $\tau \in \mathbb{R}$, we assign the following bits in $\{0,1\}$ to $|\langle a_i, x\rangle|^2$:

$$q_i = \begin{cases} 0, & \text{if } |\langle a_i, x\rangle|^2 < \tau, \\ 1, & \text{if } |\langle a_i, x\rangle|^2 \geq \tau. \end{cases}$$

Equivalently, the quantizer q_i can be computed by

$$q_i = \tfrac{1}{2}\left(\operatorname{sgn}\!\left(|\langle a_i, x\rangle|^2 - \tau\right) + 1\right). \tag{7.3}$$

With the aim of looking at the problem from the perspective of one-bit quantization, we truncate potentially larger Poisson measurements with threshold 1 and consider

$$\hat{y}_i := \min\{y_i, 1\}, \quad i \in [m].$$

Then, each measurement is represented by a random bit, and we want to interpret the Poisson measurements $\hat{y}_i$ as the result of one-bit quantization of $|\langle a_i, x\rangle|^2$. However, due to the randomness of the Poisson noise, this measurement model exhibits random bit flips. We handle this by selecting the quantization parameter τ such that the quantization approach chooses the bit from $\{0,1\}$ that is the more likely outcome of the Poisson random variable for any ground-truth $|\langle a_i, x\rangle|^2$. In an appropriate random measurement setting, this is almost surely achieved with the threshold $\tau = \log(2)$.

Lemma 7.2.1. *Let a_i, $i \in [m]$, be independent standard Gaussian random vectors. The only threshold that yields $\mathbb{P}\left(\hat{y}_i = q_i \;\middle|\; |\langle a_i, x\rangle|^2 \right) > 0.5$ almost surely is $\tau = \log(2)$.*

Proof. Given an expected value $|\langle a_i, x\rangle|^2$, the probability to observe $y_i = k$ equals

$$\mathbb{P}\left(y_i = k \;\middle|\; |\langle a_i, x\rangle|^2 \right) = \tfrac{1}{k!}\left(|\langle a_i, x\rangle|^2\right)^k \exp\left(-|\langle a_i, x\rangle|^2\right).$$

Then, for a truncated measurement $\hat{y}_i$ it holds

$$\mathbb{P}\left(\hat{y}_i = 0 \;\middle|\; |\langle a_i, x\rangle|^2 \right) = \exp\left(-|\langle a_i, x\rangle|^2\right)$$

and

$$\mathbb{P}\left(\hat{y}_i = 1 \;\middle|\; |\langle a_i, x\rangle|^2 \right) = 1 - \exp\left(-|\langle a_i, x\rangle|^2\right).$$

Let $\tau = \log(2)$. In the event that $|\langle a_i, x\rangle|^2 < \log(2)$ it is $q_i = 0$ and

$$\mathbb{P}\big(\ \hat{y}_i = q_i \ \big|\ |\langle a_i, x\rangle|^2\ \big) = \mathbb{P}\big(\ \hat{y}_i = 0 \ \big|\ |\langle a_i, x\rangle|^2\ \big) > 0.5.$$

Analogously, in the event that $|\langle a_i, x\rangle|^2 > \log(2)$ it is also $\mathbb{P}\big(\ \hat{y}_i = q_i \ \big|\ |\langle a_i, x\rangle|^2\ \big) > 0.5$.

Since the event that $|\langle a_i, x\rangle|^2 = \log(2)$ appears with probability 0, we found that $\mathbb{P}\big(\ \hat{y}_i = q_i \ \big|\ |\langle a_i, x\rangle|^2\ \big) > 0.5$ almost surely for $\tau = \log(2)$.

For $\tau \neq \log(2)$, the same calculations yield that $\mathbb{P}\big(\ \hat{y}_i = q_i \ \big|\ |\langle a_i, x\rangle|^2\ \big) < 0.5$ if $|\langle a_i, x\rangle|^2 \in (\tau, \log(2))$ or $|\langle a_i, x\rangle|^2 \in (\log(2), \tau)$, respectively. □

Note that this is not restricted to standard Gaussian random vectors. This type of measurement system was only chosen as an example for which it is guaranteed that the event $|\langle a_i, x\rangle|^2 = \log(2)$ appears with probability 0.

For a broader generality and a more convenient notation, we keep the parameter τ arbitrary in the following, but use the value $\tau = \log(2)$ in the numerical experiments.

From now on, we work with (truncated) Poisson random variables but address them with algorithms designed for one-bit quantization. That is, instead of solving the system of phaseless equations as introduced in Section 2.3 by fitting the noisy measurements y_i, we now want to find $z \in \mathbb{C}^d$ that solves the system

$$\hat{y}_i = \tfrac{1}{2}\left(\operatorname{sgn}\left(|\langle a_i, z\rangle|^2 - \tau\right) + 1\right), \quad i \in [m], \tag{7.4}$$

to obtain an approximation of the ground-truth x. We expect this to be less prone to noise overfitting as we only aim to align the signs of $|\langle a_i, z\rangle|^2 - \tau$ to the measurements $\hat{y}_i$, which gives more flexibility to the values of $|\langle a_i, z\rangle|^2$.

7.2.2. Constrained nuclear norm minimization

A first approach to solve the problem (7.4) could be to work with the PhaseLift method as explained in Section 4.3. Be reminded that this involves the linear operator $\mathcal{A} : \mathbb{C}^{d\times d} \to \mathbb{R}^m,\ X \mapsto (\langle a_i a_i^*, X\rangle_{HS})_{i\in[m]}$, satisfying $\mathcal{A}(xx^*) = |\langle a_i, x\rangle|^2$.

Without necessarily taking the low-count model or the reformulated problem (7.4) into account, we can obtain a version of (4.9) suitable for Poisson measurements as follows. For a vector of random variables $y = (y_i)_{i\in[m]}$ with independent entries $y_i \sim \mathit{Poisson}\left(|\langle a_i, x\rangle|^2\right)$ it holds that

$$\mathbb{E}_y \left\|\mathcal{A}(xx^*) - y\right\|_2^2 = \sum_{i=1}^m \mathbb{E}_{y_i}\left[\left(|\langle a_i, x\rangle|^2 - y_i\right)^2\right] = \sum_{i=1}^m |\langle a_i, x\rangle|^2 = \mathbb{E}_y\left[\sum_{i=1}^m y_i\right],$$

where we use

$$\mathbb{E}_{y_i}[y_i] = \mathbb{V}_{y_i}[y_i] = |\langle a_i, x\rangle|^2,$$

and
$$|\langle a_i, x\rangle|^4 = \mathbb{E}_{y_i}\left[y_i\right]^2 = \mathbb{E}_{y_i}\left[y_i^2\right] - \mathbb{V}_{y_i}\left[y_i\right] = \mathbb{E}_{y_i}\left[y_i^2\right] - |\langle a_i, x\rangle|^2 .$$
This suggests to use a threshold $\varepsilon = \sum_{i=1}^m y_i$ in (4.9), resulting in the constrained nuclear norm minimization problem

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & \|\mathcal{A}(X) - y\|_2^2 \le \sum_{i=1}^m y_i, \\ & X \succeq 0. \end{aligned} \tag{7.5}$$

However, if $y_i \in \{0,1\}$ for all $i \in [m]$, the zero matrix $O \in \mathbb{C}^{d\times d}$ satisfies the constraint in (7.5) since

$$\sum_{i=1}^m y_i = \sum_{i=1}^m y_i^2 = \|y\|_2^2 = \|\mathcal{A}(O) - y\|_2^2 .$$

Moreover, the zero matrix is of minimal nuclear norm. Consequently, problem (7.5) would always output the zero matrix in this scenario.

Hence, for the case of extremely low-count data, we aim to derive another feasibility constraint. Thereto, we use the reinterpretation of the low-count Poisson measurements as one-bit data and study the system (7.4).

If (7.4) is satisfied for all $i \in [m]$ by the ground-truth solution x, it holds that $2\hat{y} - 1 = \operatorname{sgn}\left(\mathcal{A}(xx^*) - \tau\right)$, hence

$$\begin{aligned} \langle 2\hat{y} - 1,\ \mathcal{A}(xx^*) - \tau\rangle &= \langle \operatorname{sgn}\left(\mathcal{A}(xx^*) - \tau\right),\ \mathcal{A}(xx^*) - \tau\rangle \\ &= \|\mathcal{A}(xx^*) - \tau\|_1 . \end{aligned}$$

Here, subtraction of a number from a vector and the signum of a vector are meant as entry-wise actions. If $\hat{y}$ is expected to approximate the ground-truth measurements $\mathcal{A}(xx^*)$ sufficiently well, we can introduce

$$\langle 2\hat{y} - 1,\ \mathcal{A}(X) - \tau\rangle = \|\hat{y} - \tau\|_1$$

as a constraint in the optimization problem.

Moreover, the Poisson random variables satisfy in expectation

$$\begin{aligned}\mathbb{E}_y \langle 2y-1,\ \mathcal{A}(xx^*)-\tau\rangle &= \sum_{i=1}^{m} (2\mathbb{E}_{y_i}[y_i]-1)\left(|\langle a_i,x\rangle|^2-\tau\right)\\ &= \sum_{i=1}^{m} 2|\langle a_i,x\rangle|^4-(2\tau+1)|\langle a_i,x\rangle|^2+\tau\\ &= \sum_{i=1}^{m} 2\mathbb{E}_{y_i}\left[y_i^2\right]-(2\tau+3)\mathbb{E}_{y_i}[y_i]+\tau,\end{aligned}$$

where we either fix a set of measurement vectors a_i or consider the expectation conditionally on the a_i if they are chosen at random. Motivated by this relation, we suggest to introduce also the constraint

$$\langle 2\hat{y}-1,\ \mathcal{A}(X)-\tau\rangle = \sum_{i=1}^{m} 2y_i^2-(2\tau+3)y_i+\tau$$

into the optimization problem.

Combining these two properties, we propose to consider a weighted version of the two conditions and work with

$$\begin{aligned}&\text{minimize} && \|X\|_* \\ &\text{subject to} && \langle 2\hat{y}-1,\ \mathcal{A}(X)-\tau\rangle = c,\\ &&& X\succeq 0,\end{aligned} \tag{7.6}$$

with $c=\alpha\cdot\sum_{i=1}^{m}(2y_i^2-(2\tau+3)y_i+\tau)+(1-\alpha)\cdot\|\hat{y}-\tau\|_1$ for $0.5\le\alpha\le 1$. Based on numerical trials, we decide to choose α as 0.5 if $\hat{y}=y$, and closer to 1 the more entries of y are larger than 1 and hence were truncated in $\hat{y}$. This heuristics of introducing the second constraint was motivated by our observation in empirical trials that the firstly derived constraint exhibits a bias towards too large values.

7.2.3. Projected subgradient descent

Our second approach to solving the system of equations (7.4) is inspired by a projected subgradient descent for one-bit compressed sensing; see [64, 67]. Again, we work with the lifted and thus linearized version of the phase retrieval problem, using the operator $\mathcal{A}:\mathbb{C}^{d\times d}\to\mathbb{R}^m, X\mapsto(\langle a_ia_i^*,X\rangle_{HS})_{i\in[m]}$.

If $z\in\mathbb{C}^d$ satisfies (7.4) for some $i\in[m]$, the observation $\hat{y}_i$ is the correct one-bit quantization of $|\langle a_i,z\rangle|^2$. In this case,

$$2\hat{y}_i-1=\operatorname{sgn}\left(\langle a_ia_i^*,zz^*\rangle_{HS}-\tau\right),$$

so that

$$(2\hat{y}_i - 1) \cdot (\langle a_i a_i^*, zz^* \rangle_{HS} - \tau) = |\langle a_i a_i^*, zz^* \rangle_{HS} - \tau| \geq 0.$$

Aiming to find a solution of (7.4), we are required to minimize the contribution of falsely quantized entries, that means, components that satisfy

$$(2\hat{y}_i - 1) \cdot (\langle a_i a_i^*, zz^* \rangle_{HS} - \tau) < 0.$$

To this end, we utilize the one-sided loss function

$$\mathcal{L}(X) = \sum_{i=1}^{m} \left[-(2\hat{y}_i - 1) \cdot (\langle a_i a_i^*, X \rangle_{HS} - \tau) \right]_+ ,$$

where $[v]_+ := \max\{v, 0\}$ for $v \in \mathbb{R}$.

Theorem 7.2.2. *The loss function $\mathcal{L}$ is convex and*

$$\nabla\mathcal{L}(X) := \sum_{i=1}^{m} \frac{1}{2} \left(\operatorname{sgn}\left(\langle a_i a_i^*, X \rangle_{HS} - \tau \right) - (2\hat{y}_i - 1) \right) a_i a_i^*$$

is a subgradient of $\mathcal{L}$.

Proof. It is $\mathcal{L}(X) = \sum_{i=1}^{m} \mathcal{L}_i(X)$ with convex functions

$$\mathcal{L}_i(X) = \begin{cases} |\langle a_i a_i^*, X \rangle_{HS} - \tau| , & \text{if } \operatorname{sgn}\left(\langle a_i a_i^*, X \rangle_{HS} - \tau \right) \neq \operatorname{sgn}\left(2\hat{y}_i - 1 \right), \\ 0, & \text{otherwise.} \end{cases}$$

For every $\mathcal{L}_i$, we individually determine the subdifferential. If $X \in \mathbb{C}^{d\times d}$, we work with Wirtinger gradients; see Appendix B. For an introduction to matrix Wirtinger calculus we refer the interested reader to [71]. Essentially, the common differentiation rules apply and we perform the computations as usual.

If $\langle a_i a_i^*, X \rangle_{HS} - \tau \neq 0$, the gradient of $\mathcal{L}_i$ is

$$\begin{aligned} \nabla\mathcal{L}_i(X) &= \begin{cases} \operatorname{sgn}\left(\langle a_i a_i^*, X \rangle_{HS} - \tau \right) a_i a_i^*, & \text{if } \operatorname{sgn}\left(\langle a_i a_i^*, X \rangle_{HS} - \tau \right) \neq \operatorname{sgn}\left(2\hat{y}_i - 1 \right), \\ 0, & \text{otherwise,} \end{cases} \\ &= \frac{1}{2} \left(\operatorname{sgn}\left(\langle a_i a_i^*, X \rangle_{HS} - \tau \right) - \operatorname{sgn}\left(2\hat{y}_i - 1 \right) \right) a_i a_i^*. \end{aligned}$$

For $\langle a_i a_i^*, X \rangle_{HS} - \tau = 0$, we can consider the subdifferential set

$$\mathcal{S}_i := \left\{ \gamma \cdot \frac{1}{2} \left(\operatorname{sgn}\left(\langle a_i a_i^*, X \rangle_{HS} - \tau \right) - \operatorname{sgn}\left(2\hat{y}_i - 1 \right) \right) a_i a_i^* : \gamma \in [0, 1] \right\}.$$

Since $\frac{1}{2}\left(\operatorname{sgn}\left(\langle a_i a_i^*, X\rangle_{HS} - \tau\right) - \operatorname{sgn}\left(2\hat{y}_i - 1\right)\right) a_i a_i^* \in \mathcal{S}_i$, we found that

$$\frac{1}{2}\left(\operatorname{sgn}\left(\langle a_i a_i^*, X\rangle_{HS} - \tau\right) - \left(2\hat{y}_i - 1\right)\right) a_i a_i^*$$

is a subgradient of $\mathcal{L}_i$, and since this holds for all $i \in [m]$,

$$\sum_{i=1}^{m} \frac{1}{2}\left(\operatorname{sgn}\left(\langle a_i a_i^*, X\rangle_{HS} - \tau\right) - \left(2\hat{y}_i - 1\right)\right) a_i a_i^*$$

indeed is a subgradient of $\mathcal{L}$. □

Consequently, a subgradient descent algorithm can be employed to find a minimizer of the loss $\mathcal{L}$.

We further use that the ground-truth solution xx^* is known to satisfy $\operatorname{rank}(xx^*) = 1$, so that the minimization of $\mathcal{L}$ can be enhanced by introducing a rank-one constraint. That means, we propose to consider the optimization problem

$$\begin{aligned} &\text{minimize} \quad && \mathcal{L}(X) \\ &\text{subject to} \quad && \operatorname{rank}(X) = 1. \end{aligned} \tag{7.7}$$

A projection onto the set of rank-one matrices, obtained by best rank-one approximation via singular value decomposition, covers the constraint. Together, the suggested algorithm is a projected subgradient descent. We summarize the procedure, involving further regularization, later in Algorithm 4.

Notes and references. *As of now, the method developed in this section lacks theoretical guarantees. During the preparation of this manuscript, a resembling approach was suggested in [27] for one-bit phaseless data, together with a recovery guarantee, albeit in a noiseless or 'bit flip free' scenario. A deeper theoretical understanding of the behavior of such methods in case of Poisson data remains a critical open question.*

7.2.4. Regularization for norm approximation

The methods discussed in Section 7.2.2 and Section 7.2.3 can be augmented with a regularization term that specifically addresses the approximation of the length of x. To this end, we again use prior knowledge about the measurements. For example, for standard Gaussian measurement vectors $a_i \sim \mathcal{N}(0, I_d)$ and random variables $y_i \sim \mathit{Poisson}\left(|\langle a_i, x\rangle|^2\right)$, it is

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} y_i\right] = \mathbb{E}_A\left[\mathbb{E}_y\left[\frac{1}{m}\sum_{i=1}^{m} y_i \;\Big|\; A\right]\right] = \mathbb{E}_A\left[\frac{1}{m}\sum_{i=1}^{m} |\langle a_i, x\rangle|^2\right]$$
$$= \frac{1}{m}\sum_{i=1}^{m} x^*\mathbb{E}_A\left[a_i a_i^*\right] x = \|x\|_2^2 = \mathrm{tr}(xx^*),$$

where A denotes the matrix with rows a_i^*. Here we use that $\mathbb{E}_{a_i}\left[a_i a_i^*\right] = I_d$ for $a_i \sim \mathcal{N}(0, I_d)$ due to $\mathbb{E}\left[a_{ij}^2\right] = 1$ for all $j \in [d]$, and $\mathbb{E}\left[a_{ij_1} a_{ij_2}\right] = 0$ for all $j_1, j_2 \in [d]$ with $j_1 \neq j_2$.

Based on this relation in expectation for the length of x, or equivalently the trace of xx^*, we suggest to equip (7.7) with an additional loss term $\left(\mathrm{tr}(X) - \frac{1}{m}\sum_{i=1}^{m} y_i\right)^2$. That means, we propose to solve the optimization problem

$$\begin{aligned} &\text{minimize} \quad \mathcal{L}(X) + \lambda\left(\mathrm{tr}(X) - \frac{1}{m}\sum_{i=1}^{m} y_i\right)^2 \\ &\text{subject to} \quad \mathrm{rank}(X) = 1, \end{aligned} \tag{7.8}$$

with a suitable regularization parameter $\lambda \geq 0$. The respective algorithm is summarized in Algorithm 4.

Algorithm 4 Projected subgradient descent for (7.8)

Input: $a_i, y_i, \hat{y}_i$ for all $i \in [m]$, step size $\mu > 0$, regularization parameter λ, stopping rule K, initialization $X_0 \in \mathbb{C}^{d\times d}$

for $k = 0$ **to** $K - 1$ **do**

$Z_{k+1} = X_k - \mu\left(\nabla\mathcal{L}(X_k) + 2\lambda I_d\left(\mathrm{tr}(X_k) - \frac{1}{m}\sum_{i=1}^{m} y_i\right)\right)$;

$X_{k+1} = \sigma_1 u_1 v_1^*$

with σ_1 first singular value, u_1 first left singular vector, and v_1 first right singular vector of Z_{k+1};

end

Output: X_K

Note that for the length approximation it is more reasonable to involve the measurements y instead of the truncated version $\hat{y}$. Hence, (7.8) combines the loss aiming to solve the sign equations (7.4) with a fitting of the non-truncated measurements. If $\hat{y} \neq y$ in case of $\max_{i\in[m]} y_i > 1$, this introduces the advantage that also the information contained in the larger measurements is utilized which was discarded in approaches as (7.7).

The same regularization term for norm approximation can also be implemented into the linear program (7.6). The respective optimization problem then reads as

$$\begin{aligned} \text{minimize} \quad & \|X\|_* + \tilde{\lambda}\left(\operatorname{tr}(X) - \frac{1}{m}\sum_{i=1}^{m} y_i\right)^2 \\ \text{subject to} \quad & \langle 2\hat{y} - 1,\ \mathcal{A}(X) - \tau\rangle = c, \\ & X \succeq 0, \end{aligned} \tag{7.9}$$

with a suitable regularization parameter $\tilde{\lambda} \geq 0$.

Up to here, we exemplarily considered the norm approximation for Gaussian measurement vectors. All other relations we built the algorithms so far upon were not limited to this case, and also the approach derived for approximating the object norm is not restricted to this setting. For a better intuition on how this can be generalized to other problem settings, let us make a small detour to the problem of ptychography. If the measurements are the output of a ptychographic experiment with Poisson noise, that is, if we have data $Y_{k,r} \sim \mathit{Poisson}\big(|\sum_{t=0}^{d-1} e^{-\frac{2\pi i t k}{d}} x_{t+r} w_t|^2\big)$ for $k, r \in [d]_0$, we find that

$$\begin{aligned} \mathbb{E}\left[\frac{1}{d}\sum_{k,r=0}^{d-1} Y_{k,r}\right] &= \frac{1}{d}\sum_{k,r=0}^{d-1}\sum_{t_1,t_2=0}^{d-1} e^{-\frac{2\pi i (t_1 - t_2) k}{d}} x_{t_1+r}\overline{x}_{t_2+r} w_{t_1}\overline{w}_{t_2} \\ &= \frac{1}{d}\sum_{r=0}^{d-1}\sum_{t_1,t_2=0}^{d-1} x_{t_1+r}\overline{x}_{t_2+r} w_{t_1}\overline{w}_{t_2}\sum_{k=0}^{d-1} e^{-\frac{2\pi i (t_1 - t_2) k}{d}} \\ &= \frac{1}{d}\sum_{r=0}^{d-1}\sum_{t=0}^{d-1} |x_{t+r}|^2 |w_t|^2 \cdot d \\ &= \sum_{t=0}^{d-1}\left(\sum_{r=0}^{d-1} |x_{t+r}|^2\right)|w_t|^2 \\ &= \|w\|_2^2 \|x\|_2^2 . \end{aligned}$$

With this expected relation between the measurements and the object norm, we can formulate optimization problems as (7.8) and (7.9) also for the problem of ptychography.

Notes and references. *In one-bit compressed sensing, it is necessary to specifically incorporate a procedure for norm approximation since quantizing linear measurements to their sign information discards all norm information. Typically, this issue is addressed by so-called dithering; see for example [68]. That means, one introduces a perturbation before quantization, similar to the threshold τ in the quantization scheme* (7.3)*. However, to achieve norm reconstruction in the compressed sensing problem, the dithering should be chosen at random. Opposed to this, the here considered quantization of phaseless measurements automatically comes with a threshold and thus inherently affects norm estimation.*

This was noted also in [27]. Nevertheless, integrating the developed regularization for norm approximation entails further improvement. Beyond that, it is an interesting question for future work whether introducing randomness in the quantization scheme (7.3) *could enhance the problem.*

7.2.5. Numerical experiments

We conclude this section with a report on numerical experiments that showcase the performance of the developed algorithms in comparison to benchmark methods discussed in [77, 81] as well as in Chapter 5, Chapter 6 and Section 7.1. As in the preceding chapters, we experiment with synthetic data simulated using Gaussian measurement vectors. However, the same experiments can also be performed for ptychographic data, resulting in similar conclusions.

In the following, we work with test objects $x \in \mathbb{C}^d$ with $d = 25$, and complex Gaussian measurement vectors $a_i \in \mathbb{C}^d,\ i \in [m]$, where we choose a relatively low oversampling scenario $m = 4d$.

We normalize the measurements to satisfy $\sum_{i=1}^m |\langle a_i, x\rangle|^2 = 1$ so that a noiseless value $|\langle a_i, x\rangle|^2$ can be interpreted as the probability of a particle hitting the ith detector pixel. Then we choose a dose $\mathscr{d} > 0$ and work with measurements y_i being realizations of random variables following the distributions $Poisson\big(\mathscr{d} \cdot |\langle a_i, x\rangle|^2\big),\ i \in [m]$. For a smaller utilized dose $\mathscr{d}$, measurements larger than 1 appear less frequently, meaning also that the one-bit quantization model approximates the Poisson model better.

Aiming to compare the performance of Algorithm 4 and other iterative phase retrieval algorithms, we require a reasonably good choice of initialization.

The initialization method commonly used in the phase retrieval community is the spectral method that we introduced in Section 4.1 and analyzed for Poisson and Bernoulli measurements in Chapter 5 and Section 7.1. We observed earlier that it requires sufficient oversampling. A slight adaptation of the spectral method was developed for Poisson measurements in [81], described as an optimized spectral method in terms of an asymptotic consideration. However, our numerical trials did reveal that in the extremely low-dose and low oversampling case there is no improvement obtained through this transform; see Figure 7.2.

Another option is to initialize with the solution of the PhaseLift problem (7.5) or with the output of our proposed alternative constrained nuclear norm minimization problem (7.9). As shown in Section 7.2.2, the problem (7.5) will always output a zero matrix if no measurement is larger than 1, corresponding to a relative reconstruction error of size 1. We observe this also in numerical experiments; see Table 7.1. Solving Algorithm 4 with a suitable parameter λ can, on the contrary, find better approximations to the ground-truth in this extreme low-count regime. Note that the value 1 is an important benchmark in the relative reconstruction error since this threshold corresponds to beating a simple zero initialization.

$\lvert\{i \in [m] : y_i > 1\}\rvert$	0	1	2	3
PhaseLift (7.5)	1.000	0.941	0.916	0.854
Our method (7.9)	0.972	0.957	0.947	0.910

Table 7.1.: Average relative reconstruction errors in experiments grouped by the amount of observed counts larger than 1. This is the result of 50 random trials.

Beyond the extreme case that $\max_{i\in[m]} y_i = 1$, Table 7.1 shows that PhaseLift with our threshold choice performs better than the constrained nuclear norm minimization (7.9). Based on this observation, we propose to use a hybrid method as initialization technique that solves (7.9) if no measurement observes more than one particle and (7.5) in all other cases. Figure 7.2 shows that this hybrid algorithm improves upon the performance of PhaseLift in very low-dose cases.

To the best of our knowledge, the literature on low-dose Poisson phase retrieval problems focused so far on methods of similar nature - the works in [74, 77, 94] all use a maximum likelihood based approach, as we did in Chapter 6. In contrast to these methods is our proposed Algorithm 4. We compare our algorithm with both the Wirtinger flow for the Poisson log-likelihood loss, with and without TV regularization, suggested in

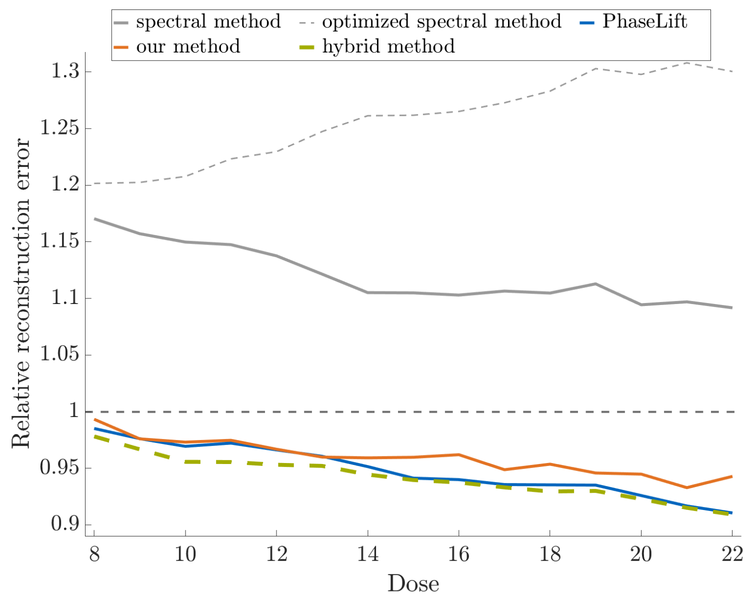


Figure 7.2.: Comparison of different initialization methods for a range of doses. Here, the spectral method is Algorithm 1 and the optimized spectral method uses the measurement transform described in [81]. Our method means solving (7.9) and PhaseLift refers to (7.5). The hybrid method combines the two nuclear norm minimization approaches, depending on the largest observed count in the set of measurements. The linear programs (7.6) and (7.5) are solved with the CVX package in MATLAB. The plot shows averages over 50 random experiments per dose.

[77], and the FIVS method which we had proposed in Chapter 6 for moderately low-count Poisson phase retrieval.

Starting from different initializations, the average reconstruction errors of the tested iterative methods are reported in Table 7.2. Even though it is usually preferable to use an initialization that takes some measurement model information into account, it can be also favorable to use a constant zero object, albeit with a small random offset, as initializer since this is computationally cheaper. For this reason, we tested the iterative algorithms also with a zero initialization, and deduce from Table 7.2 that the iterative methods are not visibly sensitive to the quality of the initialization. In fact, Algorithm 4 exhibits most potential in improving upon its initializer in case of a zero initialization, which would allow us to forgo a computationally more expensive and time intensive initialization with the hybrid method. However, even though an initialization with the hybrid method leaves only little room for improvement by Algorithm 4 in most cases, incorporating some iterations of Algorithm 4 advances potential remaining zero solutions obtained by the hybrid method in case of a suboptimal regularization parameter choice. We conclude that the best average performance is obtained by the combination of the hybrid method and Algorithm 4.

Initialization method	Zero	Spectral method	Hybrid method
Initialization error	1.000	1.132	0.948
Wirtinger flow [77]	1.234	1.233	1.237
WF with TV [77]	1.107	1.031	1.006
FIVS	1.107	1.107	1.108
Algorithm 4	0.943	1.131	**0.941**

Table 7.2.: Relative errors of reconstructions obtained via different algorithms. The table lists errors that are averages from random trials for dose parameters $\mathscr{d} \in \{7, 7.25, \dots, 19.25\}$.

Note that, in comparison to the experiments reported in [77], we do not compensate the extremely low-dose measurement with a large oversampling ratio. This is motivated by the fact that taking more measurements results in more damage, so that real-world experiments require the number of measurements to be sufficiently small to meet an overall dose restriction. In the extreme low-dose and low-oversampling regime that we consider here, our numerical experiments indicate that only the proposed methods incorporating the one-bit quantization assumption have a chance to beat the benchmark of a zero solution.

The compared methods, all of the form of Wirtinger flow methods for maximum likelihood based loss functions, result in reconstructions that are of worse quality than the respective utilized initializations. This suggests that these types of algorithms converge to suboptimal local minima of the respective highly non-convex loss functions.

An interesting observation from the experiments presented in Table 7.2 is that our Algorithm 4 seems not to be compatible with the spectral method as initialization.

In summary, we clearly notice again that the spectral method is not sufficient in this extreme scenario, and state-of-the-art phase retrieval algorithms, even if they work with a metric suitable for Poisson noise, do also not produce reasonable reconstructions. Our proposed methods based on the one-bit quantization approach, in contrast, show noticeably better performance.

Notes and references. *All experiments reported in this section were performed in MATLAB on a MacBook Pro equipped with a 2GHz Intel Core i5 and 16 GB of memory.*

In this part, we run the tests on relatively small test objects due to the higher computational effort in the lifted problem formulation. In this regard, the proposed methods need to be enhanced in future work with computationally more efficient approaches, possibly involving sketching techniques [125] or a scalable method like the iteratively least squares algorithm extensively discussed in [126].

Note that the experiments we presented in [113] used a power method for the spectral initialization with a norm approximation as given in [77]. These calculations turned out to underestimate the norm of the object, which resulted in relative reconstruction errors closer to one. Here, we instead used the spectral method in Algorithm 1 without and with the transform proposed in [81]. Through the better norm approximation (A.1) *the relative reconstruction errors appear worse than the ones reported in [113].*

8. Background removal for ptychography

Ptychographic measurements are commonly distorted by a background signal as described in Section 3.3. Driven by this real-world issue, we develop and analyze a reconstruction algorithm that removes the background from the measurements. The approach is based on the Wigner Distribution Deconvolution method introduced in Section 4.4. Based on our algorithmic approach, we establish a uniqueness of reconstruction guarantee for phase objects. In contrast to the previous chapters, we explicitly and exclusively work here with the STFT phase retrieval problem.

The results presented in this chapter are joint work with Oleh Melnyk and were first published in [87].

8.1. WDD for ptychographic data with background

We strive to find a reconstruction method that recovers $x \in \mathbb{C}^d$ from measurements

$$\tilde{Y}_{k,r} = \left|(F\left[S_{-r}x \circ w\right])_k\right|^2 + b_k, \quad k, r \in [d]_0, \tag{8.1}$$

with unknown background $b = (b_k)_{k\in[d]_0} \in \mathbb{R}^d$. We emphasize again that a background distortion is assumed to be shift-invariant, that means, it is constant with respect to the shift $r \in [d]_0$.

Applying WDD in its vanilla form, that is, Algorithm 3, to measurements with background, the reconstruction quality is expected to be impaired by the distortion. An upper bound on the reconstruction error caused by the measurement contamination is given by Theorem 4.4.6, more specifically detailed in terms of the background signal in the following corollary.

Corollary 8.1.1 (Version of [102, Theorem 1]). *Let $\delta > 2$ and $d \geq 4\delta$. For ptychographic measurements with background distortion* (8.1)*, the output $\tilde{x} \in \mathbb{C}^d$ obtained by Algorithm 3 satisfies*

$$\mathit{dist}(x, \tilde{x}) \leq 24 \frac{\|x\|_\infty}{\min_{t\in[d]_0} |x_t|^2} \left(\frac{d}{\delta}\right)^{\frac{5}{2}} \|b\|_2\, \sigma_{min}^{-1} + d^{\frac{1}{2}} \sqrt{\|b\|_2\, \sigma_{min}^{-1}}\,,$$

where σ_{min} denotes the smallest singular value of the linear operator corresponding to the ptychographic measurements.

Algorithm 3 does not use prior knowledge about the distortion of the measurements. As we assume the background to be the same for all diffraction patterns, there is some sort of redundancy in the data. This property can be used to improve upon the performance of the vanilla WDD algorithm.

Due to the constant nature of the background, the Fourier transform with respect to the shift is zero in all entries except the zeroth frequencies. Hence, we note that only the reconstructions of the zeroth Fourier coefficients $(F[x \circ S_j \overline{x}])_0$, $j \in [d]_0$, are distorted by the background signal.

Proposition 8.1.2. *Ptychographic measurements with background* (8.1) *satisfy*

$$\big(F^{-1}\tilde{Y}F\big)_{j,k} = \big(F[x \circ S_j \overline{x}]\big)_k \cdot \overline{\big(F[\overline{w} \circ S_j w]\big)_k} + d\big(F^{-1}b\big)_j \mathbb{1}_{k=0},$$

for all $j, k \in [d]_0$.

Proof. Since the background is constant with respect to the shift, the measurements (8.1) can be written as

$$\tilde{Y} = Y + \begin{pmatrix} | & & | \\ b & \cdots & b \\ | & & | \end{pmatrix},$$

with $Y = (|(F[S_{-r}x \circ w])_k|^2)_{k,r \in [d]_0}$. Applying the Fourier transforms as in Theorem 4.4.1 to the measurements $\tilde{Y}$, we obtain

$$\begin{aligned} F^{-1}\tilde{Y}F &= F^{-1}YF + F^{-1}\begin{pmatrix} b & \cdots & b \end{pmatrix} F \\ &= F^{-1}YF + F^{-1}\begin{pmatrix} db & 0 & \dots & 0 \end{pmatrix} \\ &= F^{-1}YF + d\begin{pmatrix} F^{-1}b & 0 & \dots & 0 \end{pmatrix}. \end{aligned}$$

This, together with Theorem 4.4.1, yields the statement. □

From this relation we conclude that, if Assumption 1 holds for some $0 < \gamma \leq d$, all Fourier coefficients $(F[x \circ S_j \overline{x}])_k$ for $k > 0$ and all $j \in [\gamma]_0$ are not affected by the distortion. As in Corollary 4.4.4, we derive the following reconstruction formula.

Corollary 8.1.3. *If Assumption 1 holds for some* $0 < \gamma \leq d$*, the Fourier coefficients* $(F[x \circ S_j \overline{x}])_k$ *for all* $-\gamma < j < \gamma$*, and all* $k \in [d]_0 \setminus \{0\}$*, can be recovered via*

$$(F[x \circ S_j \overline{x}])_k = \left(F^{-1}\tilde{Y}F\right)_{j,k} \Big/ \overline{(F[\overline{w} \circ S_j w])_k}.$$

However, the zeroth Fourier coefficients $(F[x \circ S_j \overline{x}])_0$, $j \in [d]_0$, are corrupted by the background, and simply proceeding with steps 2 to 4 in Algorithm 3 using the distorted

Fourier coefficients $(F[x \circ S_j\overline{x}])_0$, $j \in [d]_0$, results in an approximation $\tilde{X} := \tilde{x}\tilde{x}^*$ to the matrix $X = xx^*$ with

$$\begin{aligned}(\tilde{x} \circ S_j\overline{\tilde{x}})_\ell &= \frac{1}{d}\sum_{k=0}^{d-1} e^{\frac{2\pi i k\ell}{d}} \left[\left(\left(F^{-1}\tilde{Y}F\right)_{j,k} - d\big(F^{-1}b\big)_j \mathbb{1}_{k=0}\right) \Big/ \overline{(F[\overline{w} \circ S_j w])_k}\right] \\ &= \frac{1}{d}\sum_{k=0}^{d-1} e^{\frac{2\pi i k\ell}{d}} \left[(F[x \circ S_j\overline{x}])_k - d\big(F^{-1}b\big)_j \mathbb{1}_{k=0} \Big/ \overline{(F[\overline{w} \circ S_j w])_k}\right] \\ &= (x \circ S_j\overline{x})_\ell - \big(F^{-1}b\big)_j \Big/ \overline{(F[\overline{w} \circ S_j w])_0}\end{aligned}$$

for all $j \in [d]_0$. Concerning the algorithm step for magnitude estimation using the information given by the case $j = 0$, this relation reveals a misestimation by

$$|\tilde{x}_\ell| = \sqrt{|x_\ell|^2 - \big(F^{-1}b\big)_0 \Big/ \overline{(F[\overline{w} \circ S_0 w])_0}}. \tag{8.2}$$

We conclude that the background distortion causes an error in the magnitude estimation unless $\big(F^{-1}b\big)_0 = \frac{1}{d}\sum_{k=0}^{d-1} b_k = 0$. Less straightforward are relations for the phase estimation, but since we deal with unknown background signals such formulas are not particularly insightful for improving the reconstruction quality. We do not go into further detail about the distortion effect in the reconstruction.

Instead, we seek for a solution that removes the corruption by the background. Being aware that the background distortion affects the reconstruction of the zeroth Fourier coefficients of the diagonals, and as a result also the object reconstruction, we decide to discard these distorted Fourier coefficients $(F[x \circ S_j\overline{x}])_0$, $j \in [d]_0$.

Nevertheless, we aim to proceed with the remaining steps of Algorithm 3. Consequently, our strategy is to reconstruct the lost components $(F[x \circ S_j\overline{x}])_0$, $j \in [d]_0$, before continuing with the reconstruction algorithm as usual.

8.2. A background removing reconstruction algorithm

For the reconstruction of the lost frequencies, we make use of a relation between the diagonals of the rank-one matrix xx^*. Namely, we utilize that the diagonals are related by

$$(x \circ S_j\overline{x}) \circ S_\ell\overline{(x \circ S_j\overline{x})} = (x \circ S_\ell\overline{x}) \circ S_j\overline{(x \circ S_\ell\overline{x})} \tag{8.3}$$

for all $j, \ell \in [d]_0$. To transit from (8.3) to a property for frequencies, the Fourier transform is applied,

$$F[(x \circ S_j\overline{x}) \circ S_\ell\overline{(x \circ S_j\overline{x})}] = F[(x \circ S_\ell\overline{x}) \circ S_j\overline{(x \circ S_\ell\overline{x})}]. \tag{8.4}$$

With the convolution theorem, Theorem C.4 (i), these terms can be rewritten by

$$F[(x \circ S_p\overline{x}) \circ S_q\overline{(x \circ S_p\overline{x})}] = \tfrac{1}{d}F[x \circ S_p\overline{x}] * F[S_q\overline{(x \circ S_p\overline{x})}],$$

for both $(p,q) \in \{(j,\ell),(\ell,j)\}$.

From here on, we abbreviate the Fourier coefficients to

$$f_k^j := \left(F\left[x \circ S_j\overline{x}\right]\right)_k$$

for all $j,k \in [d]_0$. Using the properties of the shift, modulation, and reflection operators in Lemma C.1, we obtain

$$\Big(F[x \circ S_j\overline{x}] * F[S_\ell\overline{(x \circ S_j\overline{x})}]\Big)_n = \Big(f^j * M_{-\ell}R\overline{f^j}\Big)_n = \sum_{k=0}^{d-1} e^{-\frac{2\pi i\ell(n-k)}{d}} f_k^j\, \overline{f^j}_{k-n}$$

for all $n \in [d]_0$. Consequently, the relation (8.4) is equivalent to

$$\sum_{k=0}^{d-1} e^{-\frac{2\pi i\ell(n-k)}{d}} f_k^j\, \overline{f^j}_{k-n} = \sum_{k=0}^{d-1} e^{-\frac{2\pi ij(n-k)}{d}} f_k^\ell\, \overline{f^\ell}_{k-n} \tag{8.5}$$

for all $n \in [d]_0$.

If Assumption 1 is satisfied for $\gamma \leq d$ and $\ell, j \in [\gamma]_0$, the summands in (8.5) are known via Corollary 8.1.3 for all $k \in [d]_0 \setminus \{0,n\}$. Using these known summands, we aim to recover the unknown summands corresponding to $k \in \{0,n\}$. Thereto, we sort the unknown parts of (8.5) into the left-hand side and the known parts into the right-hand side, and obtain the system of linear equations

$$\begin{aligned} &e^{-\frac{2\pi ijn}{d}} f_0^\ell\, \overline{f^\ell}_{-n} + f_n^\ell\, \overline{f^\ell}_0 - e^{-\frac{2\pi i\ell n}{d}} f_0^j\, \overline{f^j}_{-n} - f_n^j\, \overline{f^j}_0 \\ &\quad = \sum_{\substack{k=1\\k\neq n}}^{d-1} e^{-\frac{2\pi i\ell(n-k)}{d}} f_k^j\, \overline{f^j}_{k-n} - e^{-\frac{2\pi ij(n-k)}{d}} f_k^\ell\, \overline{f^\ell}_{k-n} \end{aligned} \tag{8.6}$$

for all $n \in [d]_0 \setminus \{0\}$. Note that for $n = 0$ this only leads to the squared magnitudes $|f_0^j|^2$ and hence the corresponding equation is not considered.

For a fixed $j \in [\gamma]_0$, the system of equations (8.6) with $n \in [d]_0 \setminus \{0\}$ for any $\ell \in [\gamma]_0 \setminus \{j\}$ poses a linear inverse problem with four real unknowns. By solving these systems for all $j \in [\gamma]_0$ we can recover the zeroth frequencies $f_0^j = (F[x \circ S_j\overline{x}])_0$ for all $j \in [\gamma]_0$.

This computational problem can be alleviated as the number of unknowns reduces when working with $\ell = 0$. Note that $(x \circ S_0\overline{x})_t = |x_t|^2 \in \mathbb{R}$ for all $t \in [d]_0$. Hence, by

Corollary C.2, it is $f_n^0 = \overline{f^0}_{-n}$ for all $n \in [d]_0$. With this, we can rewrite (8.6) to

$$\left(e^{-\frac{2\pi ijn}{d}} + 1\right) f_n^0 \, f_0^0 - f_0^j \, \overline{f^j}_{-n} - f_n^j \, \overline{f^j}_0 = c_{0,j,n}, \quad n \in [d]_0 \setminus \{0\}, \tag{8.7}$$

where, for convenience of notation, we abbreviate the fully known right-hand side by

$$c_{\ell,j,n} := \sum_{\substack{k=1 \\ k \neq n}}^{d-1} e^{-\frac{2\pi i\ell(n-k)}{d}} f_k^j \, \overline{f^j}_{k-n} - e^{-\frac{2\pi ij(n-k)}{d}} f_k^\ell \, \overline{f^\ell}_{k-n}$$

for all $j \neq \ell \in [\gamma]_0$ and all $n \in [d]_0 \setminus \{0\}$.

If $z_{j,n} := \left(e^{-\frac{2\pi ijn}{d}} + 1\right) f_n^0 = 0$, this reduces to

$$-f_0^j \, \overline{f^j}_{-n} - f_n^j \, \overline{f^j}_0 = c_{0,j,n},$$

or, equivalently, the real-valued system

$$\begin{aligned} -\operatorname{Re} f_0^j \, [\operatorname{Re} f_{-n}^j + \operatorname{Re} f_n^j] - \operatorname{Im} f_0^j \, [\operatorname{Im} f_{-n}^j + \operatorname{Im} f_n^j] &= \operatorname{Re} c_{0,j,n}, \\ \operatorname{Re} f_0^j \, [\operatorname{Im} f_{-n}^j - \operatorname{Im} f_n^j] + \operatorname{Im} f_0^j \, [\operatorname{Re} f_n^j - \operatorname{Re} f_{-n}^j] &= \operatorname{Im} c_{0,j,n}. \end{aligned} \tag{8.8}$$

If $z_{j,n} \neq 0$, we can multiply (8.7) by $\overline{z}_{j,n}$ to get

$$f_0^0 |z_{j,n}|^2 - f_0^j \, \overline{f^j}_{-n} \overline{z}_{j,n} - f_n^j \, \overline{f^j}_0 \overline{z}_{j,n} = c_{0,j,n} \overline{z}_{j,n}. \tag{8.9}$$

It is

$$f_0^0 = \sum_{t=0}^{d-1} (x \circ S_0 \overline{x})_t \;\; = \|x\|_2^2 \geq 0, \tag{8.10}$$

hence $f_0^0 \in \mathbb{R}$ and also $f_0^0 |z_{j,n}|^2 \in \mathbb{R}$. Thus, the imaginary part of (8.9) equals

$$\operatorname{Re} f_0^j \, [\operatorname{Im}(f_{-n}^j z_{j,n}) - \operatorname{Im}(f_n^j \overline{z}_{j,n})] + \operatorname{Im} f_0^j \, [\operatorname{Re}(f_n^j \overline{z}_{j,n}) - \operatorname{Re}(f_{-n}^j z_{j,n})] = \operatorname{Im}(c_{0,j,n} \overline{z}_{j,n}). \tag{8.11}$$

By this, we reduced the problem of recovering four real unknowns from (8.6) to a linear system with just two real unknowns, namely $\operatorname{Re} f_0^j$ and $\operatorname{Im} f_0^j$, given by either (8.8) or (8.11) for $n \in [d]_0 \setminus \{0\}$.

It remains then to reconstruct f_0^0. For $n = 0$, equation (8.6) reads as

$$|f_0^0|^2 - |f_0^j|^2 = c_{0,j,0},$$

which, with $f_0^0 \geq 0$ as shown in (8.10), yields

$$f_0^0 = \sqrt{|f_0^0|^2} = \sqrt{c_{0,j,0} + |f_0^j|^2}$$

for any $j \in [\gamma]_0 \setminus \{0\}$. For a better stability, we average over all $j \in [\gamma]_0 \setminus \{0\}$, and find

$$f_0^0 = \left(\frac{1}{\gamma - 1} \sum_{j=1}^{\gamma-1} \left(c_{0,j,0} + |f_0^j|^2 \right) \right)^{1/2} . \tag{8.12}$$

With the derived strategy we can recover the lost Fourier frequencies. Subsequently, the reconstruction of x can be resumed as in Algorithm 3. The whole procedure is summarized in Algorithm 5.

Algorithm 5 Recovery algorithm for ptychography with background distortion

---

Input: Ptychographic measurements $\tilde{Y} \in \mathbb{R}^{d \times d}$ as in (8.1) with $w \in \mathbb{C}^d$ satisfying $\operatorname{supp}(w) = [\delta]_0$ for $\delta < d$ and Assumption 1 for $0 < \gamma \leq d$.

Step 1: Use Corollary 8.1.3 to compute f_k^j for all $j \in [\gamma]_0,\ k \in [d]_0 \setminus \{0\}$.

Step 2: Reconstruct f_0^j for all $j \in [\gamma]_0 \setminus \{0\}$ via the linear system consisting of a combination of equations (8.8) and (8.11) for $n \in [d]_0 \setminus \{0\}$.

Step 3: Reconstruct f_0^0 via (8.12).

Step 4: Via inverse Fourier transform applied to the Fourier coefficients obtained in Step 1 - 3, compute $(x \circ S_j \overline{x})_t$ for all $j \in [\gamma]\,,\ t \in [d]_0$, to build X as in (4.13).

Step 5: Compute the magnitudes $|\tilde{x}_t| = \sqrt{X_{t,t}},\ t \in [d]_0$.

Step 6: Compute the top eigenvector $\tilde{z}$ of $\operatorname{sgn}(X)$ and set $\operatorname{sgn}(\tilde{x}) = \operatorname{sgn}(\tilde{z})$.

Output: $\tilde{x} = |\tilde{x}| \cdot \operatorname{sgn}(\tilde{x}) \in \mathbb{C}^d$

Notes and references. *The algorithm presented in this section was inspired by the subspace completion technique proposed in [46]. It was originally designed for recovering lost Fourier coefficients that could not be obtained because Assumption 1 failed to hold in some instances, or to restore coefficients that were truncated for a regularization purpose.*

Here, we were specifically interested in reconstructing the zeroth coefficient, motivated by the background distortion problem. Clearly, any other coefficient per diagonal could be recovered with the same strategy. Beyond the here considered problem with one lost Fourier coefficient per diagonal, it would be interesting to investigate whether also multiple coefficients per diagonal could be recovered using the redundancy in the data. A theoretical analysis for the recovery strategy derived in this section is yet to be explored in future research.

8.3. A reconstruction strategy for phase objects

In this section, we focus on so-called phase objects. As briefly mentioned in Section 2.2, this means objects that are sufficiently thin so that no probe energy is absorbed and only the scattering effect of the object is measured. Since non-absorbing objects are modeled with a constant magnitude equal to one and their scattering information is reflected by the phase of the complex-valued object transmission function, the class of phase objects is defined as follows.

Definition 8.3.1. *A vector $v \in \mathbb{C}^d$ is called a phase object if*

$$v \in \mathbb{T}^d := \left\{ z \in \mathbb{C}^d : |z_t| = 1,\ t \in [d]_0 \right\}.$$

Restricting ourselves to this type of object allows us to decouple the magnitude and the phase recovery for the lost Fourier coefficients $f_0^j,\ j \in [d]_0$. For phase objects $x \in \mathbb{T}^d$, we find a closed-form expression for the magnitude of the lost Fourier coefficients f_0^j.

Proposition 8.3.1. *Let $x \in \mathbb{T}^d$. For every $j \in [d]_0$, the vector $x \circ S_j \overline{x}$ is an element of $\mathbb{T}^d$, and its Fourier transform $f^j = F\left[x \circ S_j \overline{x}\right]$ satisfies $\|f^j\|_2^2 = d^2$. Consequently, the Fourier coefficients $|f_0^j|,\ j \in [d]_0$, can be recovered by*

$$|f_0^j|^2 = d^2 - \sum_{k=1}^{d-1} |f_k^j|^2.$$

Proof. For $x \in \mathbb{T}^d$, we have

$$\left|(x \circ S_j \overline{x})_t\right| = |x_t||x_{t-j}| = 1$$

for all $j, t \in [d]_0$. Using this equality and Plancherel's identity, see (C.5), we obtain

$$\sum_{k=0}^{d-1} |f_k^j|^2 = \left\|f^j\right\|_2^2 = \|F\left[x \circ S_j \overline{x}\right]\|_2^2 = d\,\|x \circ S_j \overline{x}\|_2^2 = d \cdot \sum_{t=0}^{d-1} \left|x_t \overline{x}_{t-j}\right|^2 = d \cdot \sum_{t=0}^{d-1} 1 = d^2.$$

□

It remains to recover the phases of $f_0^j,\ j \in [d]_0$. If the formula in Proposition 8.3.1 yields $|f_0^j| = 0$, we conclude that $f_0^j = 0$. For all other cases, we have to reconstruct the argument of the lost coefficient f_0^j, in the sequel abbreviated by

$$\varphi_j := \arg(f_0^j). \tag{8.13}$$

The following result shows that these missing arguments $\varphi_j,\ j \in [d]_0$, can be found by solving a linear system.

Theorem 8.3.2. *Let $x \in \mathbb{T}^d$. For all $j \in [d]_0$ with $|f_0^j| > 0$, the argument $\varphi_j \in (0, 2\pi]$ solves the system*

$$A^j \begin{pmatrix} \cos(\varphi_j) \\ \sin(\varphi_j) \end{pmatrix} = \frac{d^2}{2|f_0^j|} \begin{pmatrix} 1 - \frac{1}{d^2}\left(|a_0^j|^2 + |f_0^j|^2\right) \\ \vdots \\ 1 - \frac{1}{d^2}\left(|a_{d-1}^j|^2 + |f_0^j|^2\right) \end{pmatrix}, \tag{8.14}$$

where

$$A^j := \begin{pmatrix} \mathrm{Re}(a_0^j) & \mathrm{Im}(a_0^j) \\ \vdots & \vdots \\ \mathrm{Re}(a_{d-1}^j) & \mathrm{Im}(a_{d-1}^j) \end{pmatrix} \quad \text{and} \quad a^j := dF^{-1} \begin{pmatrix} 0 \\ f_1^j \\ \vdots \\ f_{d-1}^j \end{pmatrix}. \tag{8.15}$$

Proof. Proposition 8.3.1 states that, for $x \in \mathbb{T}^d$, the diagonals of xx^* satisfy $x \circ S_j\overline{x} \in \mathbb{T}^d$ for all $j \in [d]_0$. Hence, we also have

$$\left|\left(F^{-1}\left(F\left[x \circ S_j\overline{x}\right]\right)\right)_t\right| = \left|\left(x \circ S_j\overline{x}\right)_t\right| = 1 \tag{8.16}$$

for all $t, j \in [d]_0$. We expand the inverse Fourier transform as

$$\left(F^{-1}\left(F\left[x \circ S_j\overline{x}\right]\right)\right)_t = \frac{1}{d}\sum_{k=0}^{d-1} e^{\frac{2\pi ikt}{d}} \left(F\left[x \circ S_j\overline{x}\right]\right)_k = \frac{1}{d}\left|f_0^j\right| e^{i\varphi_j} + \frac{1}{d}a_t^j, \tag{8.17}$$

using the definitions (8.13) and (8.15).

To recover φ_j, we combine (8.16) and (8.17), and study the system of quadratic equations

$$d^2 = \left||f_0^j| \cdot e^{i\varphi_j} + a_t^j\right|^2, \quad t \in [d]_0.$$

Expanding the squares results in

$$\begin{aligned} \left||f_0^j| \cdot e^{i\varphi_j} + a_t^j\right|^2 &= |f_0^j|^2 + 2|f_0^j| \,\mathrm{Re}\left(a_t^j e^{-i\varphi_j}\right) + |a_t^j|^2 \\ &= |f_0^j|^2 + 2|f_0^j| \left(\mathrm{Re}(a_t^j)\cos(\varphi_j) + \mathrm{Im}(a_t^j)\sin(\varphi_j)\right) + |a_t^j|^2. \end{aligned}$$

Consequently, φ_j has to satisfy the condition

$$\mathrm{Re}(a_t^j)\cos(\varphi_j) + \mathrm{Im}(a_t^j)\sin(\varphi_j) = \frac{1}{2|f_0^j|}\left(d^2 - |a_t^j|^2 - |f_0^j|^2\right)$$

for all $t \in [d]_0$. That means, the missing phase $e^{i\varphi_j} = \cos(\varphi_j) + i\sin(\varphi_j)$ is a solution of the linear system (8.14). □

Note that it is not necessary to solve the full system (8.14) due to the redundancy in the equations because of $\operatorname{rank}(A^j) \leq 2$. The $(t+1)$th equation of (8.14) is equivalent to

$$\operatorname{Re}\left(a_t^j e^{-i\varphi_j}\right) = \frac{d^2 - |a_t^j|^2 - |f_0^j|^2}{2|f_0^j|},$$

where

$$\operatorname{Re}\left(a_t^j e^{-i\varphi_j}\right) = |a_t^j| \cos(\arg(a_t^j) - \varphi_j).$$

Consequently, the missing argument φ_j satisfies

$$\varphi_j \in \bigcap_{t=0}^{d-1} \left\{ \arg(a_t^j) \pm \arccos\left(\frac{d^2 - |a_t^j|^2 - |f_0^j|^2}{2|a_t^j||f_0^j|} \right) \right\}. \tag{8.18}$$

If $\operatorname{rank}(A^j) = 2$, the intersection of sets in (8.18) is a singleton, and we can determine this solution by computing the two possible values for $t_1 \neq t_2 \in [d]_0$ satisfying

$$\frac{a_{t_1}^j}{a_{t_2}^j} \neq \frac{1 - \frac{1}{d^2}\left(|a_{t_1}^j|^2 + |f_0^j|^2\right)}{1 - \frac{1}{d^2}\left(|a_{t_2}^j|^2 + |f_0^j|^2\right)}.$$

The recovery strategy developed above, combined with the additional steps of Algorithm 3, is summarized in Algorithm 6.

8.4. Recovery guarantee

The systems (8.14) provide the recovery procedure separately for each φ_j for $j \in [d]_0$. When considered individually, φ_j is not necessarily uniquely determined by the corresponding linear system (8.14). However, all φ_j, $j \in [d]_0$, belong to the same object x, and investigating multiple systems (8.14) simultaneously allows us to establish whether x is uniquely determined or not.

We find two types of objects which cannot be recovered uniquely if the ptychographic measurements are distorted by a background signal.

Theorem 8.4.1. *(i) For all $m \in [d]_0$, there exist background signals $b^m \in \mathbb{R}^d$ such that objects*

$$x \sim x^m := \left(e^{\frac{2\pi i t m}{d}}\right)_{t \in [d]_0} \tag{8.19}$$

exhibit the same measurements (8.1).

(ii) Suppose that d is even. Let x^q for $q = 1, 2$, be defined by

$$x_t^q := e^{\frac{2\pi i t m}{d}} \cdot \begin{cases} 1, & t \text{ even}, \\ -(-1)^q e^{-(-1)^q \frac{1}{2} i\rho}, & t \text{ odd}, \end{cases} \qquad t \in [d]_0, \tag{8.20}$$

for some $m \in [d]_0$, $\rho \in (-\pi, \pi)$. If $\tilde{x}^1 \sim x^1$ and $\tilde{x}^2 \sim x^2$, there exist backgrounds b^q, $q = 1, 2$, such that $(\tilde{x}^1, b^1)$ and $(\tilde{x}^2, b^2)$ result in the same measurements (8.1).

Algorithm 6 Background removing reconstruction algorithm for phase objects

---

Input: Ptychographic measurements $\tilde{Y} \in \mathbb{R}^{d\times d}$ as in (8.1) with $w \in \mathbb{C}^d$ satisfying $\mathrm{supp}(w) = [\delta]_0$ for $\delta < d$ and Assumption 1 for $0 < \gamma \leq d$.

Step 1: Compute f_k^1 for all $k \in [d]_0 \setminus \{0\}$ via Proposition 8.1.2.
Recover $|f_0^1|$ via Proposition 8.3.1 and set up A^1.

if $\mathrm{rank}(A^1) = 0$ **then**
- Stop. Ground-truth x is of ambiguity type (8.19).

else if $\mathrm{rank}(A^1) = 1$ **then**
- Compute f_k^2 for all $k \in [d]_0 \setminus \{0\}$ via Proposition 8.1.2.
  Recover $|f_0^2|$ via Proposition 8.3.1 and set up A^2.
- **if** *d is odd* **then**
  - Solve (8.14) to recover f^2. Go to Step 2 with $\gamma = 3$.
- **else**
  - Compute the two solutions $f^{1,q}$, $q = 1, 2$, using (8.18).
  - **if** $\mathrm{rank}(A^2) = 0$ **then**
    - Stop. Ground-truth x is of ambiguity type (8.20) with ρ, m determined by $F^{-1} f^{1,q}$, $q = 1, 2$, via (8.29) and (8.30).
  - **else**
    - Solve (8.14) to recover f^2. Select $f^{1,q}$, $q = 1, 2$, which meets equality in (8.24). Go to Step 2 with $\gamma = 3$.

else
- Solve (8.14). Go to Step 2 with $\gamma = 2$.

Step 2: Build X as in (4.13) with $\gamma = 2$ or $\gamma = 3$ by applying the inverse Fourier transform to the Fourier coefficients obtained in Step 1.

Step 3: Compute the top eigenvector $\tilde{z}$ of $\mathrm{sgn}(X)$ and set $\mathrm{sgn}(\tilde{x}) = \mathrm{sgn}(\tilde{z})$.

Output: $\tilde{x} = \mathrm{sgn}(\tilde{x}) \in \mathbb{C}^d$,
with $\tilde{x} \sim x$ if $x \in \mathbb{T}^d$ is not of ambiguity type (8.19) or (8.20) and if $\gamma \geq 3$

Proof. (i) : For $m \in [d]_0$, let x^m be the object defined via (8.19), and let $\tilde{Y}^m$ be the matrix storing the respective measurements

$$\tilde{Y}^m_{k,r} := \left|(F\left[S_{-r}x^m \circ w\right])_k\right|^2 + b^m_k, \quad k,r \in [d]_0,$$

with some background $b^m \in \mathbb{R}^d$.

In the following, we will construct vectors $b^m \in \mathbb{R}^d$ which satisfy

$$(F^{-1}\tilde{Y}^m F)_{j,k} = (F^{-1}\tilde{Y}^0 F)_{j,k}$$

for all $m, j, k \in [d]_0$, and, consequently, result in $\tilde{Y}^m = \tilde{Y}^0$ for all $m \in [d]_0$.

Note that each diagonal of $x^m\,(x^m)^*$ corresponding to an object x^m is a constant vector, more precisely, its entries are

$$(x^m \circ S_j\overline{x^m})_t = e^{\frac{2\pi i j m}{d}} \tag{8.21}$$

for all $t \in [d]_0$. Hence, for all $m, j \in [d]_0$ and all $k \in [d]_0 \setminus \{0\}$,

$$f^{j,m}_k := \left(F\left[x^m \circ S_j\overline{x^m}\right]\right)_k = 0,$$

and, with Proposition 8.1.2, we conclude that

$$(F^{-1}\tilde{Y}^m F)_{j,k} = 0 = (F^{-1}\tilde{Y}^0 F)_{j,k}$$

for all $m, j \in [d]_0$ and all $k \in [d]_0 \setminus \{0\}$.

Further, according to Proposition 8.1.2,

$$(F^{-1}\tilde{Y}^m F)_{j,0} = (F^{-1}\tilde{Y}^0 F)_{j,0}$$

holds for all $j \in [d]_0$ if and only if

$$d(F^{-1}(b^m - b^0))_j = \left(f^{j,0}_0 - f^{j,m}_0\right)(\overline{F[\overline{w} \circ S_j w]})_0,$$

which is equivalent to

$$(F^{-1}(b^m - b^0))_j = \left(1 - e^{\frac{2\pi i j m}{d}}\right)(\overline{F[\overline{w} \circ S_j w]})_0, \tag{8.22}$$

since (8.21) provides $f^{j,m}_0 = d e^{\frac{2\pi i j m}{d}}$.

Via (8.22), we can define background signals b^m for all $m \in [d]_0$ that induce the same ptychographic measurements for different objects x^m. Take, for example, $(F^{-1}b^0)_j = 0$ for all $j \in [d]_0 \setminus \{0\}$. Then, it remains to choose $(F^{-1}b^0)_0 \in \mathbb{R}$ suitably. If $F^{-1}b^0$ is chosen like this, it is conjugate symmetric, and Corollary C.2 states that $b^0 = F(F^{-1}b^0) \in \mathbb{R}^d$. With (8.22) we obtain that also $F^{-1}b^m$ is conjugate symmetric: It is $(F^{-1}b^m)_0 = (F^{-1}b^0)_0 \in \mathbb{R}$

due to $e^{\frac{2\pi i j m}{d}} = 1$ for $j = 0$, and

$$\begin{aligned}\overline{(F^{-1}b^m)_{-j}} &= \overline{(F^{-1}(b^m - b^0))_{-j}} = \overline{\left(1 - e^{-\frac{2\pi i j m}{d}}\right)(F[\overline{w} \circ S_{-j}w])_0} \\ &= \left(1 - e^{\frac{2\pi i j m}{d}}\right)\overline{(F[\overline{w} \circ S_j w])_0} = (F^{-1}(b^m - b^0))_j = (F^{-1}b^m)_j\end{aligned}$$

for all $j \in [d]_0 \setminus \{0\}$, where we use that $(F^{-1}b^0)_j = 0$ and

$$(F[\overline{w} \circ S_{-j}w])_0 = \sum_{t=0}^{d-1} \overline{w}_t w_{t+j} = \sum_{t=0}^{d-1} \overline{w}_{t-j} w_t = \overline{\sum_{t=0}^{d-1} \overline{w}_t w_{t-j}} = \overline{(F[\overline{w} \circ S_j w])_0}$$

for all $j \in [d]_0 \setminus \{0\}$. Hence, with Corollary C.2, also $b^m \in \mathbb{R}^d$ for all $m \in [d]_0$.

To conform with the physical model, we should note that ptychographic measurements represent particle counts and hence $\tilde{Y}_{k,r} \geq 0$ needs to be satisfied for all $m, k, r \in [d]_0$. That means, the last requirement is to choose $(F^{-1}b^0)_0 \in \mathbb{R}$ sufficiently large so that all backgrounds b^m ensure the positivity of the measurements. Imposing

$$\begin{aligned}0 &\leq |\,(F[S_{-r}x^m \circ w])_k\,|^2 + b^m_k \\ &= |\,(F[S_{-r}x^m \circ w])_k\,|^2 + (F(F^{-1}b^m))_k \\ &= |\,(F[S_{-r}x^m \circ w])_k\,|^2 + (F^{-1}b^m)_0 + \sum_{t=1}^{d-1} e^{-\frac{2\pi i t k}{d}} (F^{-1}b^m)_t\end{aligned}$$

for all $m, k, r \in [d]_0$, requires

$$(F^{-1}b^m)_0 \geq -|\,(F[S_{-r}x^m \circ w])_k\,|^2 - \sum_{t=1}^{d-1} e^{-\frac{2\pi i t k}{d}} (F^{-1}b^m)_t \tag{8.23}$$

to hold for all $m, k, r \in [d]_0$. Here, we would like to remind the reader that $(F^{-1}b^m)_t$ for all $m \in [d]_0$ and $t \in [d]_0 \setminus \{0\}$ is fixed via (8.22) and $(F^{-1}b^0)_t = 0$ for all $t \in [d]_0 \setminus \{0\}$. Furthermore, as noted before, it is $(F^{-1}b^0)_0 = (F^{-1}b^m)_0$ for all $m \in [d]_0$. Consequently, (8.23) is attained with

$$(F^{-1}b^0)_0 = \max_{m,k,r \in [d]_0} \left(-|\,(F[S_{-r}x^m \circ w])_k\,|^2 - \sum_{t=1}^{d-1} e^{-\frac{2\pi i t k}{d}} (F^{-1}b^m)_t \right).$$

We conclude that such choices of background signals b^m lead to the same ptychographic measurements for all objects x^m, $m \in [d]_0$. This means that objects x^m and all equivalent objects with respect to the global phase ambiguity are not uniquely determined by ptychographic measurements with background distortion.

(ii) : Fix $m \in \mathbb{Z}$, $\rho \in (-\pi, \pi)$, and consider the respective objects x^q, $q = 1, 2$, as

defined in (8.20). We aim to show that there exist $b^1, b^2 \in \mathbb{R}^d$ satisfying

$$\tilde{Y}^1_{k,r} := \left|\left(F\left[S_{-r}x^1 \circ w\right]\right)_k\right|^2 + b^1_k = \left|\left(F\left[S_{-r}x^2 \circ w\right]\right)_k\right|^2 + b^2_k =: \tilde{Y}^2_{k,r}$$

for all $k, r \in [d]_0$.

The entries of the jth off-diagonal corresponding to the objects x^q, $q = 1, 2$, are

$$(x^q \circ S_j\overline{x^q})_t = \begin{cases} 1, & j \text{ even}, \\ -(-1)^q e^{\frac{2\pi i j m}{d} + (-1)^q(-1)^t \frac{1}{2} i\rho}, & j \text{ odd}, \end{cases}$$

for all $t \in [d]_0$. Hence, for j even, the Fourier coefficients equal

$$f^{j,q}_k = (F\left[x^q \circ S_j\overline{x^q}\right])_k = \begin{cases} d, & k = 0, \\ 0, & k \in [d]_0 \setminus \{0\}, \end{cases}$$

and for j odd, these are given by

$$\begin{aligned} f^{j,q}_k &= -(-1)^q e^{\frac{2\pi i j m}{d}} \left(\sum_{\substack{t=0 \\ t \text{ even}}}^{d-1} e^{-\frac{2\pi i k t}{d}} e^{(-1)^q \frac{1}{2} i\rho} + \sum_{\substack{t=0 \\ t \text{ odd}}}^{d-1} e^{-\frac{2\pi i k t}{d}} e^{-(-1)^q \frac{1}{2} i\rho} \right) \\ &= -(-1)^q e^{\frac{2\pi i j m}{d}} \left(\sum_{t=0}^{\frac{d}{2}-1} e^{-\frac{2\pi i k \frac{t}{2}}{\frac{d}{2}}} e^{(-1)^q \frac{1}{2} i\rho} + \sum_{t=0}^{\frac{d}{2}-1} e^{-\frac{2\pi i k \frac{t-1}{2}}{\frac{d}{2}}} e^{-(-1)^q \frac{1}{2} i\rho} \right) \\ &= \begin{cases} -(-1)^q e^{\frac{2\pi i j m}{d}} \cdot \frac{d}{2}\cos(\frac{\rho}{2}), & k = 0, \\ -e^{\frac{2\pi i j m}{d}} \cdot \frac{d}{2} i \sin(\frac{\rho}{2}), & k = \frac{d}{2}, \\ 0, & k \in [d]_0 \setminus \{0, \frac{d}{2}\}. \end{cases} \end{aligned}$$

In both cases, all Fourier coefficients do not depend on q. Hence, for all $j \in [d]_0$ and all $k \in [d]_0 \setminus \{0\}$, we have $f^{j,1}_k = f^{j,2}_k$, and Proposition 8.1.2 guarantees $(F^{-1}\tilde{Y}^1 F)_{j,k} = (F^{-1}\tilde{Y}^2 F)_{j,k}$. For $k = 0$, we proceed as before for objects of type (i). It is $(F^{-1}\tilde{Y}^1 F)_{j,0} = (F^{-1}\tilde{Y}^2 F)_{j,0}$ for all $j \in [d]_0$ if and only if

$$d(F^{-1}(b^2 - b^1))_j = \left(f^{j,1}_0 - f^{j,2}_0\right) (\overline{F[\overline{w} \circ S_j w]})_0.$$

We can continue as in case (i) and find that there exist background signals b^1 and b^2 with which objects x^1 and x^2 produce the same measurements (8.1). In summary, unique recovery of objects x^1 or x^2 and all representatives of the corresponding equivalence classes with respect to the global phase is also not possible. □

Note that these negative examples are neither caused by discarding the zeroth Fourier coefficients nor a weakness of our reconstruction algorithm. The proof of Theorem 8.4.1 shows that this issue appears due to the background distortion and the respective two types of objects can never be uniquely determined as long as the involved background signal is unknown.

The only non-trivial Fourier coefficients of the diagonals corresponding to the objects described in Theorem 8.4.1 (i) are the lost zeroth coefficients. Therefore, the information given by the exactly recovered Fourier coefficients f_k^j for $k \in [d]_0 \setminus \{0\}$ and all $j \in [d]_0$ is not sufficient for reconstructing the phases of the lost coefficients. However, this only happens for objects with constant diagonals, a property unique to modulations of a constantly one vector, that is precisely the objects defined by (8.19).

The second negative example resembles the conjugate reflection ambiguity well-known for Fourier phase retrieval [10] as the objects (8.20) satisfy

$$\overline{x_{-t}^1} = (-1)^t x_t^2, \quad t \in [d]_0.$$

However, in contrast to Fourier phase retrieval, in STFT phase retrieval and hence ptychography this kind of ambiguity only concerns objects of this specific type and not all $x \in \mathbb{T}^d$.

Apart from the two families of objects described in Theorem 8.4.1, all phase objects $x \in \mathbb{T}^d$ are uniquely determined by ptychographic measurements with background.

Theorem 8.4.2. *Let Assumption 1 hold with $\gamma \geq 3$. Assume that $x \in \mathbb{T}^d$ is neither an object of type (i) nor of type (ii) as described in Theorem 8.4.1. Then, x can be uniquely recovered from ptychographic measurements with background* (8.1)*.*

Proof. As mentioned above, the uniqueness of reconstruction depends on the underlying rank of the linear systems (8.14). All possible scenarios are summarized in Table 8.1.

We will show that it is sufficient to use the information given by the two cases $j = 1$ and $j = 2$ to fully characterize the uniqueness of reconstruction for ptychographic measurements with background. The proof is split into the separate analyses of the disjoint cases presented in the exhaustive list in Table 8.1.

The case $j = 0$ is not considered since

$$(F\,[x \circ S_0 \overline{x}])_0 = \sum_{t=0}^{d-1} e^{-\frac{2\pi i t \cdot 0}{d}} \cdot |x_t|^2 = d$$

is known for every phase object, and does not provide any information on the phases of the object.

We start with investigating $j = 1$.

If $\operatorname{rank}(A^1) = 2$, unique recovery is guaranteed by Lemma C.9 since 1 is coprime with any $d \in \mathbb{N}$.

rank(A^1)	rank(A^2)	d	Result	Corresponding Statement
2			unique recovery	Lemma C.9 with $j = 1$
0			negative example (i)	Lemma C.7 with $j = 1$
1	1		not possible	(★)
1	0	odd	not possible	(★★)
1	0	even	negative example (ii)	(★★)
1	2	odd	unique recovery	Lemma C.9 with $j = 2$
1	2	even	unique recovery	(★★★)

Table 8.1.: Whether (8.14) in Theorem 8.3.2 is uniquely solvable or not depends on the rank of the matrix A^j. The listed cases have to be analyzed in the proof of Theorem 8.4.2.

Let $\mathrm{rank}(A^1) = 0$. Since 1 is coprime with any $d \in \mathbb{N}$, Lemma C.7 can be used to show that the only object that satisfies this condition is the negative example (i) in Theorem 8.4.1. We showed that for this type of object unique recovery is never possible.

If $\mathrm{rank}(A^1) = 1$, considering only the information corresponding to $j = 1$ is not sufficient. This is proven in Lemma C.8. Thus, we need to involve knowledge on further diagonals with $j > 1$.

We continue with the case $j = 2$. Again, the three possible cases $\mathrm{rank}(A^2) = 0$, $\mathrm{rank}(A^2) = 1$, and $\mathrm{rank}(A^2) = 2$ have to be considered.

In the following, we use that the first and the second off-diagonal are related via

$$\begin{aligned}(F^{-1}f^2)_t &= (x \circ S_2\overline{x})_t = x_t \cdot 1 \cdot \overline{x_{t-2}} = x_t \, |x_{t-1}|^2 \, \overline{x_{t-2}} \\ &= (x \circ S_1\overline{x})_t \, (x \circ S_1\overline{x})_{t-1} = (F^{-1}f^1)_t \, (F^{-1}f^1)_{t-1}.\end{aligned} \tag{8.24}$$

We start with $\mathrm{rank}(A^2) = 1$ and prove the following statement:

The case $\mathrm{rank}(A^1) = 1$ and $\mathrm{rank}(A^2) = 1$ is not feasible. (★)

This follows from a proof by contradiction. Assume that $\mathrm{rank}(A^1) = \mathrm{rank}(A^2) = 1$. Since $\mathrm{rank}(A^1) = 1$, Lemma C.8 for the case $j = 1$ states that there exist the two distinct solutions

$$(F^{-1}f^{1,q})_t = \begin{cases} e^{i\psi_q}, & t \in \mathcal{S}, \\ e^{i(\psi_q + (-1)^q \rho)}, & t \in \mathcal{S}^c, \end{cases} \qquad q = 1, 2,$$

with $\mathcal{S}, \mathcal{S}^c, \psi_1, \psi_2$ and ρ as defined in the proof of Lemma C.8.

According to (8.24), the entries of the second diagonal can have three different values,

$$(F^{-1}f^{2,q})_t = (F^{-1}f^{1,q})_t(F^{-1}f^{1,q})_{t-1}$$

$$= \begin{cases} e^{i2\psi_q}, & t \in \mathcal{S},\ t-1 \in \mathcal{S}, \\ e^{i(2\psi_q+(-1)^q\rho)}, & t \in \mathcal{S},\ t-1 \in \mathcal{S}^c \\ & \text{or } t \in \mathcal{S}^c,\ t-1 \in \mathcal{S}, \\ e^{i(2\psi_q+(-1)^q 2\rho)}, & t \in \mathcal{S}^c,\ t-1 \in \mathcal{S}^c, \end{cases} \tag{8.25}$$

for both $q = 1$ and $q = 2$. However, $\mathrm{rank}(A^2) = 1$ together with Lemma C.8 yields that the second diagonal can only have two different values. We also know $\mathcal{S}, \mathcal{S}^c \neq \varnothing$, so there is at least one $t \in [d]_0$ with

$$\left(F^{-1}f^{1,q}\right)_t = e^{i\psi_q} \quad \text{and} \quad \left(F^{-1}f^{1,q}\right)_{t-1} = e^{i(\psi_q+(-1)^q\rho)}.$$

Thus, we conclude that the second diagonal has at least one entry of value

$$(F^{-1}f^{2,q})_t = e^{i\psi_q} \cdot e^{i(\psi_q+(-1)^q\rho)} = e^{i(2\psi_q+(-1)^q\rho)}.$$

Consequently,

$$(F^{-1}f^{2,q})_t \in \left\{e^{i(2\psi_q+(-1)^q\eta\rho)}, e^{i(2\psi_q+(-1)^q(\eta+1)\rho)}\right\}, \quad q = 1, 2,$$

where η is either zero or one for both $q = 1$ and $q = 2$, since this depends on the index set $\mathcal{S}$ which is the same for both values of q.

Lemma C.8 for $j = 1$ provides that

$$\psi_1 = \psi_2 + \rho + \pi + 2\pi\zeta_1, \tag{8.26}$$

and, for $j = 2$,

$$2\psi_1 - t\rho = 2\psi_2 + \eta\rho + \rho + \pi + 2\pi\zeta_2 \tag{8.27}$$

for some $\zeta_1, \zeta_2 \in \mathbb{Z}$.

Bringing $\psi_1 - \psi_2$ to the left–hand side and the remaining variables to the right-hand side in both (8.26) and (8.27) results in

$$\rho + \pi + 2\pi\zeta_1 = \frac{1}{2}\left((2\eta+1)\rho + \pi + 2\pi\zeta_2\right),$$

so that

$$(2\eta - 1)\rho = \pi + 2\pi(2\zeta_1 - \zeta_2).$$

For both $\eta = 0$ and $\eta = 1$, this requires $\rho \in \{-\pi, \pi\}$, which is not possible according to

Lemma C.8. Hence, the combination $\operatorname{rank}(A^1) = \operatorname{rank}(A^2) = 1$ is not feasible, and (★) is proven.

We continue with the case $\operatorname{rank}(A^2) = 0$ and $\operatorname{rank}(A^1) = 1$ and show that this happens only for objects of type (ii) in Theorem 8.4.1. From the proof of Theorem 8.4.1 it is clear that such objects can never be recovered uniquely.

Let $x^q \in \mathbb{C}^d$ be defined as in Theorem 8.4.1. We prove the following statement:

$$\begin{gathered}\text{If } \operatorname{rank}(A^1) = 1 \text{ and } \operatorname{rank}(A^2) = 0, \text{ then } d \text{ is even and } x \sim x^q \\ \text{with } q = 1 \text{ or } q = 2 \text{ for some } m \in \mathbb{Z} \text{ and } \rho \in (-\pi, \pi) \setminus \{0\}.\end{gathered} \tag{★★}$$

The condition $\operatorname{rank}(A^2) = 0$ together with Lemma C.6 provides that $F^{-1}f^2$ is a constant vector. According to Lemma C.8 for $j = 1$, the entries of $F^{-1}f^{1,q}$ can attain two different values, as seen before. Set $(F^{-1}f^{1,q})_0 := e^{i\psi_q}$ for some $\psi_q \in [0, 2\pi),\ q = 1, 2$, and the other entry value that appears at least once we set as $e^{i(\psi_q + (-1)^q \rho)}$ for $\rho \in (-\pi, \pi) \setminus \{0\}$, for both $q = 1, 2$.

Using the knowledge of the first diagonals $F^{-1}f^{1,q}$ and the fact that $F^{-1}f^2$ is constant, (8.24) yields that

$$(F^{-1}f^2)_t = e^{i(2\psi_q + (-1)^q \rho)} \tag{8.28}$$

for all $t \in [d]_0$, for either $q = 1$ or $q = 2$. This is only possible if

$$\left(F^{-1}f^{1,q}\right)_t = \begin{cases} e^{i\psi_q}, & t \text{ even}, \\ e^{i(\psi_q + (-1)^q \rho)}, & t \text{ odd}. \end{cases} \tag{8.29}$$

If d is odd, this means $x_{d-1}\overline{x}_{d-2} = e^{i\psi_q}$, and hence

$$x_0\overline{x}_{d-2} = x_0\overline{x}_{d-1}x_{d-1}\overline{x}_{d-2} = e^{i2\psi_q}.$$

Compared with (8.28), this requires $\rho \in 2\pi\mathbb{Z}$. Then, it is $\left(F^{-1}f^{1,q}\right)_t = e^{i\psi_q}$ for all $t \in [d]_0$, but a constant diagonal contradicts $\operatorname{rank}(A^1) = 1$ as proven in Lemma C.6. We conclude that $\operatorname{rank}(A^1) = 1$ and $\operatorname{rank}(A^2) = 0$ can only happen if the object's dimension d is even.

Let d be even. Since

$$\prod_{t=0}^{d-1} x_t\overline{x}_{t-1} = \prod_{t=0}^{d-1} |x_t|^2 = 1,$$

(8.29) gives

$$\left(e^{i\psi_1}\right)^{\frac{d}{2}} \left(e^{i(\psi_1 - \rho)}\right)^{\frac{d}{2}} = 1,$$

consequently we have

$$d\psi_1 - \frac{d}{2}\rho = 2\pi\zeta \tag{8.30}$$

for some $\zeta \in \mathbb{Z}$. Combined with equation (C.2) in Lemma C.8, this further provides

$$\psi_2 = -\frac{\rho}{2} - \pi + \frac{1}{d} 2\pi\zeta + 2\pi\tilde{\zeta} \tag{8.31}$$

for some $\tilde{\zeta} \in \mathbb{Z}$.

According to Lemma C.9, the first off-diagonal $x \circ S_1\overline{x}$ is sufficient to reconstruct x via the recursive relation $x_t = (x \circ S_1\overline{x})_t \cdot x_{t-1}$. Knowing this diagonal by (8.29), we derive the two possible solutions entry by entry and, recursively, find

$$x_t^q = e^{it(\psi_q + (-1)^q \rho)} \cdot \begin{cases} e^{-(-1)^q i \frac{t}{2}\rho}, & t \text{ even}, \\ e^{-(-1)^q i \frac{t+1}{2}\rho}, & t \text{ odd}, \end{cases} \qquad q = 1, 2.$$

Inserting (8.30) and (8.31), we obtain that x is equal to x^q for $q = 1$ or $q = 2$ with entries

$$x_t^q = e^{\frac{2\pi i t m}{d}} \cdot \begin{cases} 1, & t \text{ even}, \\ -(-1)^q e^{-(-1)^q \frac{1}{2} i\rho}, & t \text{ odd}, \end{cases} \qquad q = 1, 2,$$

for some $m \in \mathbb{Z}$ and $\rho \in (-\pi, \pi) \setminus \{0\}$, showing that (★★) holds true.

Note that in Theorem 8.4.1 (*ii*), we do not exclude $\rho = 0$ as in this case the objects of type (*ii*) coincide with the objects of type (*i*). Here, however, $\rho = 0$ is excluded as it is not obtained in case $\operatorname{rank}(A^1) = 1$.

The last instance to study is $\operatorname{rank}(A^2) = 2$.

If $\operatorname{rank}(A^2) = 2$ and d is odd, d is coprime with 2 and Lemma C.9 for $j = 2$ tells that x can be uniquely recovered.

In the case that d is even, we can prove the following:

If $\operatorname{rank}(A^1) = 1$, $\operatorname{rank}(A^2) = 2$, and d is even,
the ground-truth solution x can be uniquely recovered. (★★★)

If $\operatorname{rank}(A^2) = 2$, the system (8.14) for $j = 2$ has one solution, hence f_0^2 can be recovered uniquely and the second diagonal $F^{-1}f^2$ can be reconstructed.

Since d is even, it is not coprime with $j = 2$ and there are multiple vectors $\tilde{x}$ satisfying the condition given by the second diagonal, that is, $\tilde{x} \circ S_2\overline{\tilde{x}} = F^{-1}f^2$; see Remark C.10. However, we can use f^2 to select the ground-truth solution out of the two possible solutions $f^{1,1} \neq f^{1,2}$ obtained via Lemma C.8 for $j = 1$. In the following, we show that this is indeed possible.

Assume that, via (8.24), $f^{1,1}$ and $f^{1,2}$ result in the same diagonal $F^{-1}f^2$, that means, they cannot be distinguished by using the extra information given by f^2. The entries of $F^{-1}f^2$ can be constructed from $f^{1,1}$ and $f^{1,2}$ and can attain the three possible values as described in (8.25) for both $q = 1$ and $q = 2$.

As $\operatorname{rank}(A^2) \neq 0$, according to Lemma C.6 it is not possible that $F^{-1}f^2$ is constant. Hence, there is at least one $t \in [d]_0$ with

$$(F^{-1}f^2)_t \in \left\{e^{i2\psi_q}, e^{i(2\psi_q+(-1)^q 2\rho)}\right\}$$

for both $q = 1, 2$. Since we assumed that $F^{-1}f^2$ is obtained from both $f^{1,1}$ and $f^{1,2}$, either

$$(i) \quad e^{i2\psi_1} = e^{i2\psi_2} \quad \text{or} \quad (ii) \quad e^{i(2\psi_1-2\rho)} = e^{i(2\psi_2+2\rho)}$$

must hold. Suppose (i) is true. Then

$$2\psi_1 = 2\psi_2 + 2\pi\zeta$$

for some $\zeta \in \mathbb{Z}$, that is,

$$\psi_1 = \psi_2 + \pi\zeta.$$

If ζ is even, this means that $\psi_1 = \psi_2$ as $\psi_1, \psi_2 \in (0, 2\pi]$. This, however, is impossible according to Lemma C.8.

For ζ odd, we get

$$\psi_1 = \psi_2 + \pi + 2\pi\zeta_1$$

with $\zeta_1 \in \mathbb{Z}$. Combining this with (C.2) yields $\rho \in 2\pi\mathbb{Z}$, which is again not possible by Lemma C.8.

If property (ii) would be true, then

$$2\psi_1 - 2\rho = 2\psi_2 + 2\rho + 2\pi\zeta_2$$

for some $\zeta_2 \in \mathbb{Z}$, and combined with (C.2) we find that $\rho \in \pi\mathbb{Z}$. Once again, this is not feasible according to Lemma C.8.

As both (i) and (ii) are not possible, only one of $f^{1,1}$ or $f^{1,2}$ is compatible with the uniquely recovered f^2.

Therefore, involving $j = 2$ helps to distinguish the two possible diagonals caused by $\operatorname{rank}(A^1) = 1$ into the diagonal corresponding to the ground-truth solution and the wrong one. Then, knowing the first off-diagonal, x can be recovered uniquely; see Lemma C.9 (ii).

We conclude that this exhausts all possible events, and hence only from $j = 1$ and $j = 2$ it can be decided whether x can be recovered uniquely, up to the global phase, and which objects cannot be uniquely recovered from ptychographic data with background. Theorem 8.4.2 summarizes these results. □

Comparing Theorem 8.4.2 with its analog for distortion-free ptychographic measurements reported in Theorem 4.4.5, it turns out that for measurements with background more diagonals have to be employed to guarantee uniqueness of reconstruction. More precisely, the Fourier coefficients of the second off-diagonal ($j = 2$) have to be additionally taken into account to compensate for the coefficients discarded due to the distortion. Note also that the main diagonal ($j = 0$) only characterizes the magnitudes which are a priori known for phase objects. In contrast to the background-free scenario where unique reconstruction is guaranteed for all non-vanishing objects by Theorem 4.4.5, there are two types of phase objects that are non-vanishing but are not uniquely determined by their ptychographic measurements with background. We would like to stress that this non-uniqueness cannot be circumvented by involving further diagonals ($j > 2$) into the algorithm as Theorem 8.4.1 assures.

8.5. Numerical experiments

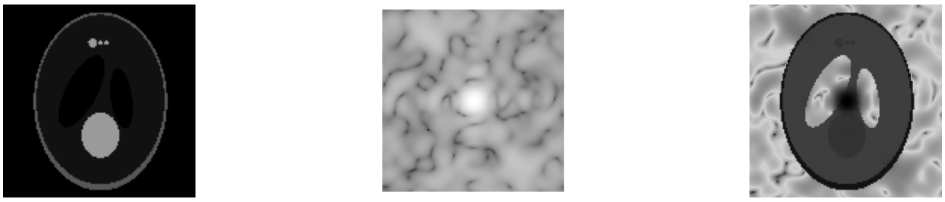
In this section, we extensively test the performance of the developed methods, Algorithm 5 and Algorithm 6, in a variety of numerical experiments.

Up to now, we focused on a one-dimensional setting for the theoretical consideration to avoid overcomplicating the notation. However, the algorithms can easily be extended to a two-dimensional version, as done for example in [32], using the 2D analogies of the occurring operators. Being driven by real-world imaging applications, we investigate in the following experiments for 2D images. It is not yet known whether the uniqueness guarantee in Theorem 8.4.2 can be generalized to higher dimensions, but the experiments discussed in this section give a preliminary conjecture for this question.

We test the algorithms on the synthetic data shown in Figure 8.1. The ground-truth object is a 128×128 image with the boat as the magnitude and the cameraman as the argument. For the phase object, we use the same argument but replace the magnitude with a matrix of ones. To enable sufficiently fast computation for the experiments, smaller versions of the same objects were used for the experiments reported in Tables 8.2 and 8.3, Figure 8.5 and Table 8.4. All displayed reconstructions use the color scales given in Figure 8.1. The window is localized on a 16×16 nonzero block, using the randomly perturbed Gaussian bell shown in Figure 8.1 (c). The perturbation is required to guarantee Assumption 1, which does not hold for symmetric windows; see Remark 4.4.3.

(a) Magnitude and argument (b) Argument (c) Magnitude and argument

Figure 8.1.: (a) Ground-truth object, (b) ground-truth phase object and (c) nonzero block of window.



(a) (b) (c)

Figure 8.2.: (a) Background, (b) diffraction pattern and (c) diffraction pattern distorted by this background (in logarithmic scale).

We distort the measurements by adding the Shepp-Logan phantom in Figure 8.2 (a) as a background signal to each diffraction pattern. Figure 8.2 (c) illustrates an exemplary diffraction pattern with background, while (b) shows the corresponding clear diffraction pattern (in a different scale).

Various distortion levels are simulated by adding different multiples of the phantom to the same diffraction patterns. The level of distortion by a background or of potential additional random noise is quantified via the ratio

$$\|Y - \tilde{Y}\|_F / \|Y\|_F \,,$$

where Y is distortion-free ptychographic data (4.10) and $\tilde{Y}$ is a distorted version of Y.

To compare reconstruction results, we use the two metrics relative reconstruction error and relative measurement error, that is,

$$\frac{dist(x, \tilde{x})}{\|x\|_2} \quad \text{and} \quad \frac{\|Y - Y^{rec}\|_F}{\|Y\|_F},$$

where x is the ground-truth, $\tilde{x}$ is the reconstruction, and $Y^{rec}_{k,r} := |(F\,[S_{-r}\tilde{x} \circ w])_k|^2$, for all $k, r \in [d]_0$, are the corresponding simulated measurements.

We start with comparing the performance of the 'vanilla' WDD algorithm (Algorithm 3) with the results of our algorithms, described in Algorithm 5 for general objects and in Algorithm 6 for phase objects.

In the two-dimensional setting, the parameter γ in Assumption 1 for any $\gamma < \delta$ includes all diagonals corresponding to tuples $j = (j_1, j_2)$ in

$$\mathcal{D}_\gamma := \{(j_1, j_2) \in \mathbb{Z}^2 : 0 \leq j_1 < \gamma,\ -\gamma < j_2 < \gamma,\ -\gamma < j_1 + j_2 < \gamma\}.$$

If all diagonals $\{(j_1, j_2) \in \mathbb{Z}^2 : 0 \leq j_1 < \delta,\ -\delta < j_2 < \delta\}$ are used, we denote this as 'all'.

First, we showcase the performance of Algorithm 5 with an experiment involving the object in Figure 8.1 (a). We compare the reconstructions obtained by this approach with the results of Algorithm 3. In addition, we implemented the preprocessing algorithm proposed in [130] to reduce the background distortion from the measurements, and apply Algorithm 3 to its output. The results are shown in Figure 8.3. As anticipated, the results

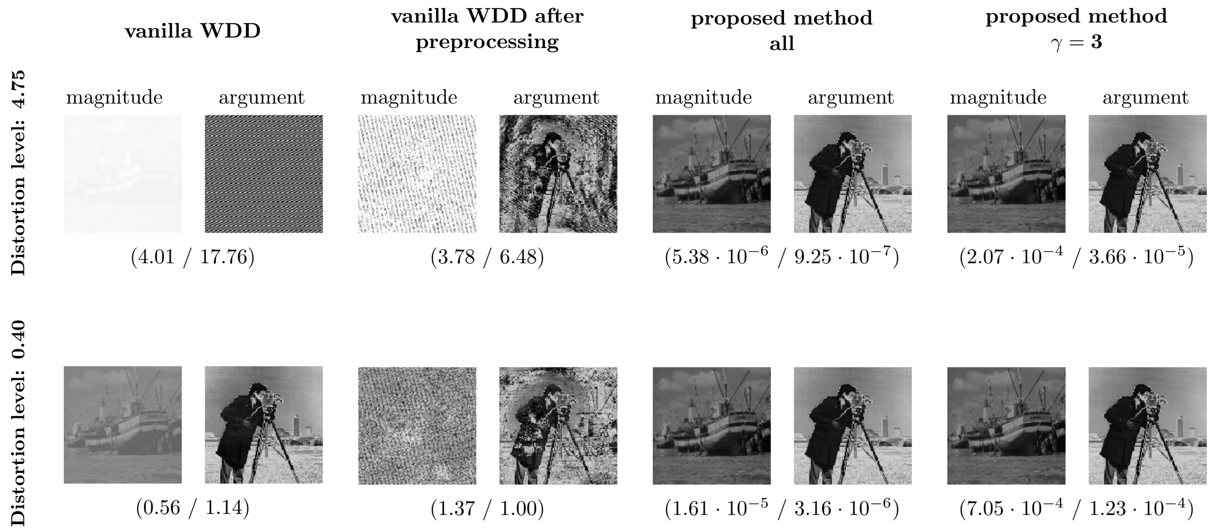


Figure 8.3.: Reconstruction results of the vanilla WDD algorithm (Algorithm 3), without and with preprocessing, and Algorithm 5 for different distortion levels and different amounts of diagonals incorporated into the experiment (all or $\gamma = 3$). We compare the results in terms of (relative reconstruction error / relative measurement error).

of vanilla WDD are visibly affected by the background distortion. We recognize here the constant additive error in the (squared) magnitudes as described in (8.2). The preprocessing approach seems to have the potential to improve the reconstruction. At least in the case of a higher distortion level, we observe better quality metrics, and details of the cameraman are visible after preprocessing while all information was lost in the result of the plain vanilla WDD. However, the example with a lower distortion level shows that the preprocessing can also deteriorate the reconstruction result. This can be explained by the fact that the same algorithm parameters were selected for the different distortion levels, which gives the impression that the method is highly sensitive to the parameter choice. In contrast, the proposed method (Algorithm 5) discards the background distorted information and hence is independent of the distortion level. This is verified by the experiment, also in the case where only the diagonals in $\mathcal{D}_3$ are employed for the reconstruction.

Next, we run Algorithm 6 on data corresponding to the phase object in Figure 8.1 (b). Again, the reconstructions are compared with the outcome of Algorithm 3 with and without additional preprocessing [130]. Since these methods are not expected to return phase objects, we project the obtained reconstructions onto $\mathbb{T}^{d \times d}$ before determining the quality metrics. The results reported in Figure 8.4 corroborate that the proposed algorithm works well in terms of background removal.

Additionally, we compare the performance of our algorithms with the ADP algorithm proposed in [24]. ADP stands for ADMM (alternating direction method of multipliers) denoising for phase retrieval and is designed for object reconstruction from ptychographic

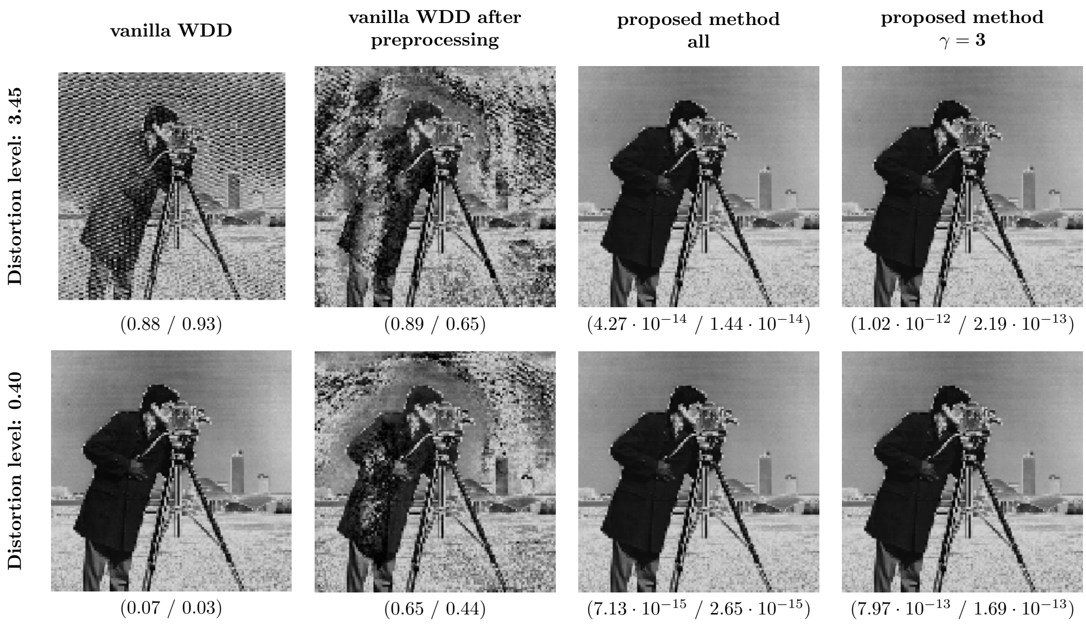


Figure 8.4.: Reconstruction results of the vanilla WDD algorithm (Algorithm 3), without and with preprocessing, and Algorithm 6 for different distortion levels and different amounts of diagonals incorporated into the experiment (all or $\gamma = 3$). We compare the results in terms of (relative reconstruction error / relative measurement error).

data with background signal and additional random noise. The ADP algorithm is built on an optimization approach. We run ADP and Algorithm 5 for the test object in Figure 8.1 (a) and list the observed run times, relative reconstruction and measurement errors in Table 8.2. As before, we compare the two versions of Algorithm 5 where either all diagonals are used or only those in $\mathcal{D}_3$. In both cases, our algorithm is more precise, and working only with the data corresponding to the set $\mathcal{D}_3$ requires the shortest run time. Analogous trials are performed for phase objects together with Algorithm 6. The respective results are summarized in Table 8.3. The experiments show that in most cases WDD recovers the ground-truth object up to a 64-bit double-precision, while in some trials the errors are worse. We suspect that this is a result of a numerical error accumulation from the larger number of diagonals used or a suboptimal phase synchronization. On the contrary, ADP always requires more iterations and computation time to match the same precision.

Since in real-world ptychographic experiments the measurements are additionally affected by particle counting noise, as discussed in Section 3.1, we want to also test the performance of the WDD based methods for background distorted measurements corrupted by Poisson noise. More precisely, we aim to reconstruct x from

$$\left(\tilde{Y}^{\mathrm{P}}\right)_{k,r} \sim \frac{1}{\kappa} Poisson\left(\kappa \tilde{Y}_{k,r}\right), \quad k, r \in [d]_0, \tag{8.32}$$

where the scaling parameter $\kappa > 0$ controls the distortion level.

d	δ	run time (in seconds)			rel. reconstruction error			rel. measurement error		
		WDD (all)	WDD ($\gamma = 3$)	ADP	WDD (all)	WDD ($\gamma = 3$)	ADP	WDD (all)	WDD ($\gamma = 3$)	ADP
64	8	17	5	61	$4.9 \cdot 10^{-14}$	$2.3 \cdot 10^{-13}$	0.91	$1.3 \cdot 10^{-14}$	$4.5 \cdot 10^{-14}$	0.88
	16	67	9	66	$2.0 \cdot 10^{-13}$	$2.1 \cdot 10^{-12}$	0.85	$4.1 \cdot 10^{-14}$	$6.7 \cdot 10^{-13}$	0.44
96	8	52	36	942	$1.6 \cdot 10^{-5}$	$9.2 \cdot 10^{-4}$	0.91	$4.5 \cdot 10^{-6}$	$2.5 \cdot 10^{-4}$	0.87
	16	172	54	865	$4.7 \cdot 10^{-6}$	$6.9 \cdot 10^{-4}$	0.90	$9.5 \cdot 10^{-7}$	$2.0 \cdot 10^{-4}$	0.36

Table 8.2.: Comparison of Algorithm 5 (WDD) with 10 iterations of the ADP algorithm initialized with a constant 1 matrix [24] for ptychographic measurements with background causing a distortion level of size approximately 3.5.

d	δ	run time (in seconds)			rel. reconstruction error			rel. measurement error		
		WDD (all)	WDD ($\gamma = 3$)	ADP	WDD (all)	WDD ($\gamma = 3$)	ADP	WDD (all)	WDD ($\gamma = 3$)	ADP
64	8	5	4	65	$2.9 \cdot 10^{-14}$	$1.6 \cdot 10^{-13}$	0.94	$1.1 \cdot 10^{-14}$	$4.6 \cdot 10^{-14}$	0.37
	16	15	6	62	$4.5 \cdot 10^{-4}$	$1.4 \cdot 10^{-13}$	0.83	$3.0 \cdot 10^{-4}$	$5.7 \cdot 10^{-14}$	0.53
96	8	25	21	913	$3.0 \cdot 10^{-14}$	$9.0 \cdot 10^{-14}$	0.97	$9.6 \cdot 10^{-15}$	$1.8 \cdot 10^{-14}$	0.34
	16	46	27	775	$3.0 \cdot 10^{-4}$	$3.2 \cdot 10^{-13}$	0.89	$1.4 \cdot 10^{-4}$	$8.7 \cdot 10^{-14}$	0.45

Table 8.3.: Comparison of Algorithm 6 (WDD) with 10 iterations of the ADP algorithm initialized with a constant 1 matrix [24] for ptychographic measurements of a phase object with background causing a distortion level of size approximately 3.5.

From now on, we focus on presenting further examinations of the algorithm performance only for the more general version Algorithm 5 and the object Figure 8.1 (a).

We compare our proposed algorithm with a background subtraction approach combined with vanilla WDD. For the background subtraction, we produce an extra 'empty' diffraction pattern, which means in experimental sense that we perform one illumination without a specimen placed in the setup. This generates measurements

$$Y_k^{\text{extra}} \sim \frac{1}{\kappa} Poisson\left(\kappa(|(Fw)_k|^2 + b_k)\right), \quad k \in [d]_0.$$

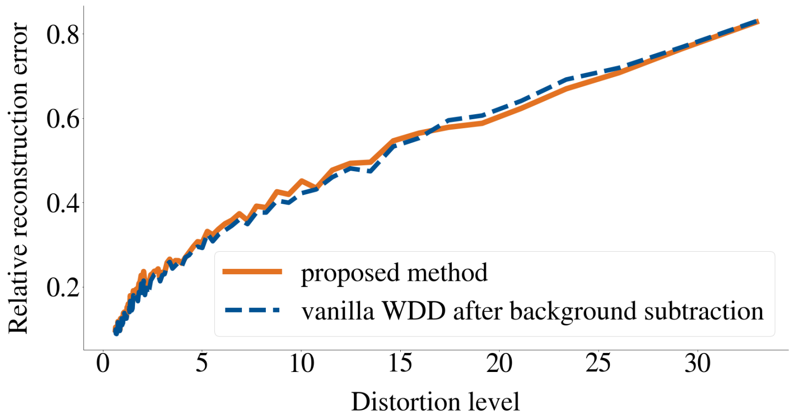


Figure 8.5.: Comparison of the performance of the vanilla WDD algorithm after background subtraction and Algorithm 5 for ptychographic measurements (8.32) with background and Poisson noise for different distortion levels.

Since we assume that the window w is known, we can compute $|(Fw)_k|^2$ to (approximately) recover the background b using Y^{extra} and then remove it from all diffraction patterns. As the Poisson noise affects also this step, we need to test how stable this approach is.

The results reported in Figure 8.5 indicate that vanilla WDD after background subtraction and our proposed method perform comparably well. However, the algorithms proposed in this paper are advantageous in that they do not require to measure the additional diffraction pattern for the background subtraction step. While this is not a concern in synthetic experiments, this can impact the experiment duration in a real-world application significantly. For a better noise stability of the reconstruction algorithms, we used a regularization technique based on singular value truncation for both WDD methods. This method is explained in detail in [85, Section 3.6.2.2].

The output of the non-iterative WDD algorithm can also be used as an initialization for iterative methods like the ADP algorithm. The benefit of ADP is that it is based on an optimization approach which incorporates the Poisson distribution, meaning that it is designed specifically for ptychographic measurements not only with background but

Initialization method / Errors after	Constant 1 object	Vanilla WDD without background subtraction	Proposed method Algorithm 5
initialization	(1 / 1)	(6.84 / 65.91)	(0.81 / 0.84)
50 iterations ADP	(0.97 / 0.99)	(1.07 / 1.01)	(0.39 / 0.39)

Table 8.4.: ADP for ptychographic data with background and Poisson noise (overall distortion level 32.96) with different initializations. We compare the results in terms of (relative reconstruction error / relative measurement error).

also with Poisson noise. Comparing the performance of ADP for different initializations, we found that consecutively applying the proposed background removing WDD method (Algorithm 5) and the ADP algorithm yields the best result; see Table 8.4. We reason that, especially in case of dominating Poisson noise, that is, in a low-count scenario as discussed in Section 3.2, it is advisable to combine the WDD approach with some iterations of ADP to enhance the reconstruction quality.

In Section 2.4 and Section 4.4, we briefly addressed that real-world applications can involve shifts of size L larger than one. While our theoretical results do no longer hold in this case, we can nevertheless perform numerical experiments to test how the reconstruction ability persists in such scenarios. We work with ptychographic data $\tilde{Y}_{k,r}$ subsampled to shifts $r \in \{0, L, 2L, \ldots\} \subset [d]_0$. Theoretical guarantees for WDD algorithms for $L > 1$ require the object to be bandlimited [32, 96] or piecewise constant [85, Section 3.6.5.1]. Neither of these configurations applies to the natural images considered here. However, a version of WDD for piecewise constant objects combined with our background removal strategy can be applied as a heuristic. Thereto, the object is divided into blocks of $L \times L$ pixels and all pixels within the same block are set to be equal. The result is a low-resolution image that can be reconstructed using the version of WDD proposed in [85, Section 3.6.5.1]. The suggested background removal step can be incorporated in this algorithm in the same way as it was implemented into the vanilla WDD (Algorithm 3). Not only results this version of WDD in low-resolution images, also the reconstruction of the discarded Fourier coefficients is no longer exact as the other Fourier coefficients are also affected by the approximation to a piecewise constant object. Hence, the reconstructions obtained by a WDD method might not be satisfactory, but initializing an iterative method that does not rely on full shift data, like ADP, with these results clearly improves the reconstruction quality; see Figure 8.6. We also compare the results for ptychographic experiments with background with the results obtained by the employed WDD heuristic for the associated background-free measurements and observe that the background distortion together with the background removal step results in only negligibly larger errors. However, it seems that increasing the step size L goes in hand with a slightly deteriorating reconstruction quality in the presence of a background signal compared to background-free measurements. We note in the displayed reconstructions that pixels corresponding to dark pieces of the boat in the magnitude image are prone to severe artifacts in the phase reconstruction. This can be explained by the fact that the respective magnitude values are small, causing a less stable approach. To improve the phase reconstruction also for these parts, more iterations of ADP could be performed.

In summary, the numerical experiments support the proposed algorithms, showing that they improve reconstruction by removing the unknown background signal. Furthermore, we observed that also in case of additional corruption or shifts by more than one pixel our suggested methods can be seen as an advancement, although they are recommended to be used in combination with an iterative method for further refinement.

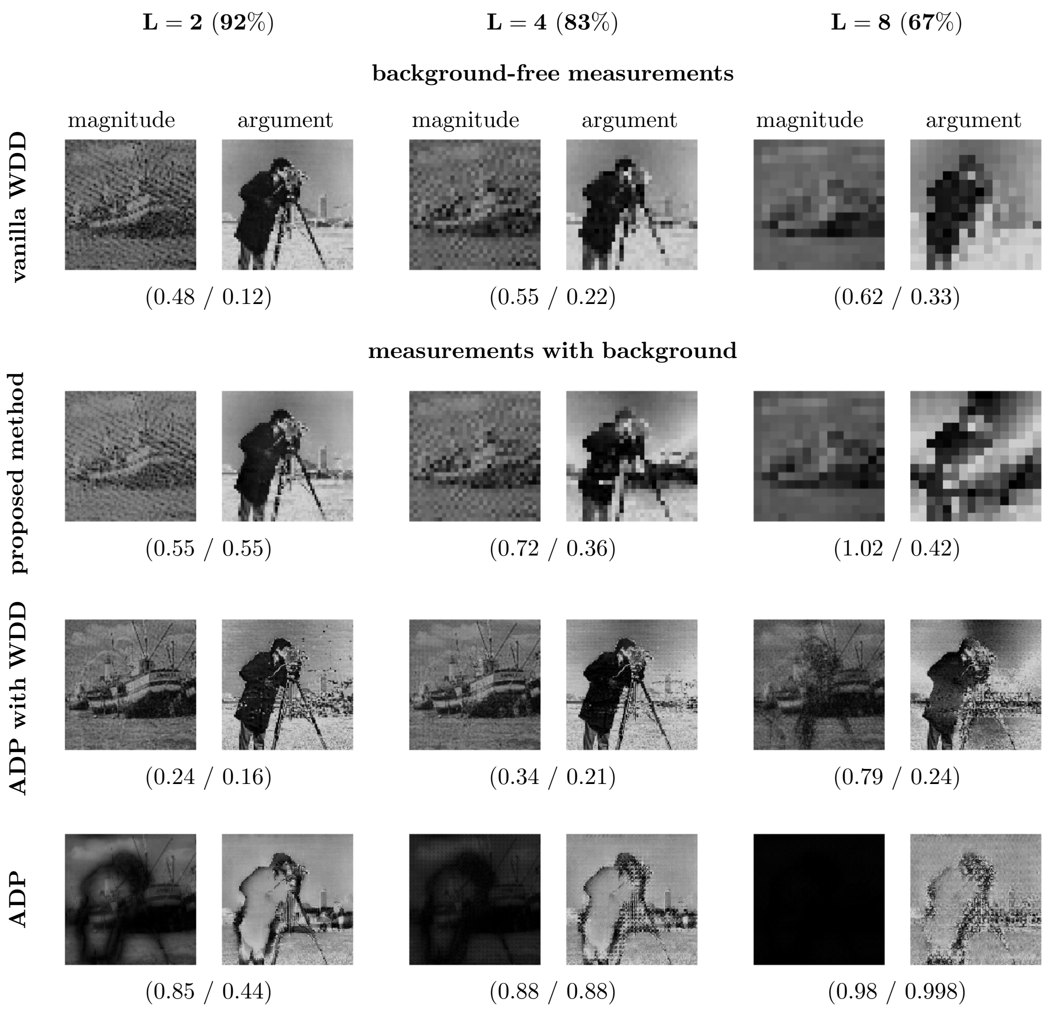


Figure 8.6.: Ptychographic experiment with different shift sizes L for the general object Figure 8.1 (a). Here, we use a window as in Figure 8.1 (c) but with support size 24×24 to satisfy Assumption 1 with $\gamma \geq 3$ also for shifts of size $L = 8$. Next to the shift sizes L, we state the overlap ratio $1 - L/\delta$ between two adjacent illumination regions. We compare the performance of vanilla WDD in case of no background distortion with the performance of Algorithm 5 (proposed method) in case of background distortion. A regularization technique with singular value truncation is applied in all cases for better robustness. Subsequently, we run ADP initialized with the result of Algorithm 5 and compare it with ADP initialized with a constant 1 object as proposed in [24]. We compare the results in terms of (relative reconstruction error / relative measurement error).

Notes and references. *All numerical experiments reported in this section were performed in Python on a MacBook Pro equipped with a 2GHz Intel Core i5 and 16 GB of memory. The code for all presented experiments is available at [88].*

The figures and tables shown in this section were first published in [87], published by the Society for Industrial and Applied Mathematics (SIAM).

9. Conclusion

We conclude this thesis with a short summary of the discussed results and, finally, pose some open questions for future research.

The first pillar of this thesis explores phaseless reconstruction from data corrupted by Poisson noise, driven by the limitations in biological imaging applications caused by the discrete nature of particle counts. We investigated state-of-the-art phase retrieval methods under Poisson noise in scenarios with moderate energy as well as in an extremely low-dose regime. For the classical spectral method, we provided recovery guarantees for Poisson measurements.

Noticing that the spectral method might not produce self-sufficient reconstructions without enormous oversampling, especially in lower dose cases, we aimed at developing a suitable version of a Wirtinger gradient descent algorithm to refine the spectral method output with some iterations of this technique. In this regard, we found an improved variance-stabilizing transformation optimized for small counts, and incorporated this to formulate a loss function appropriate for low-dose Poisson phase retrieval. For gradient descent algorithms used in combination with the considered loss functions, we established bounds on the step size required to guarantee descent in the loss.

Furthermore, we inspected the extreme regime of only zero or one counts, relevant in real-world experiments since detectors have a certain idle time. To address this extreme case, we analyzed the spectral method also for Bernoulli measurements, which can be understood as truncated Poisson measurements. To the best of our knowledge, we provided the first theoretical guarantee for phase retrieval from Bernoulli measurements. From our recovery guarantee we concluded that a tremendous oversampling would be required to obtain satisfactorily reconstructions in case of a very small dose, but also for rather large dose values as in this case information gets lost due to the truncation of the measurements. While we could discover that reconstruction is possible for noticeably lower oversampling ratios than predicted by the theory, also our empirical studies led us to the conclusion that, in extremely low-dose cases, there was still room for improvement of the reconstruction quality. Motivated by this observation, we designed new algorithmic approaches for phaseless reconstruction by interpreting the low-count measurements as one-bit data, building on methods known from the field of one-bit compressed sensing.

A key direction for future work is to examine how the information loss due to the dead time of the detector in larger dose experiments, that is, the truncation effect, could be approached in algorithmic designs. Since some of the theoretical guarantees for phase retrieval with Poisson noise were in this thesis only obtained for random frames, there is a demand for extending these concepts to the concrete applications of ptychography. Finally, we propose the analysis of the designed one-bit algorithms as an open problem for future research.

The second pillar of this manuscript focused on the distortion of ptychographic measurements by a background signal, an additional main source of measurement corruption occuring in imaging experiments. We designed two novel background removing algorithms, one for arbitrary objects and a simplified version for phase objects, and proved a uniqueness of reconstruction guarantee for the class of phase objects. Empirically, we could verify that the algorithms are able to remove the impact of the background distortion in reconstructions. For data sets that are not covered by the theoretical results, for example with additional Poisson noise or in case of undersampled data, the numerical experiments clearly demonstrated that initializing an iterative algorithm with the results of our proposed methods improves significantly upon a sole use of the iterative method with an uninformed initialization.

It is left for future work to establish a theoretical guarantee for non-phase objects. Further interesting open problems concern uniqueness guarantees and arising ambiguities if multiple coefficients per diagonal were lost due to erroneous data.

Bibliography

[1] Rima Alaifari and Matthias Wellershoff. "Stability estimates for phase retrieval from discrete Gabor measurements". In: *Journal of Fourier Analysis and Applications* 27 (2021), pp. 1–31. DOI: 10.1007/s00041-020-09802-1.

[2] Boris Alexeev et al. "Phase Retrieval with Polarization". In: *SIAM Journal on Imaging Sciences* 7.1 (2014), pp. 35–66. DOI: 10.1137/12089939X.

[3] Francis J Anscombe. "The transformation of Poisson, binomial and negative-binomial data". In: *Biometrika* 35.3/4 (1948), pp. 246–254. DOI: 10.2307/2332343.

[4] Lindsay A. Baker and John L. Rubinstein. "Radiation damage in electron cryomicroscopy". In: *Methods in enzymology*. Vol. 481. Elsevier, 2010, pp. 371–388. DOI: 10.1016/S0076-6879(10)81015-8.

[5] Radu Balan, Pete Casazza, and Dan Edidin. "On signal reconstruction without phase". In: *Applied and Computational Harmonic Analysis* 20.3 (2006), pp. 345–356. DOI: 10.1016/j.acha.2005.07.001.

[6] Afonso S. Bandeira et al. "Saving phase: Injectivity and stability for phase retrieval". In: *Applied and Computational Harmonic Analysis* 37.1 (2014), pp. 106–125. DOI: 10.1016/j.acha.2013.10.002.

[7] Arya Bangun et al. "Wigner Distribution Deconvolution Adaptation for Live Ptychography Reconstruction". In: *Microscopy and Microanalysis* 29.3 (Mar. 2023), pp. 994–1008. DOI: 10.1093/micmic/ozad021.

[8] Maurice S Bartlett. "The square root transformation in analysis of variance". In: *Supplement to the Journal of the Royal Statistical Society* 3.1 (1936), pp. 68–78. DOI: 10.2307/2983678.

[9] Tobias Beetz and Chris Jacobsen. "Soft X-ray radiation-damage studies in PMMA using a cryo-STXM". In: *Journal of Synchrotron Radiation* 10.3 (2003), pp. 280–283. DOI: 10.1107/S0909049503003261.

[10] Robert Beinert and Gerlind Plonka. "Ambiguities in one-dimensional discrete phase retrieval from Fourier magnitudes". In: *Journal of Fourier Analysis and Applications* 21 (2015), pp. 1169–1198. DOI: 10.1007/s00041-015-9405-2.

[11] Tamir Bendory, Chi-Yu Cheng, and Dan Edidin. “Near-optimal bounds for signal recovery from blind phaseless periodic short-time Fourier transform”. In: *Journal of Fourier Analysis and Applications* 29.1 (2023), p. 1. DOI: 10.1007/s00041-022-09983-x.

[12] Tamir Bendory, Dan Edidin, and Yonina C. Eldar. “Blind phaseless short-time Fourier transform recovery”. In: *IEEE Transactions on Information Theory* 66.5 (2019), pp. 3232–3241. DOI: 10.1109/TIT.2019.2947056.

[13] Mario Bertero et al. “A discrepancy principle for Poisson data”. In: *Inverse problems* 26.10 (2010), p. 105004. DOI: 10.1088/0266-5611/26/10/105004.

[14] Francesca Bevilacqua et al. “Nearly exact discrepancy principle for low-count Poisson image restoration”. In: *Journal of Imaging* 8.1 (2021). DOI: 10.3390/jimaging8010001.

[15] Oier Bikondoa and Dina Carbone. “On Compton scattering as a source of background in coherent diffraction imaging experiments”. In: *Journal of Synchrotron Radiation* 28.2 (2021), pp. 538–549. DOI: 10.1107/S1600577521000722.

[16] Irena Bojarovska and Axel Flinth. “Phase retrieval from Gabor measurements”. In: *Journal of Fourier Analysis and Applications* 22.3 (2016), pp. 542–567. DOI: 10.1007/s00041-015-9431-0.

[17] Pantelis Bouboulis. “Wirtinger’s calculus in general Hilbert spaces”. In: *arXiv preprint arXiv:1005.5170* (2010).

[18] Oliver Bunk et al. “Influence of the overlap parameter on the convergence of the ptychographical iterative engine”. In: *Ultramicroscopy* 108.5 (2008), pp. 481–487.

[19] Emmanuel J. Candeè, Xiaodong Li, and Mahdi Soltanolkotabi. “Phase retrieval via Wirtinger flow: Theory and algorithms”. In: *IEEE Transactions on Information Theory* 61.4 (2015), pp. 1985–2007. DOI: 10.1109/TIT.2015.2399924.

[20] Emmanuel J. Candès and Xiaodong Li. “Solving quadratic equations via PhaseLift when there are about as many equations as unknowns”. In: *Foundations of Computational Mathematics* 14 (2014), pp. 1017–1026. DOI: 10.1007/s10208-013-9162-z.

[21] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming”. In: *Communications on Pure and Applied Mathematics* 66.8 (2013), pp. 1241–1274. DOI: 10.1002/cpa.21432.

[22] Emmanuel J. Candès et al. “Phase retrieval via matrix completion”. In: *SIAM review* 57.2 (2015), pp. 225–251. DOI: 10.1137/151005099.

[23] Huibin Chang, Pablo Enfedaque, and Stefano Marchesini. "Blind ptychographic phase retrieval via convergent alternating direction method of multipliers". In: *SIAM Journal on Imaging Sciences* 12.1 (2019), pp. 153–185. DOI: 10.1137/18M1188446.

[24] Huibin Chang et al. "Advanced denoising for X-ray ptychography". In: *Optics express* 27.8 (2019), pp. 10395–10418. DOI: 10.1364/OE.27.010395.

[25] Huibin Chang et al. "Total variation-based phase retrieval for Poisson noise removal". In: *SIAM Journal on Imaging Sciences* 11.1 (2018), pp. 24–55. DOI: 10.1137/16M1103270.

[26] Henry N. Chapman. "Phase-retrieval X-ray microscopy by Wigner-distribution deconvolution". In: *Ultramicroscopy* 66.3-4 (1996), pp. 153–172. DOI: 10.1016/S0304-3991(96)00084-8.

[27] Junren Chen and Ming Yuan. "One-Bit Phase Retrieval: Optimal Rates and Efficient Algorithms". In: *arXiv preprint arXiv:2405.04733* (2024).

[28] Yuxin Chen and Emmanuel J Candès. "Solving random quadratic systems of equations is nearly as easy as solving linear systems". In: *Communications on pure and applied mathematics* 70.5 (2017), pp. 822–883. DOI: 10.1002/cpa.21638.

[29] Ziyi Chen et al. "Generalized-Smooth Nonconvex Optimization is As Efficient As Smooth Nonconvex Optimization". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 5396–5427.

[30] Richard M. Clare et al. "Characterization of near-field ptychography". In: *Optics express* 23.15 (2015), pp. 19728–19742. DOI: 10.1364/OE.23.019728.

[31] Aldo Conca et al. "An algebraic characterization of injectivity in phase retrieval". In: *Applied and Computational Harmonic Analysis* 38.2 (2015), pp. 346–356. DOI: 10.1016/j.acha.2014.06.005.

[32] Cyril Cordor et al. "Fast 2D Phase Retrieval using Bandlimited Masks". In: *28th European Signal Processing Conference (EUSIPCO 2020)*. IEEE, 2020, pp. 980–984. DOI: 10.23919/Eusipco47968.2020.9287439.

[33] Benedikt Diederichs, Frank Filbir, and Patricia Römer. "Wirtinger gradient descent methods for low-dose Poisson phase retrieval". In: *Inverse Problems* 40.12 (2024), p. 125030. DOI: 10.1088/1361-6420/ad97d7.

[34] Benedikt Diederichs et al. "Exact inversion of partially coherent dynamical electron scattering for picometric structure retrieval". In: *Nature Communications* 15.1 (2024), p. 101. DOI: 10.1038/s41467-023-44268-x.

[35] Sjoerd Dirksen, Hans Christian Jung, and Holger Rauhut. “One-bit compressed sensing with partial Gaussian circulant matrices”. In: *Information and Inference: A Journal of the IMA* 9.3 (2020), pp. 601–626. DOI: 10.1093/imaiai/iaz017.

[36] Sjoerd Dirksen et al. “Recovery guarantees for low-count Poisson and Bernoulli phase retrieval”. In: *15th International Conference on Sampling Theory and Applications*. 2025.

[37] Sjoerd Dirksen et al. “Spectral method for low-dose Poisson and Bernoulli phase retrieval”. In: *arXiv preprint arXiv:2502.13263* (2025).

[38] Bradley Efron. “Transformation theory: How normal is a family of distributions?” In: *The Annals of Statistics* (1982), pp. 323–339.

[39] Yonina C. Eldar and Shahar Mendelson. “Phase retrieval: Stability and recovery guarantees”. In: *Applied and Computational Harmonic Analysis* 36.3 (2014), pp. 473–494. DOI: 10.1016/j.acha.2013.08.003.

[40] Yonina C. Eldar et al. “Sparse phase retrieval from short-time Fourier measurements”. In: *IEEE Signal Processing Letters* 22.5 (2014), pp. 638–642. DOI: 10.1109/LSP.2014.2364225.

[41] Albert Fannjiang and Pengwen Chen. “Blind ptychography: uniqueness and ambiguities”. In: *Inverse Problems* 36.4 (2020), p. 045005. DOI: 10.1088/1361-6420/ab6504.

[42] William Feller. “An Introduction to Probability Theory and Its Applications”. In: *Wiley* 1 (1957). DOI: 10.1017/S0020268100037586.

[43] Frank Filbir, Felix Krahmer, and Oleh Melnyk. “On recovery guarantees for angular synchronization”. In: *Journal of Fourier Analysis and Applications* 27.2 (2021), p. 31. DOI: 10.1007/s00041-021-09834-1.

[44] Frank Filbir and Oleh Melnyk. “Image Recovery for Blind Polychromatic Ptychography”. In: *SIAM Journal on Imaging Sciences* 16.3 (2023), pp. 1308–1337. DOI: 10.1137/22M1527155.

[45] Alessandro Foi. “Optimization of variance-stabilizing transformations”. In: *Preprint, 2009b* 94 (2009), pp. 1809–1814.

[46] Anton Forstner et al. “Well-conditioned ptychographic imaging via lost subspace completion”. In: *Inverse Problems* 36.10 (2020), p. 105009. DOI: 10.1088/1361-6420/abaf3a.

[47] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013. DOI: 10.1007/978-0-8176-4948-7.

[48] Murray F. Freeman and John W. Tukey. “Transformations related to the angular and the square root”. In: *The Annals of Mathematical Statistics* (1950), pp. 607–611.

[49] Bing Gao et al. “Perturbed amplitude flow for phase retrieval”. In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 5427–5440. DOI: 10.1109/TSP.2020.3022817.

[50] Klaus Giewekemeyer et al. “Quantitative biological imaging by ptychographic x-ray diffraction microscopy”. In: *Proceedings of the National Academy of Sciences* 107.2 (2010), pp. 529–534. DOI: 10.1073/pnas.0905846107.

[51] Robert M. Glaeser. “Limitations to significant information in biological electron microscopy as a result of radiation damage”. In: *Journal of ultrastructure research* 36.3-4 (1971), pp. 466–482. DOI: 10.1016/S0022-5320(71)80118-1.

[52] Joseph W. Goodman. *Introduction to Fourier optics*. Roberts and Company publishers, 2005.

[53] Joseph W. Goodman. *Statistical optics*. John Wiley & Sons, 2015.

[54] Philipp Grohs, Sarah Koppensteiner, and Martin Rathmair. “Phase retrieval: uniqueness and stability”. In: *SIAM Review* 62.2 (2020), pp. 301–350. DOI: 10.1137/19M1256865.

[55] Peter W. Hawkes and John C. H. Spence. *Springer Handbook of Microscopy*. Cham: Springer Nature, 2019. DOI: 10.1007/978-3-030-00069-1.

[56] Reiner Hegerl and Walter Hoppe. “Dynamische Theorie der Kristallstrukturanalyse durch Elektronenbeugung im inhomogenen Primärstrahlwellenfeld”. In: *Berichte der Bunsengesellschaft für physikalische Chemie* 74.11 (1970), pp. 1148–1154. DOI: 10.1002/bbpc.19700741112.

[57] Thorsten Hohage and Frank Werner. “Inverse problems with Poisson data: statistical regularization theory, applications and algorithms”. In: *Inverse Problems* 32.9 (2016), p. 093001. DOI: 10.1088/0266-5611/32/9/093001.

[58] Malcolm R. Howells et al. “An assessment of the resolution limitation due to radiation-damage in x-ray diffraction microscopy”. In: *Journal of electron spectroscopy and related phenomena* 170.1-3 (2009), pp. 4–12. DOI: 10.1016/j.elspec.2008.10.008.

[59] Meng Huang and Yang Wang. “Linear convergence of randomized Kaczmarz method for solving complex-valued phaseless equations”. In: *SIAM Journal on Imaging Sciences* 15.2 (2022), pp. 989–1016. DOI: 10.1137/21M1450537.

[60] Mark Iwen, Felix Krahmer, and Aditya Viswanathan. “A Minor Correction of Theorem 1.3 From [1]”. In: *https://users.math.msu.edu/users/iwenmark/Papers/PhaseLiftproof.pdf* (2015).

[61] Mark A. Iwen, Michael Perlmutter, and Mark Philip Roach. "Toward fast and provably accurate near-field ptychographic phase retrieval". In: *Sampling Theory, Signal Processing, and Data Analysis* 21.1 (2023), p. 6. DOI: 10.1007/s43670-022-00045-3.

[62] Mark A. Iwen, Aditya Viswanathan, and Yang Wang. "Fast phase retrieval from local correlation measurements". In: *SIAM Journal on Imaging Sciences* 9.4 (2016), pp. 1655–1688. DOI: 10.1137/15M1053761.

[63] Mark A. Iwen et al. "Phase retrieval from local measurements: Improved robustness via eigenvector-based angular synchronization". In: *Applied and Computational Harmonic Analysis* 48.1 (2020), pp. 415–444. DOI: 10.1016/j.acha.2018.06.004.

[64] Laurent Jacques et al. "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors". In: *IEEE transactions on information theory* 59.4 (2013), pp. 2082–2102. DOI: 10.1109/TIT.2012.2234823.

[65] Kishore Jaganathan, Yonina C. Eldar, and Babak Hassibi. "STFT phase retrieval: Uniqueness guarantees and recovery algorithms". In: *IEEE Journal of selected topics in signal processing* 10.4 (2016), pp. 770–781. DOI: 10.1109/JSTSP.2016.2549507.

[66] Daen Jannis et al. "Event driven 4D STEM acquisition with a Timepix3 detector: Microsecond dwell time and faster scans for high precision and low dose applications". In: *Ultramicroscopy* 233 (2022), p. 113423. DOI: 10.1016/j.ultramic.2021.113423.

[67] Hans Christian Jung. "Estimation of low–complexity signals using structured and quantized observations". PhD thesis. RWTH Aachen University, 2022.

[68] Karin Knudson, Rayan Saab, and Rachel Ward. "One-bit compressive sensing with norm estimation". In: *IEEE Transactions on Information Theory* 62.5 (2016), pp. 2748–2758. DOI: 10.1109/TIT.2016.2527637.

[69] Ritesh Kolte and Ayfer Özgür. "Phase retrieval via incremental truncated Wirtinger flow". In: *arXiv preprint arXiv:1606.03196* (2016).

[70] Adam P. Konijnenberg, Wim M.J. Coene, and Horst P. Urbach. "Study of cost functionals for ptychographic phase retrieval to improve the robustness against noise, and a proposal for another noise-robust ptychographic phase retrieval scheme". In: *arXiv preprint arXiv:1703.10892* (2017).

[71] Kelvin Koor et al. "A short tutorial on Wirtinger Calculus with applications in quantum information". In: *arXiv preprint arXiv:2312.04858* (2023).

[72] Berk Küçükoğlu et al. "Low-dose cryo-electron ptychography of proteins at sub-nanometer resolution". In: *Nature Communications* 15.1 (2024), p. 8062. DOI: 10.1038/s41467-024-52403-5.

[73] Triet Le, Rick Chartrand, and Thomas J. Asaki. “A variational approach to reconstructing images corrupted by Poisson noise”. In: *Journal of mathematical imaging and vision* 27.3 (2007), pp. 257–263. DOI: 10.1007/s10851-007-0652-y.

[74] Max Leo Leidl et al. “Influence of loss function and electron dose on ptychography of 2D materials using the Wirtinger flow”. In: *Micron* 185 (2024), p. 103688. DOI: 10.1016/j.micron.2024.103688.

[75] Peng Li, Tega B. Edo, and John M. Rodenburg. “Ptychographic inversion via Wigner distribution deconvolution: noise suppression and probe design”. In: *Ultramicroscopy* 147 (2014), pp. 106–113. DOI: 10.1016/j.ultramic.2014.07.004.

[76] Peng Li and Andrew Maiden. “Multi-slice ptychographic tomography”. In: *Scientific reports* 8.1 (2018), p. 2049. DOI: 10.1038/s41598-018-20530-x.

[77] Zongyu Li, Kenneth Lange, and Jeffrey A Fessler. “Poisson phase retrieval in very low-count regimes”. In: *IEEE transactions on computational imaging* 8 (2022), pp. 838–850. DOI: 10.1109/TCI.2022.3209936.

[78] Stefan Loock and Gerlind Plonka. “Phase retrieval for Fresnel measurements using a shearlet sparsity constraint”. In: *Inverse Problems* 30.5 (2014), p. 055005. DOI: 10.1088/0266-5611/30/5/055005.

[79] Yue M Lu and Gen Li. “Phase transitions of spectral initialization for high-dimensional non-convex estimation”. In: *Information and Inference: A Journal of the IMA* 9.3 (2020), pp. 507–541. DOI: 10.1093/imaiai/iaz020.

[80] Qi Luo, Hongxia Wang, and Shijian Lin. “Phase retrieval via smoothed amplitude flow”. In: *Signal Processing* 177 (2020), p. 107719. DOI: 10.1016/j.sigpro.2020.107719.

[81] Wangyu Luo, Wael Alghamdi, and Yue M Lu. “Optimal spectral initialization for signal recovery with applications to phase retrieval”. In: *IEEE Transactions on Signal Processing* 67.9 (2019), pp. 2347–2356. DOI: 10.1109/TSP.2019.2904918.

[82] Stefano Marchesini, Yu-Chao Tu, and Hau-Tieng Wu. “Alternating projection, ptychographic imaging and phase synchronization”. In: *Applied and Computational Harmonic Analysis* 41.3 (2016), pp. 815–851. DOI: 10.1016/j.acha.2015.06.005.

[83] Stefano Marchesini et al. “Augmented projections for ptychographic imaging”. In: *Inverse Problems* 29.11 (2013), p. 115009. DOI: 10.1088/0266-5611/29/11/115009.

[84] Oleh Melnyk. “Convergence properties of gradient methods for blind ptychography”. In: *arXiv preprint arXiv:2306.08750* (2023).

[85] Oleh Melnyk. “Phase Retrieval from Short-Time Fourier Measurements and Applications to Ptychography”. PhD thesis. Technische Universität München, 2023.

[86] Oleh Melnyk. "Stochastic Amplitude Flow for phase retrieval, its convergence and doppelgängers". In: *arXiv preprint arXiv:2212.04916* (2022).

[87] Oleh Melnyk and Patricia Römer. "Background Removal for Ptychography via Wigner Distribution Deconvolution". In: *SIAM Journal on Imaging Sciences* 17.3 (2024), pp. 1978–2014. DOI: 10.1137/24M1642433.

[88] Oleh Melnyk and Patricia Römer. *WDDBackground Github repository.* https://github.com/Sniper2k/WDDBackground.

[89] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. "Phase retrieval using alternating minimization". In: *Advances in Neural Information Processing Systems* 26 (2013). DOI: 10.1109%2FTSP.2015.2448516.

[90] Gergely Odor et al. "Frank-Wolfe works for non-Lipschitz continuous gradient objectives: Scalable poisson phase retrieval". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 6230–6234. DOI: https://doi.org/10.1109/ICASSP.2016.7472875.

[91] Michal Odstrcil et al. "Ptychographic imaging with a compact gas–discharge plasma extreme ultraviolet light source". In: *Optics letters* 40.23 (2015), pp. 5574–5577. DOI: 10.1364/OL.40.005574.

[92] Colin Ophus. "Four-Dimensional Scanning Transmission Electron Microscopy (4D-STEM): From Scanning Nanodiffraction to Ptychography and Beyond". In: *Microscopy and Microanalysis* 25.3 (2019), pp. 563–582. DOI: 10.1017/S1431927619000497.

[93] Philipp Pelz. "Low-dose computational phase contrast transmission electron microscopy via electron ptychography". PhD thesis. Universität Hamburg Hamburg, 2018.

[94] Philipp Michael Pelz et al. "Low-dose cryo electron ptychography via non-convex Bayesian optimization". In: *Scientific reports* 7.1 (2017), p. 9883. DOI: 10.1038/s41598-017-07488-y.

[95] Michael Perlmutter et al. "A Provably Accurate Algorithm for Recovering Compactly Supported Smooth Functions from Spectrogram Measurements". In: *28th European Signal Processing Conference (EUSIPCO 2020)*. IEEE, 2020, pp. 970–974. DOI: 10.23919/Eusipco47968.2020.9287698.

[96] Michael Perlmutter et al. "Inverting spectrogram measurements via aliased Wigner distribution deconvolution and angular synchronization". In: *Information and Inference: A Journal of the IMA* 10.4 (2021), pp. 1491–1531. DOI: 10.1093/imaiai/iaaa023.

[97] Götz E. Pfander and Palina Salanevich. "Robust Phase Retrieval Algorithm for Time-Frequency Structured Measurements". In: *SIAM Journal on Imaging Sciences* 12.2 (2019), pp. 736–761. DOI: 10.1137/18M1205522.

[98] Franz Pfeiffer. "X-ray ptychography". In: *Nature Photonics* 12.1 (2018), pp. 9–17. DOI: 10.1038/s41566-017-0072-5.

[99] Yaniv Plan and Roman Vershynin. "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach". In: *IEEE Transactions on Information Theory* 59.1 (2012), pp. 482–494. DOI: 10.1109/TIT.2012.2207945.

[100] Gerlind Plonka et al. *Numerical Fourier Analysis*. Cham: Springer, 2018. DOI: 10.1007/978-3-030-04306-3.

[101] Siméon-Denis Poisson. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Bachelier, 1837.

[102] Brian Preskitt. "Phase retrieval from locally supported measurements". PhD thesis. University of California, San Diego, 2018.

[103] Brian Preskitt and Rayan Saab. "Admissible measurements and robust algorithms for ptychography". In: *Journal of Fourier Analysis and Applications* 27 (2021), pp. 1–39. DOI: 10.1007/s00041-021-09811-8.

[104] Juliane Reinhardt. "Resonant hard X-ray ptychography for high-sensitivity imaging with chemical contrast". PhD thesis. Hamburg University, 2017.

[105] Juliane Reinhardt et al. "Beamstop-based low-background ptychography to image weakly scattering objects". In: *Ultramicroscopy* 173 (2017), pp. 52–57. DOI: 10.1016/j.ultramic.2016.11.005.

[106] M. S. Richman, T. W. Parks, and R. G. Shenoy. "Discrete-time, discrete-frequency, time-frequency analysis". In: *IEEE Transactions on Signal Processing* 46.6 (1998), pp. 1517–1527. DOI: 10.1109/78.678465.

[107] John M. Rodenburg. "Ptychography and related diffractive imaging methods". In: *Advances in imaging and electron physics* 150 (2008), pp. 87–184. DOI: 10.1016/S1076-5670(07)00003-1.

[108] John M. Rodenburg and R. H. T. Bates. "The theory of super-resolution electron microscopy via Wigner-distribution deconvolution". In: *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences* 339.1655 (1992), pp. 521–553. DOI: 10.1098/rsta.1992.0050.

[109] John M. Rodenburg and Helen M. L. Faulkner. "A phase retrieval algorithm for shifting illumination". In: *Applied physics letters* 85.20 (2004), pp. 4795–4797. DOI: 10.1063/1.1823034.

[110] John M. Rodenburg and Andrew Maiden. "Ptychography". In: *Springer Handbook of Microscopy*. Springer, 2019, pp. 819–904. DOI: 10.1007/978-3-030-00069-1_17.

[111] Patricia Römer. *LowDosePoissonPR Github repository*. https://github.com/patricia-roemer/LowDosePoissonPR.

[112] Patricia Römer, Frank Filbir, and Felix Krahmer. "On the randomized Kaczmarz algorithm for phase retrieval". In: *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE. 2021, pp. 847–851. DOI: 10.1109/IEEECONF53345.2021.9723291.

[113] Patricia Römer and Felix Krahmer. "A one-bit quantization approach for low-dose Poisson phase retrieval". In: *2024 International Workshop on the Theory of Computational Sensing and its Applications to Radar, Multimodal Sensing and Imaging (CoSeRa)*. IEEE. 2024, pp. 42–46. DOI: 10.1109/CoSeRa60846.2024.10720330.

[114] Palina Salanevich. "Stability of phase retrieval problem". In: *2019 13th International Conference on Sampling Theory and Applications (SampTA)*. IEEE. 2019, pp. 1–4. DOI: 10.1109/SampTA45681.2019.9031013.

[115] Tim Salditt, Alexander Egner, and D. Russell Luke. *Nanoscale Photonic Imaging*. Cham: Springer Nature, 2020. DOI: 10.1007/978-3-030-34413-9.

[116] David Sayre and Henry N Chapman. "X-ray microscopy". In: *Acta Crystallographica Section A: Foundations of Crystallography* 51.3 (1995), pp. 237–252. DOI: 10.1107/S0108767394011803.

[117] Mark J Schervish and Morris H DeGroot. *Probability and statistics*. 4th ed. Pearson Education London, UK, 2012.

[118] Andreas Schropp et al. "PtyNAMi: ptychographic nano-analytical microscope". In: *Journal of applied crystallography* 53.4 (2020), pp. 957–971. DOI: 10.1107/S1600576720008420.

[119] Xiaowen Shi et al. "X-ray ptychography on low-dimensional hard-condensed matter materials". In: *Applied Physics Reviews* 6.1 (2019), p. 011306. DOI: 10.1063/1.5045131.

[120] Nada Sissouno et al. "A direct solver for the phase retrieval problem in ptychographic imaging". In: *Mathematics and Computers in Simulation* 176 (2020), pp. 292–300. DOI: 10.1016/j.matcom.2019.11.003.

[121] K. Stenn and Günther F. Bahr. "Specimen damage caused by the beam of the transmission electron microscope, a correlative reconsideration". In: *Journal of ultrastructure research* 31.5-6 (1970), pp. 526–550. DOI: 10.1016/S0022-5320(70)90167-X.

[122] Yan Shuo Tan and Roman Vershynin. "Phase retrieval via randomized Kaczmarz: theoretical guarantees". In: *Information and Inference: A Journal of the IMA* 8.1 (2019), pp. 97–123. DOI: 10.1093/imaiai/iay005.

[123] Pierre Thibault and Manuel Guizar-Sicairos. "Maximum-likelihood refinement for coherent diffractive imaging". In: *New Journal of Physics* 14.6 (2012), p. 063004. DOI: 10.1088/1367-2630/14/6/063004.

[124] LHC Tippett. "2-statistical methods in textile research. Part 2-Uses of the binomial and poisson distributions". In: *Journal of the Textile Institute Transactions* 26.1 (1935), T13–T50. DOI: 10.1080/19447023508661636.

[125] Joel A Tropp et al. "Practical sketching algorithms for low-rank matrix approximation". In: *SIAM Journal on Matrix Analysis and Applications* 38.4 (2017), pp. 1454–1485. DOI: 10.1137/17M1111590.

[126] Claudio Mayrink Verdun. "Scalability in Ill-posed Machine Learning Problems: Bridging Least Squares Methods with (Non-) convex Algorithms". PhD thesis. Universitätsbibliothek der TU München, 2023.

[127] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science.* Vol. 47. Cambridge University Press, 2018. DOI: 10.1017/9781108231596.

[128] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *arXiv preprint arXiv:1011.3027* (2010).

[129] Aditya Viswanathan and Mark Iwen. "Fast angular synchronization for phase retrieval via incomplete information". In: *Wavelets and Sparsity XVI.* Vol. 9597. SPIE. 2015, pp. 281–288. DOI: 10.1117/12.2186336.

[130] Chunpeng Wang et al. "Background noise removal in x-ray ptychography". In: *Applied optics* 56.8 (2017), pp. 2099–2111. DOI: 10.1364/AO.56.002099.

[131] Gang Wang, Georgios B Giannakis, and Jie Chen. "Scalable solvers of random quadratic equations via stochastic truncated amplitude flow". In: *IEEE Transactions on Signal Processing* 65.8 (2017), pp. 1961–1974. DOI: 10.1109/TSP.2017.2652392.

[132] Gang Wang, Georgios B. Giannakis, and Yonina C. Eldar. "Solving systems of random quadratic equations via truncated amplitude flow". In: *IEEE Transactions on Information Theory* 64.2 (2017), pp. 773–794. DOI: 10.1109/TIT.2017.2756858.

[133] Max O. Wiedorn et al. "Post-sample aperture for low background diffraction experiments at X-ray free-electron lasers". In: *Journal of synchrotron radiation* 24.6 (2017), pp. 1296–1298. DOI: 10.1107/s1600577517011961.

[134] Wilhelm Wirtinger. "Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen". In: *Mathematische Annalen* 97.1 (1927), pp. 357–375. DOI: 10.1007/BF01447872.

[135] Rui Xu et al. "Accelerated Wirtinger Flow: A fast algorithm for ptychography". In: *arXiv preprint arXiv:1806.05546* (2018).

[136] Wenhui Xu et al. "Super-resolution near-field ptychography". In: *Optics Express* 28.4 (2020), pp. 5164–5178. DOI: 10.1364/OE.383986.

[137] Guan Yu. "Variance stabilizing transformations of Poisson, binomial and negative binomial distributions". In: *Statistics & Probability Letters* 79.14 (2009), pp. 1621–1629. DOI: 10.1016/j.spl.2009.04.010.

[138] Bo Zhang, Jalal M Fadili, and Jean-Luc Starck. "Wavelets, ridgelets, and curvelets for Poisson noise removal". In: *IEEE Transactions on image processing* 17.7 (2008), pp. 1093–1108. DOI: 10.1109/TIP.2008.924386.

[139] Huishuai Zhang, Yingbin Liang, and Yuejie Chi. "A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms". In: *Journal of Machine Learning Research* 18.141 (2017), pp. 1–35.

[140] Yongbing Zhang, Pengming Song, and Qionghai Dai. "Fourier ptychographic microscopy using a generalized Anscombe transform approximation of the mixed Poisson-Gaussian likelihood". In: *Optics express* 25.1 (2017), pp. 168–179. DOI: 10.1364/OE.25.000168.

Appendix

A. Probabilistic tools and auxiliary lemmas for Chapter 5 and Section 7.1

In the following, we introduce basic concepts of random matrix theory required to prove the results presented in Chapter 5 and Section 7.1, where we work with standard Gaussian random vectors.

Definition A.1. *A random variable ε is called Rademacher random variable if*

$$\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}.$$

Lemma A.1 (Symmetrization [35, Lemma C.3])**.** *Let T be a countable set. Let (Y_t), $t \in T$, be a family of random variables, and let (Y_t') be an independent copy of (Y_t). Then, for $a, b > 0$,*

$$\mathbb{P}\left(\sup_{t\in T} |Y_t| \geq a + b\right) \leq \mathbb{P}\left(\sup_{t\in T} |Y_t - Y_t'| \geq a\right) + \sup_{t\in T} \mathbb{P}\left(|Y_t| \geq b\right).$$

Definition A.2. *The p-Schatten norm of a $d \times d$ matrix A is defined as*

$$\|A\|_{C_p^d} = \left(\sum_{i=1}^{d} \sigma_i(A)^p\right)^{\frac{1}{p}},$$

that is, the ℓ_p-norm of the singular values of A.

Lemma A.2 ([128])**.** *For $p \geq \log(d)$, the p-Schatten norm and the spectral norm are equivalent. More precisely, for $p = \log(d)$,*

$$\|A\|_{C_p^d} \leq \|A\| \leq e \|A\|_{C_p^d},$$

and for $p = \infty$ they are equal.

Theorem A.3 (Non-commutative Khintchine inequality [128, Theorem 5.26])**.** *Let A_i, $i \in [m]$, be self-adjoint $d \times d$ matrices and ε_i, $i \in [m]$, independent Rademacher random vari-*

ables. For every $2 \le p < \infty$, it holds

$$\left(\mathbb{E} \left\| \sum_{i=1}^{m} \varepsilon_i A_i \right\|_{C_p^d}^p \right)^{\frac{1}{p}} \le C\sqrt{p} \left\| \sum_{i=1}^{m} A_i^2 \right\|_{C_p^d}^{\frac{1}{2}},$$

where $C > 0$ is an absolute constant.

Definition A.3. *A random variable Y is called a sub-exponential random variable if there exists $K > 0$ such that*

$$\mathbb{E}\left[\exp\left(\frac{|Y|}{K}\right)\right] \le 2.$$

The sub-exponential norm is defined as

$$\|Y\|_{\psi_1} := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{|Y|}{t}\right)\right] \le 2\right\}.$$

Example A.1. *An important example of a sub-exponential random variable is g^2 for a Gaussian random variable $g \sim \mathcal{N}(\nu, \sigma^2)$.*

Theorem A.4 (Bernstein's inequality [127, Theorem 2.8.1])**.** *Let Y_i, $i \in [m]$, be independent, mean zero, sub-exponential random variables. Then, for every $t \ge 0$, it is*

$$\mathbb{P}\left(\left|\sum_{i=1}^{m} Y_i\right| \ge t\right) \le 2\exp\left(-C\min\left\{\frac{t^2}{\sum_{i=1}^{m}\|Y_i\|_{\psi_1}^2}, \frac{t}{\max_{i\in[m]}\|Y_i\|_{\psi_1}}\right\}\right),$$

where $C > 0$ is an absolute constant.

Definition A.4. *A random variable Y is called sub-gaussian if there exists $K > 0$ such that*

$$\mathbb{E}\left[\exp\left(\frac{Y^2}{K^2}\right)\right] \le 2.$$

The sub-gaussian norm is defined as

$$\|Y\|_{\psi_2} := \inf\left\{t > 0 : \mathbb{E}\left[\exp\left(\frac{Y^2}{t^2}\right)\right] \le 2\right\}.$$

A random vector $a \in \mathbb{R}^d$ is called sub-gaussian if all $\langle a, z\rangle$ with $z \in \mathbb{R}^d$ are sub-gaussian.

Definition A.5. *A random vector $a \in \mathbb{R}^d$ is called isotropic if $\mathbb{E}(aa^{\mathrm{T}}) = I_d$.*

Theorem A.5 ([128, Theorem 5.39])**.** *Let A be an $m \times d$ random matrix with independent sub-gaussian isotropic random vectors in $\mathbb{R}^d$ as rows. Then, for every $t \ge 0$, it is*

$$\sqrt{m} - C\sqrt{d} - t \le \sigma_{\min}(A) \le \sigma_{\max}(A) \le \sqrt{m} + C\sqrt{d} + t,$$

with probability at least $1 - 2\exp(-ct^2)$, where $C, c > 0$ are absolute constants.

Theorem A.6 (Hoeffding's inequality [127, Theorem 2.6.3])**.** *Let Y_i, $i \in [m]$, be independent, mean zero, sub-gaussian random variables, and $z \in \mathbb{R}^m$. Then, for every $t \geq 0$, it holds*

$$\mathbb{P}\left(\left|\sum_{i=1}^{m} z_i Y_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\max_{i\in[m]} \|Y_i\|_{\psi_2}^2 \|z\|_2^2}\right).$$

Theorem A.7 ([47, Proposition 7.11])**.** *If a random variable Y satisfies*

$$(\mathbb{E}|Y|^p)^{\frac{1}{p}} \leq ap^{\frac{1}{b}} \quad \text{for all } p \in [p_0, p_1],$$

for some constants $a, b, p_1 > p_0 > 0$, then

$$\mathbb{P}\left(|Y| \geq e^{\frac{1}{b}} au\right) \leq \exp\left(-\frac{u^b}{b}\right) \quad \text{for all } u \in [p_0^{\frac{1}{b}}, p_1^{\frac{1}{b}}].$$

Lemma A.8 ([35, Lemma C.2])**.** *Let Y be a random variable and Y' an independent copy of Y. If $\mathbb{E}\left(Y - \mathbb{E}\left[Y\right]\right)^2 < \infty$, it is*

$$\mathbb{P}\left(|Y - \mathbb{E}\left[Y\right]| \geq t\right) \leq 2\mathbb{P}\left(|Y - Y'| \geq t - \left(\mathbb{E}\left(Y - \mathbb{E}\left[Y\right]\right)^2\right)^{\frac{1}{2}}\right).$$

Furthermore, we need the following properties for standard Gaussian random vectors $a_i \in \mathbb{R}^d$, $i \in [m]$, and a deterministic vector $x \in \mathbb{R}^d$.

It is

$$\langle a_i, x\rangle \sim \mathcal{N}(0, \|x\|_2^2),$$

since the sum of Gaussian random variables is Gaussian,

$$\mathbb{E}\left[\langle a_i, x\rangle\right] = \sum_{j=1}^{d} \mathbb{E}\left[a_{ij}\right] x_j = 0, \quad \text{and} \quad \mathbb{V}\left[\langle a_i, x\rangle\right] = \sum_{j=1}^{d} \mathbb{V}\left[a_{ij}\right] x_j^2 = \|x\|_2^2.$$

Moreover,

$$\mathbb{E}\left[\|a_i\|_2^2\right] = \sum_{j=1}^{d} \mathbb{E}\left[a_{ij}^2\right] = d$$

and

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} |\langle a_i, x\rangle|^2\right] = x^{\mathrm{T}}\mathbb{E}\left[a_1 a_1^{\mathrm{T}}\right] x = x^{\mathrm{T}} I_d x = \|x\|_2^2, \tag{A.1}$$

due to $\mathbb{E}\left[a_{1j}^2\right] = 1$ for all $j \in [d]$, and $\mathbb{E}\left[a_{1j_1} a_{1j_2}\right] = 0$ for all $j_1, j_2 \in [d]$ with $j_1 \neq j_2$.

For the proofs presented in Chapter 5 and Section 7.1, we require also the following auxiliary lemmas. The results are based on joint work with Sjoerd Dirksen, Felix Krahmer and Palina Salanevich; see also [37].

Lemma A.9. *For independent random vectors* $a_i \sim \mathcal{N}(0, I_d),\ i \in [m]$, *it holds*

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \left|\langle a_i, x\rangle\right|^2 a_i a_i^{\mathrm{T}}\right] = 2xx^{\mathrm{T}} + \|x\|_2^2 \cdot I_d.$$

Proof. It is

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} \left|\langle a_i, x\rangle\right|^2 a_i a_i^{\mathrm{T}}\right] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[|\langle a_i, x\rangle|^2 a_i a_i^{\mathrm{T}}\right] = \mathbb{E}\left[\left|\langle a_1, x\rangle\right|^2 a_1 a_1^{\mathrm{T}}\right].$$

We determine the expectation of the random matrix entry-wise. That means, for all $k, \ell \in [d]$, we need to find

$$\left(\mathbb{E}\left[\left|\langle a_1, x\rangle\right|^2 a_1 a_1^{\mathrm{T}}\right]\right)_{k,\ell} = \begin{cases} \mathbb{E}\left[\left|\langle a_1, x\rangle\right|^2 \langle a_1, e_k\rangle^2\right], & k = \ell, \\ \mathbb{E}\left[\left|\langle a_1, x\rangle\right|^2 \langle a_1, e_k\rangle \langle a_1, e_\ell\rangle\right], & k \neq \ell, \end{cases}$$

where e_k denotes the kth standard unit vector.

Inspired by [99], we use a trick to rewrite the random variables $\langle a_1, e_k\rangle, \langle a_1, e_\ell\rangle$ and $\langle a_1, x\rangle$. For vectors $u, v \in \mathbb{R}^d$ with $\|u\|_2 = \|v\|_2 = 1$, the inner products $\langle a_1, u\rangle$ and $\langle a_1, v\rangle$ with standard Gaussian random vector a_1 are both standard Gaussian random variables with covariance

$$\begin{aligned} \operatorname{Cov}\left(\langle a_1, u\rangle, \langle a_1, v\rangle\right) &= \mathbb{E}\left(\langle a_1, u\rangle \cdot \langle a_1, v\rangle\right) - \mathbb{E}\langle a_1, u\rangle \cdot \mathbb{E}\langle a_1, v\rangle \\ &= u^{\mathrm{T}}\mathbb{E}\left(a_1 a_1^{\mathrm{T}}\right) v \\ &= \langle u, v\rangle. \end{aligned}$$

Let $g, h \sim \mathcal{N}(0,1)$ be independent. Setting $\langle a_1, u\rangle = g$, we can rewrite the dependent random variable $\langle a_1, v\rangle$ as $\langle a_1, v\rangle = \gamma_1 g + \gamma_2 h$. The values for γ_1, γ_2 can be derived using that $\mathbb{V}(\langle a_1, v\rangle) = 1$ and $\operatorname{Cov}\left(\langle a_1, u\rangle, \langle a_1, v\rangle\right) = \langle u, v\rangle$. We have

$$\begin{aligned} \operatorname{Cov}\left(g, \gamma_1 g + \gamma_2 h\right) &= \mathbb{E}\left(\gamma_1 g^2 + \gamma_2 g h\right) - \mathbb{E}(g) \cdot \mathbb{E}(\gamma_1 g + \gamma_2 h) \\ &= \gamma_1 \mathbb{E}(g^2) = \gamma_1. \end{aligned}$$

That is, $\gamma_1 = \langle u, v\rangle$. Furthermore, it holds

$$\mathbb{V}(\gamma_1 g + \gamma_2 h) = \gamma_1^2 + \gamma_2^2,$$

so that $\gamma_2 = \sqrt{1 - \gamma_1^2} = \sqrt{1 - \langle u, v\rangle^2}$.

Consequently, we denote in the following $g := \left\langle a_1, \frac{x}{\|x\|_2} \right\rangle$, and can rewrite

$$\langle a_1, e_k \rangle = \left\langle \frac{x}{\|x\|_2}, e_k \right\rangle g + \sqrt{1 - \left\langle \frac{x}{\|x\|_2}, e_k \right\rangle^2} \, h_1$$

and

$$\langle a_1, e_\ell \rangle = \left\langle \frac{x}{\|x\|_2}, e_\ell \right\rangle g + \sqrt{1 - \left\langle \frac{x}{\|x\|_2}, e_\ell \right\rangle^2} \, h_2,$$

with $g, h_1, h_2 \sim \mathcal{N}(0,1)$, where g and h_1 are independent and g and h_2 are independent as well. Note that $\mathbb{E}\left(a_{1k} a_{1\ell}\right) = 0$ if $k \neq \ell$. With this, we find that

$$\mathbb{E}(h_1 h_2) = -\frac{\left\langle \frac{x}{\|x\|_2}, e_k \right\rangle \left\langle \frac{x}{\|x\|_2}, e_\ell \right\rangle}{\sqrt{1 - \left\langle \frac{x}{\|x\|_2}, e_k \right\rangle^2} \sqrt{1 - \left\langle \frac{x}{\|x\|_2}, e_\ell \right\rangle^2}} \, . \tag{A.2}$$

Rewriting the random variables accordingly, using $\mathbb{E}[g^4] = 3$ and the independence of g and h_1, we can compute

$$\begin{aligned}
\mathbb{E}\left[|\langle a_1, x \rangle|^2 \langle a_1, e_k \rangle^2 \right] &= \mathbb{E}\left[\|x\|_2^2 \, g^2 \left(\left\langle \frac{x}{\|x\|_2}, e_k \right\rangle g + \sqrt{1 - \left\langle \frac{x}{\|x\|_2}, e_k \right\rangle^2} \, h_1 \right)^2 \right] \\
&= \mathbb{E}\left[\langle x, e_k \rangle^2 g^4 \right] + \mathbb{E}\left[\left(\|x\|_2^2 - \langle x, e_k \rangle^2 \right) g^2 h_1^2 \right] \\
&= 3 \langle x, e_k \rangle^2 + \left(\|x\|_2^2 - \langle x, e_k \rangle^2 \right) \\
&= 2 x_k^2 + \|x\|_2^2 \, ,
\end{aligned}$$

and, using also (A.2), we obtain

$$
\begin{aligned}
&\mathbb{E}\left[|\langle a_1, x\rangle|^2 \langle a_1, e_k\rangle \langle a_1, e_\ell\rangle\right] \\
&= \mathbb{E}\left[\|x\|_2^2\, g^2 \left(\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle g + \sqrt{1-\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle^2}\, h_1\right)\cdot\right. \\
&\qquad \left.\cdot\left(\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle g + \sqrt{1-\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle^2}\, h_2\right)\right] \\
&= \mathbb{E}\left[\langle x, e_k\rangle\langle x, e_\ell\rangle g^4\right] + \mathbb{E}\left[\|x\|_2^2\sqrt{1-\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle^2}\sqrt{1-\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle^2}\, g^2 h_1 h_2\right] \\
&= 3\langle x, e_k\rangle\langle x, e_\ell\rangle - \langle x, e_k\rangle\langle x, e_\ell\rangle \\
&= 2x_k x_\ell.
\end{aligned}
$$

Combining all entries yields the statement. □

Lemma A.10. *If $y_i \sim$ Poisson$\left(|\langle a_i, x\rangle|^2\right)$, $i \in [m]$, with independent random vectors $a_i \sim \mathcal{N}(0, I_d)$, $i \in [m]$, it holds*

$$
\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m y_i a_i a_i^{\mathrm{T}}\right] = 2xx^{\mathrm{T}} + \|x\|_2^2 \cdot I_d.
$$

Proof. Conditioning on the vectors a_i, $i \in [m]$, the expectation of $S := \frac{1}{m}\sum_{i=1}^m y_i a_i a_i^{\mathrm{T}}$ is

$$
\mathbb{E}_y\left[S \mid A\right] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{y_i}\left[y_i \mid A\right] a_i a_i^{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^m |\langle a_i, x\rangle|^2 a_i a_i^{\mathrm{T}},
$$

where A denotes the matrix with rows a_i^{T}. Taking the expectation of this yields

$$
\mathbb{E}\left[S\right] = \mathbb{E}_A\left[\mathbb{E}_y\left[S \mid A\right]\right] = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_A\left[|\langle a_i, x\rangle|^2 a_i a_i^{\mathrm{T}}\right] = \mathbb{E}_{a_1}\left[|\langle a_1, x\rangle|^2 a_1 a_1^{\mathrm{T}}\right].
$$

Now, continue as in the proof of Lemma A.9 to obtain the result. □

Lemma A.11. *If $y_i \sim$ Bernoulli$\left(1-\exp\left(-|\langle a_i, x\rangle|^2\right)\right)$, $i \in [m]$, with independent random vectors $a_i \sim \mathcal{N}(0, I_d)$, $i \in [m]$, it holds*

$$
\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m y_i a_i a_i^{\mathrm{T}}\right] = \frac{2}{(2\|x\|_2^2+1)^{\frac{3}{2}}} \cdot xx^{\mathrm{T}} + \left(1 - \frac{1}{(2\|x\|_2^2+1)^{\frac{1}{2}}}\right) \cdot I_d.
$$

Proof. We start again by computing $\mathbb{E}[S]$ for $S := \frac{1}{m}\sum_{i=1}^{m} y_i a_i a_i^{\mathrm{T}}$. We condition on the vectors $a_i,\ i \in [m]$, and obtain

$$\mathbb{E}_y[S \mid A] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{y_i}[y_i \mid A]\, a_i a_i^{\mathrm{T}} = \frac{1}{m}\sum_{i=1}^{m} \left(1 - \exp(-|\langle a_i, x\rangle|^2)\right) a_i a_i^{\mathrm{T}},$$

where A denotes the matrix with rows a_i^{T}. The expectation of this equals

$$\begin{aligned}\mathbb{E}[S] &= \mathbb{E}_A\left[\mathbb{E}_y[S \mid A]\right] \\ &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_A\left[\left(1 - \exp(-|\langle a_i, x\rangle|^2)\right) a_i a_i^{\mathrm{T}}\right] \\ &= \mathbb{E}_{a_1}\left[\left(1 - \exp(-|\langle a_1, x\rangle|^2)\right) a_1 a_1^{\mathrm{T}}\right].\end{aligned}$$

To obtain the expectation of the random matrix S, we study its entries

$$\begin{aligned}&\left(\mathbb{E}\left[\left(1 - \exp(-|\langle a_1, x\rangle|^2)\right) a_1 a_1^{\mathrm{T}}\right]\right)_{k,\ell} \\ &= \begin{cases}\mathbb{E}\left[\left(1 - \exp(-|\langle a_1, x\rangle|^2)\right) \langle a_1, e_k\rangle^2\right], & k = \ell, \\ \mathbb{E}\left[\left(1 - \exp(-|\langle a_1, x\rangle|^2)\right) \langle a_1, e_k\rangle \langle a_1, e_\ell\rangle\right], & k \neq \ell,\end{cases}\end{aligned}$$

for all $k, \ell \in [d]$, where e_k denotes the kth standard unit vector.

We use the same trick as in the proof of Lemma A.9 to rewrite the random variables $\langle a_1, e_k\rangle, \langle a_1, e_\ell\rangle$ and $\langle a_1, x\rangle$. However, for the Bernoulli model, we need to further determine

$$\begin{aligned}\mathbb{E}\left[\exp(-\|x\|_2^2\, g^2)\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp(-\|x\|_2^2\, t^2)\exp\left(-\frac{t^2}{2}\right) dt \\ &= \frac{1}{\sqrt{2\pi}}\int_{\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\sqrt{2\|x\|_2^2 + 1}\cdot t\right)^2\right) dt \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) du \cdot \frac{1}{\sqrt{2\|x\|_2^2 + 1}} \\ &= \frac{1}{(2\|x\|_2^2 + 1)^{\frac{1}{2}}}\end{aligned}$$

and

$$
\begin{aligned}
&\mathbb{E}\left[g^2 \exp(-\|x\|_2^2 g^2)\right] \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 \exp(-\|x\|_2^2 t^2) \exp\left(-\frac{t^2}{2}\right) dt \\
&= \frac{1}{\sqrt{2\pi}} \left(\left[-\frac{1}{2\|x\|_2^2 + 1} \exp\left(-\left(\|x\|_2^2 + \frac{1}{2} \right) t^2 \right) \cdot t \right]_{-\infty}^{\infty} \right. \\
&\quad \left. + \frac{1}{2\|x\|_2^2 + 1} \cdot \int_{\infty}^{\infty} \exp\left(-\frac{2\left(\|x\|_2^2 + \frac{1}{2}\right) t^2}{2} \right) dt \right) \\
&= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2\|x\|_2^2 + 1} \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) du \cdot \frac{1}{\sqrt{2\|x\|_2^2 + 1}} \\
&= \frac{1}{(2\|x\|_2^2 + 1)^{\frac{3}{2}}}.
\end{aligned}
$$

Now, we can compute

$$
\begin{aligned}
&\mathbb{E}\left[\left(1 - \exp(-|\langle a_1, x\rangle|^2)\right) \langle a_1, e_k\rangle^2\right] \\
&= \mathbb{E}\left[\left(1 - \exp(-\|x\|_2^2 g^2)\right) \left(\left\langle \frac{x}{\|x\|_2}, e_k \right\rangle g + \sqrt{1 - \left\langle \frac{x}{\|x\|_2}, e_k \right\rangle^2} h_1 \right)^2 \right] \\
&= \mathbb{E}\left[\left\langle \frac{x}{\|x\|_2}, e_k \right\rangle^2 \left(1 - \exp(-\|x\|_2^2 g^2)\right) g^2 \right] \\
&\quad + \mathbb{E}\left[\left(1 - \left\langle \frac{x}{\|x\|_2}, e_k \right\rangle^2 \right) \left(1 - \exp(-\|x\|_2^2 g^2)\right) h_1^2 \right] \\
&= \left(1 - \frac{1}{\left(2\|x\|_2^2 + 1\right)^{\frac{3}{2}}} \right) \left\langle \frac{x}{\|x\|_2}, e_k \right\rangle^2 + \left(1 - \frac{1}{\left(2\|x\|_2^2 + 1\right)^{\frac{1}{2}}} \right) \left(1 - \left\langle \frac{x}{\|x\|_2}, e_k \right\rangle^2 \right) \\
&= 2x_k^2 \cdot \frac{1}{\left(2\|x\|_2^2 + 1\right)^{\frac{3}{2}}} + \left(1 - \frac{1}{\left(2\|x\|_2^2 + 1\right)^{\frac{1}{2}}} \right),
\end{aligned}
$$

and, using (A.2),

$$
\begin{aligned}
&\mathbb{E}\left[\left(1-\exp(-\left|\langle a_1, x\rangle\right|^2)\right)\langle a_1, e_k\rangle\langle a_1, e_\ell\rangle\right] \\
&= \mathbb{E}\left[\left(1-\exp(-\|x\|_2^2\, g^2)\right)\left(\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle g + \sqrt{1-\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle^2}\, h_1\right)\cdot\right. \\
&\qquad\left.\cdot\left(\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle g + \sqrt{1-\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle^2}\, h_2\right)\right] \\
&= \mathbb{E}\left[\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle\left(1-\exp(-\|x\|_2^2\, g^2)\right) g^2\right] \\
&\quad + \mathbb{E}\left[\sqrt{1-\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle^2}\sqrt{1-\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle^2}\left(1-\exp(-\|x\|_2^2\, g^2)\right) h_1 h_2\right] \\
&= \left(1-\frac{1}{\left(2\|x\|_2^2+1\right)^{\frac{3}{2}}}\right)\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle \\
&\quad - \left(1-\frac{1}{\left(2\|x\|_2^2+1\right)^{\frac{1}{2}}}\right)\left\langle \frac{x}{\|x\|_2}, e_k\right\rangle\left\langle \frac{x}{\|x\|_2}, e_\ell\right\rangle \\
&= 2x_k x_\ell\cdot\frac{1}{\left(2\|x\|_2^2+1\right)^{\frac{3}{2}}}.
\end{aligned}
$$

The respective last equalities use

$$
\begin{aligned}
\frac{1}{\left(2\|x\|_2^2+1\right)^{\frac{1}{2}}}-\frac{1}{\left(2\|x\|_2^2+1\right)^{\frac{3}{2}}} &= \frac{1}{\left(2\|x\|_2^2+1\right)^{\frac{1}{2}}}\cdot\left(1-\frac{1}{2\|x\|_2^2+1}\right) \\
&= \frac{1}{\left(2\|x\|_2^2+1\right)^{\frac{1}{2}}}\cdot\frac{2\|x\|_2^2}{2\|x\|_2^2+1} \\
&= \frac{2\|x\|_2^2}{\left(2\|x\|_2^2+1\right)^{\frac{3}{2}}}.
\end{aligned}
$$

Combining the entries, we obtain the statement. □

Furthermore, we need the following concentration results.

Lemma A.12. *Let $m \geq d$. If $y_i \sim$ Poisson $\left(|\langle a_i, x\rangle|^2\right)$, $i \in [m]$, with independent random vectors $a_i \sim \mathcal{N}(0, I_d)$, $i \in [m]$, there exist absolute constants $C, c > 0$, such that, with probability at least*

$$1 - 24m\exp\left(-cd\right) - \frac{6}{d} - \frac{18}{m},$$

it holds that

$$\|S - \mathbb{E}\left[S\right]\| \leq C\lceil\|x\|_2^2\rceil\sqrt{\log(d)}\log(m)\sqrt{\frac{d}{m}},$$

where $S := \frac{1}{m}\sum_{i=1}^m y_i a_i a_i^{\mathrm{T}}$.

Proof. As the Poisson random variables are unbounded, we start by deriving a concentration bound for $\max_{i\in[m]} y_i$. For this, we use that, for any $\xi \geq 0$ and $t \in \mathbb{N}\setminus\{0\}$,

$$\begin{aligned}
\sum_{k=0}^{t-1}\frac{\xi^k}{k!} &= \exp(\xi) - \sum_{k=t}^{\infty}\frac{\xi^k}{k!} = \exp(\xi) - \xi^t\sum_{k=0}^{\infty}\frac{\xi^k}{(k+t)!} \\
&= \exp(\xi) - \xi^t\sum_{k=0}^{\infty}\frac{\xi^k}{k!}\prod_{s=k+1}^{k+t}\frac{1}{s} \geq \exp(\xi) - \frac{\xi^t}{t!}\sum_{k=0}^{\infty}\frac{\xi^k}{k!} \\
&= \exp(\xi)\left(1 - \frac{\xi^t}{t!}\right),
\end{aligned}$$

and hence, for all $i \in [m]$ and some $t \in \mathbb{N}\setminus\{0\}$, we have, conditional on the a_i, $i \in [m]$,

$$\begin{aligned}
\mathbb{P}_y\left(y_i < t\right) &= \sum_{k=0}^{t-1}\frac{\exp\left(-|\langle a_i, x\rangle|^2\right)\left(|\langle a_i, x\rangle|^2\right)^k}{k!} \\
&= \exp\left(-|\langle a_i, x\rangle|^2\right)\sum_{k=0}^{t-1}\frac{\left(|\langle a_i, x\rangle|^2\right)^k}{k!} \\
&\geq 1 - \frac{\left(|\langle a_i, x\rangle|^2\right)^t}{t!}.
\end{aligned}$$

Thus, conditional on the a_i, $i \in [m]$, we get

$$\begin{aligned}
\mathbb{P}_y\left(\max_{i\in[m]} y_i \geq t\right) &= \mathbb{P}_y\left(\exists i \in [m] : y_i \geq t\right) \leq \sum_{i=1}^m \mathbb{P}_y\left(y_i \geq t\right) \leq \sum_{i=1}^m\frac{\left(|\langle a_i, x\rangle|^2\right)^t}{t!} \\
&\leq m\cdot\frac{\left(\max_{i\in[m]}|\langle a_i, x\rangle|^2\right)^t}{t!}.
\end{aligned}$$

Let A be the matrix with rows a_i^{T}. With Hoeffding's inequality (Theorem A.6) we obtain that, for all $t > 0$, it is

$$\begin{aligned}
\mathbb{P}_A\left(\max_{i\in[m]} |\langle a_i, x\rangle|^2 \geq t\right) &= \mathbb{P}_A\left(\exists i \in [m] : |\langle a_i, x\rangle|^2 \geq t\right) \leq \sum_{i=1}^{m} \mathbb{P}_A\left(|\langle a_i, x\rangle|^2 \geq t\right) \\
&= \sum_{i=1}^{m} \mathbb{P}_A\left(\left|\sum_{j=1}^{d} a_{ij}x_j\right| \geq \sqrt{t}\right) \\
&\leq \sum_{i=1}^{m} 2\exp\left(-\frac{ct}{\left(\max_{j\in[d]} \|a_{ij}\|_{\psi_2}\right)^2 \|x\|_2^2}\right) \\
&= 2m\exp\left(-\frac{ct}{\|x\|_2^2}\right)
\end{aligned}$$

for an absolute constant $c > 0$. Hence, we get that there exists a constant $\hat{c} > 0$ with

$$\mathbb{P}_A\left(\max_{i\in[m]} |\langle a_i, x\rangle|^2 \geq 2\hat{c}\,\|x\|_2^2 \log(m)\right) \leq 2m\exp\left(-2\log(m)\right) = \frac{2}{m}.$$

We continue with the bound on y_i, and find that for all $\tau > 0$ it is

$$\begin{aligned}
&\mathbb{P}_y\left(\max_{i\in[m]} y_i \geq \lceil \tau\log(m)\rceil \;\middle|\; \max_{i\in[m]} |\langle a_i, x\rangle|^2 \leq 2\hat{c}\,\|x\|_2^2 \log(m)\right) \\
&\leq m \cdot \frac{\left(2\hat{c}\,\|x\|_2^2 \log(m)\right)^{\lceil \tau\log(m)\rceil}}{\lceil \tau\log(m)\rceil!} \\
&= m\exp\left(\lceil \tau\log(m)\rceil \log\left(2\hat{c}\,\|x\|_2^2\log(m)\right) - \log\left(\lceil \tau\log(m)\rceil!\right)\right).
\end{aligned}$$

Using Stirling's approximation, that is, $\log(\zeta!) = \zeta\log(\zeta) - \zeta + \mathcal{O}(\log(\zeta))$, we get

$$\begin{aligned}
&\lceil \tau\log(m)\rceil \log\left(2\hat{c}\,\|x\|_2^2\log(m)\right) - \log\left(\lceil \tau\log(m)\rceil!\right) \\
&= \lceil \tau\log(m)\rceil \left(\log\left(\frac{2\hat{c}\,\|x\|_2^2\log(m)}{\lceil \tau\log(m)\rceil}\right) + 1\right) - \mathcal{O}\left(\log(\lceil \tau\log(m)\rceil\right).
\end{aligned}$$

Choosing $\tau = 2\hat{c}e^2\,\|x\|_2^2$, we obtain

$$\log\left(\frac{2\hat{c}\,\|x\|_2^2\, e^2 e^{-2}\log(m)}{\lceil 2\hat{c}e^2\,\|x\|_2^2\log(m)\rceil}\right) \leq -2,$$

so that we have

$$\begin{aligned}
&\mathbb{P}_y\left(\max_{i\in[m]} y_i \geq 2\hat{c}e^2\lceil\|x\|_2^2\rceil\log(m) \;\middle|\; \max_{i\in[m]}|\langle a_i, x\rangle|^2 \leq 2\hat{c}\,\|x\|_2^2\log(m)\right)\\
&\quad\leq m\exp\left(-2\hat{c}e^2\lceil\|x\|_2^2\rceil\log(m)\right)\\
&\quad\leq \frac{1}{m}
\end{aligned}$$

if $\hat{c}$ is sufficiently large.

Let us set

$$\gamma := \max\left\{2\hat{c}e^2, \frac{2}{\lceil\|x\|_2^2\rceil}\right\}$$

and define the following events of good concentration of the Poisson measurements y_i and the squared inner products $|\langle a_i, x\rangle|^2$:

$$E_1 := \left\{\max_{i\in[m]} y_i \leq \gamma\lceil\|x\|_2^2\rceil\log(m)\right\}$$

and

$$E_2 := \left\{\max_{i\in[m]}|\langle a_i, x\rangle|^2 \leq 2\hat{c}\,\|x\|_2^2\log(m)\right\}.$$

Let $E := E_1 \cap E_2$. From the above, we know that its complement E^c satisfies

$$\begin{aligned}
\mathbb{P}(E^c) &= \mathbb{P}(E_1^c \cup E_2^c) = \mathbb{P}((E_1^c \cap E_2) \cup E_2^c)\\
&\leq \mathbb{P}(E_1^c \cap E_2) + \mathbb{P}(E_2^c) \leq \mathbb{P}(E_1^c \mid E_2) + \mathbb{P}(E_2^c)\\
&\leq \frac{3}{m}.
\end{aligned}$$

Now, we can bound $\|S - \mathbb{E}[S]\|$. Using a symmetrization trick, see Lemma A.1, we obtain

$$\begin{aligned}
&\mathbb{P}(\|S - \mathbb{E}[S]\| \geq t)\\
&\quad= \mathbb{P}\left(\sup_{\|z\|_2=1}\left|\frac{1}{m}\sum_{i=1}^m y_i|\langle a_i, z\rangle|^2 - z^{\mathrm{T}}\mathbb{E}[S]\,z\right| \geq t\right)\\
&\quad\leq \mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\left(y_i|\langle a_i, z\rangle|^2 - y_i'|\langle a_i', z\rangle|^2\right)\right| \geq \frac{t}{2}\right)\\
&\qquad+ \sup_{\|z\|_2=1}\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m y_i|\langle a_i, z\rangle|^2 - z^{\mathrm{T}}\mathbb{E}[S]\,z\right| \geq \frac{t}{2}\right),
\end{aligned}$$

where $y_i', a_i',\ i \in [m]$, are independent copies of $y_i, a_i,\ i \in [m]$. Here, we use that the supremum can be reduced to a supremum over a countable dense subset due to continuity of the argument and hence Lemma A.1 can be applied.

With a Rademacher random vector $\varepsilon = (\varepsilon_i)_{i\in[m]}$, independent of $y_i', a_i', y_i, a_i,\ i \in [m]$, the first term can be rewritten and bounded by

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\left(y_i\left|\langle a_i,z\rangle\right|^2 - y_i'\left|\langle a_i',z\rangle\right|^2\right)\right| \geq \frac{t}{2}\right)\\
&\quad= \mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\varepsilon_i\left(y_i\left|\langle a_i,z\rangle\right|^2 - y_i'\left|\langle a_i',z\rangle\right|^2\right)\right| \geq \frac{t}{2}\right)\\
&\quad\leq 2\mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\varepsilon_i y_i\left|\langle a_i,z\rangle\right|^2\right| \geq \frac{t}{4}\right),
\end{aligned}$$

using the triangle inequality in the second step. Now, we consider these tail estimates on the event E and its complement and get

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\left(y_i\left|\langle a_i,z\rangle\right|^2 - y_i'\left|\langle a_i',z\rangle\right|^2\right)\right| \geq \frac{t}{2}\right)\\
&\qquad\leq 2\mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\varepsilon_i y_i\mathbb{1}_E\left|\langle a_i,z\rangle\right|^2\right| \geq \frac{t}{4}\right) + 2\mathbb{P}(E^c).
\end{aligned}$$

The first term equals

$$\mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\varepsilon_i y_i\mathbb{1}_E\left|\langle a_i,z\rangle\right|^2\right| \geq \frac{t}{4}\right) = \mathbb{P}\left(\left\|\frac{1}{m}\sum_{i=1}^m\varepsilon_i y_i\mathbb{1}_E a_i a_i^{\mathrm{T}}\right\| \geq \frac{t}{4}\right).$$

Let $p \geq \log(d)$. With the equivalence of the p-Schatten norm and the spectral norm for $p \geq \log(d)$, see Lemma A.2, and the non-commutative Khintchine inequality, Theorem A.3, we obtain

$$\begin{aligned}
\mathbb{E}_\varepsilon\left\|\frac{1}{m}\sum_{i=1}^m\varepsilon_i y_i\mathbb{1}_E a_i a_i^{\mathrm{T}}\right\|^p &\leq e^p\cdot\mathbb{E}_\varepsilon\left\|\frac{1}{m}\sum_{i=1}^m\varepsilon_i y_i\mathbb{1}_E a_i a_i^{\mathrm{T}}\right\|_{C_p^d}^p\\
&\leq \left(C\sqrt{p}\cdot\frac{1}{m}\left\|\sum_{i=1}^m y_i^2\mathbb{1}_E a_i a_i^{\mathrm{T}} a_i a_i^{\mathrm{T}}\right\|_{C_p^d}^{\frac{1}{2}}\right)^p\\
&\leq \left(C\sqrt{p}\cdot\frac{1}{m}\left\|\sum_{i=1}^m y_i^2\mathbb{1}_E\|a_i\|_2^2 a_i a_i^{\mathrm{T}}\right\|^{\frac{1}{2}}\right)^p,
\end{aligned}$$

with an absolute constant $C > 0$. Furthermore, it is

$$\left\| \sum_{i=1}^{m} y_i^2 \mathbb{1}_E \|a_i\|_2^2 a_i a_i^{\mathrm{T}} \right\|^{\frac{1}{2}} = \sup_{\|z\|_2=1} \left(\sum_{i=1}^{m} y_i^2 \mathbb{1}_E \|a_i\|_2^2 \langle a_i, z\rangle^2 \right)^{\frac{1}{2}}$$
$$\leq \gamma \lceil \|x\|_2^2 \rceil \log(m) \sup_{\|z\|_2=1} \left(\sum_{i=1}^{m} \|a_i\|_2^2 \langle a_i, z\rangle^2 \right)^{\frac{1}{2}},$$

using the restriction onto the event E.

We use $\mathbb{E}_A \|a_i\|_2^2 = d$ and expand

$$\|a_i\|_2^2 = \left| \|a_i\|_2^2 - \mathbb{E}_A \|a_i\|_2^2 + d \right|.$$

Since the random vectors a_i are standard Gaussian, and, hence, their squared entries $a_{ik}^2,\ k \in [d]$, are independent sub-exponential random variables with norm $\|a_{ik}^2\|_{\psi_1} = 1$, we can use Bernstein's inequality (Theorem A.4) to show that

$$\mathbb{P}_A \left(| \|a_i\|_2^2 - \mathbb{E}_A \|a_i\|_2^2 | \geq d \right) = \mathbb{P}_A \left(\left| \sum_{k=1}^{d} a_{ik}^2 - \mathbb{E}_A a_{ik}^2 \right| \geq d \right)$$
$$\leq 2 \exp \left(-c \min \left\{ \frac{d^2}{\sum_{k=1}^{d} \|a_{ik}^2\|_{\psi_1}^2}, \frac{d}{\max_{k\in[d]} \|a_{ik}^2\|_{\psi_1}} \right\} \right)$$
$$= 2 \exp(-cd),$$

with an absolute constant $c > 0$. Consequently, with probability at least $1 - 2\exp(-cd)$ with respect to $a_i,\ i \in [m]$,

$$\|a_i\|_2^2 = \left| \|a_i\|_2^2 - \mathbb{E}_A \|a_i\|_2^2 + d \right| \leq 2d,$$

and $\mathbb{P}\left(\max_{i\in[m]} \|a_i\|_2^2 \leq 2d \right) \geq 1 - 2m \exp(-cd)$. It remains to bound $\left\| \sum_{i=1}^{m} a_i a_i^{\mathrm{T}} \right\|$. We use Theorem A.5 to show that, with probability at least $1 - 2\exp(-\tilde{c}d)$ with respect to $a_i,\ i \in [m]$,

$$\left\| \sum_{i=1}^{m} a_i a_i^{\mathrm{T}} \right\|^{\frac{1}{2}} = \sup_{\|z\|_2=1} \left(\sum_{i=1}^{m} \langle a_i, z\rangle^2 \right)^{\frac{1}{2}} = \sup_{\|z\|_2=1} \|Az\|_2 = \sigma_{\max}(A) \leq \left(\sqrt{m} + \tilde{C}\sqrt{d} \right),$$

where $\tilde{C}, \tilde{c} > 0$ are absolute constants.

Using that $m \geq d$, we summarize that there exist absolute constants $C, c > 0$ so that, for all $p \geq \log(d)$,

$$\left(\mathbb{E}_\varepsilon \left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i \mathbb{1}_E a_i a_i^{\mathrm{T}}\right\|^p\right)^{\frac{1}{p}} \leq C\gamma\lceil\|x\|_2^2\rceil \log(m)\sqrt{p}\sqrt{d}\cdot\frac{1}{m}\cdot\sqrt{m}$$

with probability at least $1 - 4m\exp(-cd)$ with respect to a_i, $i \in [m]$. Consequently, Theorem A.7 yields that

$$\begin{aligned}\mathbb{P}&\left(\left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i \mathbb{1}_E a_i a_i^{\mathrm{T}}\right\| \geq \frac{t}{4}\right)\\ &\leq \exp\left(-\frac{t^2}{2\cdot 16e\cdot C^2\gamma^2\lceil\|x\|_2^2\rceil^2\log^2(m)\frac{d}{m}}\right) + 4m\exp(-cd)\end{aligned}$$

for all $t \geq 4\sqrt{e}C\gamma\lceil\|x\|_2^2\rceil\log(m)\sqrt{\frac{d}{m}}\cdot\sqrt{\log(d)}$, with absolute constants $C, c > 0$. Choosing $t = \sqrt{2}\cdot 4\sqrt{e}C\gamma\lceil\|x\|_2^2\rceil\log(m)\sqrt{\frac{d}{m}}\cdot\sqrt{\log(d)}$ and using that there exists a constant $\hat{C} > 0$ with

$$\gamma\lceil\|x\|_2^2\rceil \leq \left(2\hat{c}e^2 + \frac{2}{\lceil\|x\|_2^2\rceil}\right)\lceil\|x\|_2^2\rceil = 2\hat{c}e^2\lceil\|x\|_2^2\rceil + 2 \leq \hat{C}\lceil\|x\|_2^2\rceil,$$

we found that there exist absolute constants $\tilde{C}, c > 0$ so that

$$\begin{aligned}\mathbb{P}&\left(\|S - \mathbb{E}[S]\| \geq \tilde{C}\lceil\|x\|_2^2\rceil\sqrt{\log(d)}\log(m)\sqrt{\tfrac{d}{m}}\right)\\ &\leq \frac{2}{d} + 8m\exp(-cd) + \frac{6}{m}\\ &\quad + \sup_{\|z\|_2=1}\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m y_i|\langle a_i, z\rangle|^2 - z^{\mathrm{T}}\mathbb{E}[S]z\right| \geq \frac{1}{2}\cdot\tilde{C}\lceil\|x\|_2^2\rceil\sqrt{\log(d)}\log(m)\sqrt{\tfrac{d}{m}}\right).\end{aligned}$$

The remaining step is to bound the last term on the right-hand side. Using Lemma A.8, we get that, for any $t > 0$,

$$\begin{aligned}\mathbb{P}&\left(\left|\frac{1}{m}\sum_{i=1}^m y_i|\langle a_i, z\rangle|^2 - z^{\mathrm{T}}\mathbb{E}[S]z\right| \geq t\right)\\ &\leq 2\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m y_i|\langle a_i, z\rangle|^2 - y_i'|\langle a_i', z\rangle|^2\right| \geq t - V\right),\end{aligned}$$

where $y_i', a_i',\ i \in [m]$, are independent copies of $y_i, a_i,\ i \in [m]$, and

$$V = \left(\mathbb{E} \left[\frac{1}{m} \sum_{i=1}^{m} y_i \left| \langle a_i, z \rangle \right|^2 - \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} \left(y_i \left| \langle a_i, z \rangle \right|^2 \right) \right]^2 \right)^{\frac{1}{2}}.$$

Since for any $z \in \mathbb{R}^d$ with $\|z\|_2 = 1$ it is

$$\begin{aligned} &\mathbb{P} \left(\left| \frac{1}{m} \sum_{i=1}^{m} y_i \left| \langle a_i, z \rangle \right|^2 - y_i' \left| \langle a_i', z \rangle \right|^2 \right| \geq t - V \right) \\ &\leq \mathbb{P} \left(\sup_{\|v\|_2 = 1} \left| \frac{1}{m} \sum_{i=1}^{m} y_i \left| \langle a_i, v \rangle \right|^2 - y_i' \left| \langle a_i', v \rangle \right|^2 \right| \geq t - V \right), \end{aligned}$$

we can proceed as before to find that

$$\begin{aligned} &\sup_{\|z\|_2 = 1} \mathbb{P} \left(\left| \frac{1}{m} \sum_{i=1}^{m} y_i \left| \langle a_i, z \rangle \right|^2 - z^{\mathrm{T}} \mathbb{E} \left[S \right] z \right| \geq \frac{1}{2} C \lceil \|x\|_2^2 \rceil \sqrt{\log(d)} \log(m) \sqrt{\tfrac{d}{m}} \right) \\ &\leq 4 \mathbb{P}_{\varepsilon} \left(\left\| \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i y_i \mathbb{1}_E a_i a_i^{\mathrm{T}} \right\| \geq \frac{1}{2} \left(C \lceil \|x\|_2^2 \rceil \sqrt{\log(d)} \sqrt{\frac{d}{m}} - V \right) \right) + \frac{12}{m}. \end{aligned}$$

To analyze V, we observe that

$$\begin{aligned} &\mathbb{E} \left[\frac{1}{m} \sum_{i=1}^{m} y_i \left| \langle a_i, z \rangle \right|^2 - \frac{1}{m} \sum_{i=1}^{m} \mathbb{E} \left(y_i \left| \langle a_i, z \rangle \right|^2 \right) \right]^2 \\ &= \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^{m} y_i \left| \langle a_i, z \rangle \right|^2 \right]^2 - \left(\mathbb{E} \left[y_1 \left| \langle a_1, z \rangle \right|^2 \right] \right)^2 \\ &= \mathbb{E} \left[\frac{1}{m^2} \sum_{i=1}^{m} y_i^2 |\langle a_i, z \rangle|^4 \right] + \mathbb{E} \left[\frac{1}{m^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} y_i y_j |\langle a_i, z \rangle|^2 |\langle a_j, z \rangle|^2 \right] - \left(\mathbb{E} \left[y_1 \left| \langle a_1, z \rangle \right|^2 \right] \right)^2 \\ &= \frac{1}{m} \mathbb{E} \left(y_1^2 | \langle a_1, z \rangle |^4 \right) + \frac{1}{m^2} \cdot (m^2 - m) \left(\mathbb{E} \left[y_1 \left| \langle a_1, z \rangle \right|^2 \right] \right)^2 - \left(\mathbb{E} \left[y_1 \left| \langle a_1, z \rangle \right|^2 \right] \right)^2 \\ &= \frac{1}{m} \left(\mathbb{E} \left(y_1^2 | \langle a_1, z \rangle |^4 \right) - \left(\mathbb{E} \left[y_1 \left| \langle a_1, z \rangle \right|^2 \right] \right)^2 \right), \end{aligned}$$

and use

$$\begin{aligned}\mathbb{E}\left(y_1^2 |\left\langle a_1, z\right\rangle|^4\right) &= \mathbb{E}_A\left[\mathbb{E}_y\left[y_1^2 |\left\langle a_1, z\right\rangle|^4 \;\middle|\; A\right]\right] \\ &= \mathbb{E}_A\left[\left(\left(\mathbb{E}_y\left[y_1 \mid A\right]\right)^2 + \mathbb{V}_y\left[y_1 \mid A\right]\right) |\left\langle a_1, z\right\rangle|^4\right] \\ &= \mathbb{E}_A\left[\left(|\left\langle a_1, x\right\rangle|^4 + |\left\langle a_1, x\right\rangle|^2\right) |\left\langle a_1, z\right\rangle|^4\right]\end{aligned}$$

as well as

$$\begin{aligned}\left(\mathbb{E}\left[y_1 |\langle a_1, z\rangle|^2\right]\right)^2 &= \left(\mathbb{E}_A\left[\mathbb{E}_y\left[y_1 |\langle a_1, z\rangle|^2 \;\middle|\; A\right]\right]\right)^2 \\ &= \left(\mathbb{E}_A\left[|\langle a_1, x\rangle|^2 |\langle a_1, z\rangle|^2\right]\right)^2.\end{aligned}$$

As in the proof of Lemma A.9, we rewrite $g = \left\langle a_1, \frac{x}{\|x\|_2}\right\rangle$ and

$$\langle a_1, z\rangle = \left\langle \frac{x}{\|x\|_2}, z\right\rangle g + \sqrt{1 - \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2} h,$$

for $g, h \sim \mathcal{N}(0,1)$ with g, h independent. Then,

$$\begin{aligned}&\mathbb{E}\left[\left(|\left\langle a_1, x\right\rangle|^4 + |\left\langle a_1, x\right\rangle|^2\right) |\left\langle a_1, z\right\rangle|^4\right] \\ &= \mathbb{E}\left[\left(\|x\|_2^4 g^4 + \|x\|_2^2 g^2\right)\left(\left\langle \frac{x}{\|x\|_2}, z\right\rangle g + \sqrt{1 - \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2} h\right)^4\right] \\ &= 105\|x\|_2^4 \left\langle \frac{x}{\|x\|_2}, z\right\rangle^4 + 90\|x\|_2^4 \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2 \left(1 - \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2\right) \\ &\quad + 9\|x\|_2^4 \left(1 - \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2\right)^2 + 15\|x\|_2^2 \left\langle \frac{x}{\|x\|_2}, z\right\rangle^4 \\ &\quad + 18\|x\|_2^2 \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2 \left(1 - \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2\right) + 3\|x\|_2^2 \left(1 - \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2\right)^2 \\ &= 24 \left\langle x, z\right\rangle^4 + 72\|x\|_2^2 \left\langle x, z\right\rangle^2 + 9\|x\|_2^4 + 12 \left\langle x, z\right\rangle^2 + 3\|x\|_2^2\end{aligned}$$

and

$$\mathbb{E}\left[|\langle a_1, x\rangle|^2 \, |\langle a_1, z\rangle|^2\right] = \mathbb{E}\left[\|x\|_2^2 g^2 \left(\left\langle \frac{x}{\|x\|_2}, z\right\rangle g + \sqrt{1 - \left\langle \frac{x}{\|x\|_2}, z\right\rangle^2} h\right)^2\right]$$
$$= 2\left\langle x, z\right\rangle^2 + \|x\|_2^2.$$

Hence, using $|\langle x, z\rangle| \leq \|x\|_2 \|z\|_2$, we find that for any $z \in \mathbb{R}^d$ with $\|z\|_2 = 1$ we can bound

$$\begin{aligned} V &= \frac{1}{\sqrt{m}} \left(20\left\langle x, z\right\rangle^4 + 68\|x\|_2^2 \left\langle x, z\right\rangle^2 + 8\|x\|_2^4 + 12\left\langle x, z\right\rangle^2 + 3\|x\|_2^2\right)^{\frac{1}{2}} \\ &\leq \frac{1}{\sqrt{m}} \left(96\|x\|_2^4 + 15\|x\|_2^2\right)^{\frac{1}{2}} \\ &\leq \frac{1}{\sqrt{m}} \left(10\|x\|_2^2 + 4\|x\|_2\right) \\ &\leq \frac{14}{\sqrt{m}} \lceil \|x\|_2^2 \rceil. \end{aligned}$$

We summarize that there exists a constant $\hat{C} > 0$ so that

$$\begin{aligned} &\mathbb{P}_\varepsilon\left(\left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i \mathbb{1}_E a_i a_i^{\mathrm{T}}\right\| \geq \frac{1}{2}\left(\hat{C}\lceil \|x\|_2^2\rceil \sqrt{\log(d)}\sqrt{\frac{d}{m}} - V\right)\right) \\ &\leq \mathbb{P}_\varepsilon\left(\left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i \mathbb{1}_E a_i a_i^{\mathrm{T}}\right\| \geq \frac{1}{2}\left(\hat{C}\lceil \|x\|_2^2\rceil \sqrt{\log(d)}\sqrt{\frac{d}{m}} - \frac{14}{\sqrt{m}}\lceil \|x\|_2^2\rceil\right)\right) \\ &\leq \mathbb{P}_\varepsilon\left(\left\|\frac{1}{m}\sum_{i=1}^m \varepsilon_i y_i \mathbb{1}_E a_i a_i^{\mathrm{T}}\right\| \geq \frac{1}{2}\left(C\lceil \|x\|_2^2\rceil \sqrt{\log(d)}\sqrt{\frac{d}{m}}\right)\right), \end{aligned}$$

which we already know is bounded from above by $\frac{1}{d} + 4m\exp(-cd)$ for an absolute constant $c > 0$.

In summary, we found that there exist absolute constants $C, c > 0$ so that

$$\mathbb{P}\left(\|S - \mathbb{E}[S]\| \geq C\lceil \|x\|_2^2\rceil \sqrt{\log(d)}\log(m)\sqrt{\tfrac{d}{m}}\right) \leq 24m\exp(-cd) + \frac{6}{d} + \frac{18}{m}.$$

$\square$

Lemma A.13. *Let $m \geq d\log(d)$. If $y_i \sim$ Bernoulli$\left(1 - \exp\left(-\left|\langle a_i, x\rangle\right|^2\right)\right)$, $i \in [m]$, with independent random vectors $a_i \sim \mathcal{N}(0, I_d)$, $i \in [m]$, there exist absolute constants $C, c_1, c_2 > 0$, such that, with probability at least*

$$1 - \frac{2}{d} - 8m\exp\left(-c_1 d\right) - 2\exp\left(-c_2 d\log(d)\right),$$

it holds that

$$\|S - \mathbb{E}\left[S\right]\| \leq C\sqrt{\log(d)}\sqrt{\frac{d}{m}},$$

where $S := \frac{1}{m}\sum_{i=1}^m y_i a_i a_i^{\mathrm{T}}$.

Proof. We follow the same steps as in the proof of Lemma A.12. Again, we find

$$\begin{aligned}
&\mathbb{P}\left(\|S - \mathbb{E}\left[S\right]\| \geq t\right) \\
&\leq \mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\left(y_i\left|\langle a_i, z\rangle\right|^2 - y_i'\left|\langle a_i', z\rangle\right|^2\right)\right| \geq \frac{t}{2}\right) \\
&\quad + \sup_{\|z\|_2=1}\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m y_i\left|\langle a_i, z\rangle\right|^2 - z^{\mathrm{T}}\mathbb{E}\left[S\right]z\right| \geq \frac{t}{2}\right),
\end{aligned}$$

where y_i', a_i', $i \in [m]$, are independent copies of y_i, a_i, $i \in [m]$. Using a Rademacher random vector $\varepsilon = (\varepsilon_i)_{i\in[m]}$, independent of y_i', a_i', y_i, a_i, $i \in [m]$, the first term satisfies

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\left(y_i\left|\langle a_i, z\rangle\right|^2 - y_i'\left|\langle a_i', z\rangle\right|^2\right)\right| \geq \frac{t}{2}\right) \\
&= \mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\varepsilon_i\left(y_i\left|\langle a_i, z\rangle\right|^2 - y_i'\left|\langle a_i', z\rangle\right|^2\right)\right| \geq \frac{t}{2}\right) \\
&\leq 2\mathbb{P}\left(\frac{1}{m}\sup_{\|z\|_2=1}\left|\sum_{i=1}^m\varepsilon_i y_i\left|\langle a_i, z\rangle\right|^2\right| \geq \frac{t}{4}\right).
\end{aligned}$$

Let $p \geq \log(d)$. Using the equivalence of the p-Schatten norm and the spectral norm for $p \geq \log(d)$ and the Theorem A.3, we obtain

$$\begin{aligned}
\mathbb{E}_\varepsilon\left\|\frac{1}{m}\sum_{i=1}^m\varepsilon_i y_i a_i a_i^{\mathrm{T}}\right\|^p &\leq e^p\cdot\mathbb{E}_\varepsilon\left\|\frac{1}{m}\sum_{i=1}^m\varepsilon_i y_i a_i a_i^{\mathrm{T}}\right\|_{C_p^d}^p \leq \left(C\sqrt{p}\cdot\frac{1}{m}\left\|\sum_{i=1}^m y_i^2 a_i a_i^{\mathrm{T}} a_i a_i^{\mathrm{T}}\right\|_{C_p^d}^{\frac{1}{2}}\right)^p \\
&\leq \left(C\sqrt{p}\cdot\frac{1}{m}\left\|\sum_{i=1}^m y_i^2\|a_i\|_2^2 a_i a_i^{\mathrm{T}}\right\|^{\frac{1}{2}}\right)^p,
\end{aligned}$$

with an absolute constant $C > 0$.

Again, we are required to bound

$$\left\| \sum_{i=1}^m y_i^2 \|a_i\|_2^2 a_i a_i^{\mathrm{T}} \right\|^{\frac{1}{2}} = \sup_{\|z\|_2=1} \left(\sum_{i=1}^m y_i^2 \|a_i\|_2^2 \langle a_i, z\rangle^2 \right)^{\frac{1}{2}} \leq \sup_{\|z\|_2=1} \left(\sum_{i=1}^m \|a_i\|_2^2 \langle a_i, z\rangle^2 \right)^{\frac{1}{2}},$$

where we use that $y_i \in \{0,1\}$.

As in the proof of Lemma A.12, we find that, with probability at least $1 - 2\exp(-cd)$ with respect to $a_i,\ i \in [m]$, it holds

$$\left| \|a_i\|_2^2 - \mathbb{E}_A \|a_i\|_2^2 + d \right| \leq 2d,$$

and, with probability at least $1 - 2\exp(-\tilde{c}d)$ on the $a_i,\ i \in [m]$,

$$\sup_{\|z\|_2=1} \left(\sum_{i=1}^m \langle a_i, z\rangle^2 \right)^{\frac{1}{2}} \leq \tilde{C}\sqrt{m},$$

where $\tilde{C}, \tilde{c} > 0$ are absolute constants. Here, we use that $m \geq d\log(d) > d$ for d sufficiently large. Consequently, there exist absolute constants $C, c > 0$ so that, for all $p \geq \log(d)$,

$$\left(\mathbb{E}_\varepsilon \left\| \frac{1}{m} \sum_{i=1}^m \varepsilon_i y_i a_i a_i^{\mathrm{T}} \right\|^p \right)^{\frac{1}{p}} \leq C\sqrt{p}\frac{\sqrt{d}}{\sqrt{m}}$$

with probability at least $1 - 4m\exp(-cd)$ with respect to $a_i,\ i \in [m]$.

Due to Theorem A.7, it holds

$$\mathbb{P}\left(\left\| \frac{1}{m} \sum_{i=1}^m \varepsilon_i y_i a_i a_i^{\mathrm{T}} \right\| \geq \frac{t}{4} \right) \leq \exp\left(-\frac{t^2}{32eC^2 \frac{d}{m}} \right) + 4m\exp(-cd)$$

for all $t \geq 4\sqrt{e}C\sqrt{\frac{d}{m}} \cdot \sqrt{\log(d)}$, with absolute constants $C, c > 0$. Choosing $t = \sqrt{2} \cdot 4\sqrt{e}C\sqrt{\frac{d}{m}} \cdot \sqrt{\log(d)}$, we obtain that there exist absolute constants $\tilde{C}, c > 0$ so that

$$\begin{aligned} &\mathbb{P}\left(\|S - \mathbb{E}[S]\| \geq \tilde{C}\sqrt{\log(d)}\sqrt{\tfrac{d}{m}} \right) \\ &\leq \frac{2}{d} + 8m\exp(-cd) \\ &\quad + \sup_{\|z\|_2=1} \mathbb{P}\left(\left| \frac{1}{m}\sum_{i=1}^m y_i |\langle a_i, z\rangle|^2 - z^{\mathrm{T}}\mathbb{E}[S]\,z \right| \geq \frac{1}{2}\tilde{C}\sqrt{\log(d)}\sqrt{\frac{d}{m}} \right). \end{aligned}$$

It remains to bound the latter term on the right-hand side. We could proceed again as in the proof of Lemma A.12. However, for the Bernoulli model, we can also state an alternative proof using Bernstein's inequality (Theorem A.4), which involves less tedious computations than applying Lemma A.8. Fix $z \in \mathbb{R}^d$ with $\|z\|_2 = 1$. The product of the squared Gaussian and the bounded random variable y_i is sub-exponential with $\|y_i \left|\langle a_i, z\rangle\right|^2 \|_{\psi_1} \leq 1$. Bernstein's inequality for independent mean zero sub-exponential random variables provides that, for every $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^m y_i \left|\langle a_i, z\rangle\right|^2 - m z^{\mathrm{T}} \mathbb{E}\left[S\right] z\right| \geq \frac{tm}{2}\right) \leq 2\exp\left(-c\min\left\{\frac{t^2m^2}{4m}, \frac{tm}{2}\right\}\right),$$

where $c > 0$ is an absolute constant. For $t = \tilde{C}\sqrt{\log(d)}\sqrt{\frac{d}{m}}$, we obtain

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m y_i \left|\langle a_i, z\rangle\right|^2 - z^{\mathrm{T}} \mathbb{E}\left[S\right] z\right| \geq \frac{1}{2}\tilde{C}\sqrt{\log(d)}\sqrt{\frac{d}{m}}\right)$$
$$\leq 2\exp\left(-cm\cdot\min\left\{\frac{1}{4}\tilde{C}^2\log(d)\frac{d}{m}, \frac{1}{2}\tilde{C}\sqrt{\log(d)}\sqrt{\frac{d}{m}}\right\}\right),$$

for an absolute constant $c > 0$. As the bound does not depend on z, we found that

$$\sup_{\|z\|_2=1}\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m y_i \left|\langle a_i, z\rangle\right|^2 - z^{\mathrm{T}} \mathbb{E}\left[S\right] z\right| \geq \frac{1}{2}\tilde{C}\sqrt{\log(d)}\sqrt{\frac{d}{m}}\right)$$
$$\leq 2\exp\left(-cm\cdot\min\left\{\frac{1}{4}\tilde{C}^2\log(d)\frac{d}{m}, \frac{1}{2}\tilde{C}\sqrt{\log(d)}\sqrt{\frac{d}{m}}\right\}\right).$$

Summarizing all bounds, we found that there exist absolute constants $C, c_1, c_2 > 0$ so that

$$\mathbb{P}\left(\|S - \mathbb{E}\left[S\right]\| \geq C\sqrt{\log(d)}\sqrt{\frac{d}{m}}\right)$$
$$\leq \frac{2}{d} + 8m\exp\left(-c_1 d\right) + 2\exp\left(-c_2 m\cdot\min\left\{\sqrt{\log(d)}\sqrt{\frac{d}{m}}, \log(d)\frac{d}{m}\right\}\right).$$

Since $m \geq d\log(d)$, we conclude

$$\mathbb{P}\left(\|S - \mathbb{E}\left[S\right]\| \geq C\sqrt{\log(d)}\sqrt{\tfrac{d}{m}}\right) \leq \frac{2}{d} + 8m\exp\left(-c_1 d\right) + 2\exp\left(-c_2 d\log(d)\right).$$

$\square$

B. Wirtinger calculus, gradient descent and auxiliary lemmas for Section 6.4

In this chapter, we provide some insights into the theory of Wirtinger calculus, attributed to the seminal work of Wirtinger on functions with complex variables [134]. The following overview is based on [17].

Consider a function

$$f(z) = u(x, y) + iv(x, y), \ z = x + iy, \ x, y \in \mathbb{R}^d,$$

with real-valued and differentiable functions u and v. Such functions are not necessarily holomorphic, and hence the standard complex derivative does not exist for all functions f. However, f can also be considered as a function of the conjugate variables $z = x+iy$ and $\bar{z} = x - iy$, treating them as independent. Fixing $\bar{z}$, the function $f(z, \bar{z})$ is holomorphic with respect to z for a fixed $\bar{z}$ since u and v are differentiable. The same holds true for exchanged roles of z and $\bar{z}$.

Definition B.1. *The Wirtinger derivative of f is defined as*

$$\partial_z f = \tfrac{1}{2} \left(\partial_x f - i\partial_y f\right),$$

and the conjugate Wirtinger derivative is

$$\partial_{\bar{z}} f = \tfrac{1}{2} \left(\partial_x f + i\partial_y f\right).$$

If f is complex differentiable in the standard sense, the Wirtinger derivatives can be shown to coincide with the respective complex derivatives.

The Wirtinger derivatives $\partial_z f$ and $\partial_{\bar{z}} f$ can also be expressed as

$$\partial_z f = \partial_z f(z, \bar{z})\big|_{\bar{z} \text{ constant}} = [\partial_{z_1} f(z, \bar{z}), \ldots, \partial_{z_n} f(z, \bar{z})] \big|_{\bar{z} \text{ constant}},$$

$$\partial_{\bar{z}} f = \partial_{\bar{z}} f(z, \bar{z})\big|_{z \text{ constant}} = [\partial_{\bar{z}_1} f(z, \bar{z}), \ldots, \partial_{\bar{z}_n} f(z, \bar{z})] \big|_{z \text{ constant}},$$

and the common differentiation rules apply. Accordingly, the Wirtinger gradient and Wirtinger Hessian are given as

$$\nabla f(z) = \begin{pmatrix} (\partial_z f)^* \\ (\partial_{\bar{z}} f)^* \end{pmatrix}, \quad \nabla^2 f(z) = \begin{pmatrix} \partial_z (\partial_z f)^* & \partial_{\bar{z}} (\partial_z f)^* \\ \partial_z (\partial_{\bar{z}} f)^* & \partial_{\bar{z}} (\partial_{\bar{z}} f)^* \end{pmatrix}.$$

The Wirtinger derivatives satisfy

$$\overline{\partial_z f} = \partial_{\bar{z}} \bar{f}, \quad \text{and} \quad \overline{\partial_{\bar{z}} f} = \partial_z \bar{f}.$$

Hence, if f is a real-valued function, that is, $f(z) = u(x, y)$, we have

$$\partial_{\bar{z}} f = \overline{\partial_z f}, \quad \partial_{\bar{z}}(\partial_{\bar{z}} f)^* = \overline{\partial_z (\partial_z f)^*}, \quad \partial_z(\partial_{\bar{z}} f)^* = \overline{\partial_{\bar{z}}(\partial_z f)^*}.$$

Consequently, we can use the simplified notation

$$\nabla_z f := (\partial_z f)^*, \quad \nabla^2_{z,z} f := \partial_z(\partial_z f)^*, \text{ and } \quad \nabla^2_{z,\bar{z}} f := \partial_z(\partial_{\bar{z}} f)^*$$

for real-valued functions.

The second-order Taylor polynomial of f at a point z_0 is

$$P_f(v, z_0) = f(z_0) + (\nabla f(z_0))^* \begin{pmatrix} v \\ \bar{v} \end{pmatrix} + \begin{pmatrix} v \\ \bar{v} \end{pmatrix}^* \nabla^2 f(z) \begin{pmatrix} v \\ \bar{v} \end{pmatrix}.$$

For a real-valued function f, the quadratic term of the Taylor polynomial P_f can be written as

$$\begin{pmatrix} v \\ \bar{v} \end{pmatrix}^* \nabla^2 f(z) \begin{pmatrix} v \\ \bar{v} \end{pmatrix} = 2\mathrm{Re}\left(v^* \nabla^2_{z,z} f(z)\, v\right) + 2\mathrm{Re}\left(v^*\, \nabla^2_{\bar{z},z} f(z)\, \bar{v}\right). \tag{B.1}$$

In this thesis, we use Wirtinger derivatives in algorithms for first-order optimization problems involving real-valued functions f. The direction of the Wirtinger gradient $\nabla_z f$ is the direction of ascent of f. Hence, for minimizing a function f, we can perform a gradient descent algorithm with update rule

$$z_{k+1} = z_k - \mu_k\, \nabla_z f(z_k), \tag{B.2}$$

using an appropriate initial vector $z_0 \in \mathbb{C}^d$. The parameter $\mu_k > 0$ is commonly known as step size or learning rate. It can be chosen to be constant or adaptive, preferably in a way that ensures descent in every iteration, that is, $f(z_{k+1}) \leq f(z_k)$ for all $k \geq 0$.

Convergence of such Wirtinger gradient descent algorithms can be proven by means of the following result, a proof of which can be found in [44].

Proposition B.1. *Let* $b \in \mathbb{R}$, $f : \mathbb{C}^d \to [b, \infty)$, *be a twice Wirtinger differentiable function with a uniformly bounded Hessian, that means*

$$\begin{pmatrix} v \\ \bar{v} \end{pmatrix}^* \nabla^2 f(z) \begin{pmatrix} v \\ \bar{v} \end{pmatrix} \leq L \left\| \begin{pmatrix} v \\ \bar{v} \end{pmatrix} \right\|_2^2$$

for all $z, v \in \mathbb{C}^d$, *with a constant* $L > 0$ *independent of* z. *Let the sequence* $(z_k)_{k\geq 0} \subset \mathbb{C}^d$ *be generated by the update scheme* (B.2) *with an arbitrary initialization* $z_0 \in \mathbb{C}^d$. *If* $0 < \mu \leq L^{-1}$, *then*

$$f(z_k) - f(z_{k+1}) \geq \mu \left\| \nabla_z f(z_{k+1}) \right\|_2^2$$

for all $k \geq 0$. *Consequently, if* f *has compact sublevel sets* $L_s(f) = \{z \in \mathbb{C}^d : f(z) \leq s\}$, *the Wirtinger flow algorithm* (B.2) *is guaranteed to converge to a stationary point of* f.

We conclude this chapter with two auxiliary lemmas used in the proofs presented in Chapter 6. They are based on joint work with Benedikt Diederichs and Frank Filbir and have been reported in [33].

We aim to specify the requirements of Proposition B.1 to loss functions that are given as a composition $\ell \circ \varphi$ of a smooth function $\ell : \mathbb{R} \to \mathbb{R}$ and $\varphi : \mathbb{C}^d \to [0, \infty)$, $\varphi(z) = |\langle a, z\rangle|^2$. We obtain the following bound on the Hessian.

Lemma B.2. *Let* $\ell : \mathbb{R} \to \mathbb{R}$ *be twice differentiable and* $\varphi : \mathbb{C}^d \to [0, \infty)$, $\varphi(z) = |\langle a, z\rangle|^2$ *with* $a \in \mathbb{C}^d$. *Then*

$$\begin{pmatrix} v \\ \overline{v} \end{pmatrix}^* \nabla^2(\ell \circ \varphi)(z) \begin{pmatrix} v \\ \overline{v} \end{pmatrix} \leq \left(2\ell''\left(|\langle a, z\rangle|^2\right) |\langle a, z\rangle|^2 + \ell'\left(|\langle a, z\rangle|^2\right)\right) \|a\|_2^2 \left\| \begin{pmatrix} v \\ \overline{v} \end{pmatrix} \right\|_2^2$$

for all $z, v \in \mathbb{C}^d$.

Proof. The Wirtinger derivative of $(\ell \circ \varphi)(z, \overline{z}) = \ell(\overline{z}^{\mathrm{T}} a a^* z)$ is

$$\partial_z(\ell \circ \varphi)(z, \overline{z}) = \partial_z(\ell \circ \varphi)(z, \overline{z}) \,|_{\overline{z} \text{ constant}} = \ell'(\overline{z}^{\mathrm{T}} a a^* z)\overline{z}^{\mathrm{T}} a a^* = \ell'(|\langle a, z\rangle|^2) z^* a a^*.$$

The second derivatives are

$$\partial_z\Big(\partial_z(\ell \circ \varphi)(z, \overline{z})\Big)^* = \partial_z\Big(\ell'(|\langle a, z\rangle|^2) a a^* z\Big) = \ell''(|\langle a, z\rangle|^2) \, |\langle a, z\rangle|^2 \, a a^* + \ell'(|\langle a, z\rangle|^2) a a^*,$$

and

$$\partial_{\overline{z}}\Big(\partial_z(\ell \circ \varphi)(z, \overline{z})\Big)^* = \partial_{\overline{z}}\Big(\ell'(|\langle a, z\rangle|^2) a a^* z\Big) = \ell''(|\langle a, z\rangle|^2) \, \langle a, z\rangle^2 \, a a^{\mathrm{T}}.$$

With (B.1) we obtain

$$\begin{aligned}
\begin{pmatrix} v \\ \overline{v} \end{pmatrix}^* \nabla^2(\ell \circ \varphi)(z) \begin{pmatrix} v \\ \overline{v} \end{pmatrix} &= 2\Big(\ell''(|\langle a, z\rangle|^2) \, |\langle a, z\rangle|^2 + \ell'(|\langle a, z\rangle|^2)\Big) \, |\langle a, v\rangle|^2 \\
&\quad + 2\mathrm{Re}\Big(\ell''(|\langle a, z\rangle|^2) \, \langle a, z\rangle^2 \, \langle a, v\rangle^2\Big) \\
&\leq 2\Big(2\ell''(|\langle a, z\rangle|^2) \, |\langle a, z\rangle|^2 + \ell'(|\langle a, z\rangle|^2)\Big) \, |\langle a, v\rangle|^2 \\
&\leq \Big(2\ell''(|\langle a, z\rangle|^2) \, |\langle a, z\rangle|^2 + \ell'(|\langle a, z\rangle|^2)\Big) \, \|a\|_2^2 \left\| \begin{pmatrix} v \\ \overline{v} \end{pmatrix} \right\|_2^2,
\end{aligned}$$

where we used $\mathrm{Re}(\alpha^2 \, \beta^2) \leq |\alpha|^2 \, |\beta|^2$ for $\alpha, \beta \in \mathbb{C}$. □

Consequently, we can apply the convergence result of Proposition B.1 if we find an upper bound for $2\,\ell''(t)\,t + \ell'(t)$ that holds uniformly for all $t \in [0,\infty)$, and if the level sets of $\ell \circ \varphi$ are compact. The second condition is addressed in the following lemma.

Lemma B.3. *Let* $f : \mathbb{C}^d \to \mathbb{R}$, $z \mapsto \sum_{i=1}^m |\langle a_i, z\rangle|^2$ *with* $a_i \in \mathbb{C}^d$, $i \in [m]$. *Denote with* A *the matrix with rows* a_i^*, $i \in [m]$. *For any* $z_0 \in \mathbb{C}^d$, f *restricted to* $z_0 + \operatorname{range}(A^*)$ *has compact level sets.*

Proof. In this proof, we use that $\sum_{i=1}^m |\langle a_i, z\rangle|^2 = \|Az\|_2^2$.

Consider any $z \in z_0 + \operatorname{range}(A^*)$. Decompose $z_0 = \hat{z}_0 + \tilde{z}_0$ with $\hat{z}_0 \in \ker(A)$ and $\tilde{z}_0 \in \operatorname{range}(A^*)$, and define $\tilde{z} \coloneqq z - \hat{z}_0$. The vector $\tilde{z}$ is the orthogonal projection of z onto the range of A^*, and hence is an element of the orthogonal complement of the kernel of A.

Let $A = U\Sigma V^*$ be the singular value decomposition of A with unitary matrix $U \in \mathbb{C}^{m\times m}$, unitary matrix $V \in \mathbb{C}^{d\times d}$ with columns v_i, $i \in [d]$, and diagonal matrix $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{m\times d}$ with singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$. Since $\hat{z}_0 \in \ker(A)$ and U is unitary, we obtain

$$\|Az\|_2^2 = \|A\tilde{z}\|_2^2 = \|U\Sigma V^*\tilde{z}\|_2^2 = \|\Sigma V^*\tilde{z}\|_2^2 = \sum_{i=1}^d \sigma_i^2 \,|\langle v_i, \tilde{z}\rangle|^2 .$$

As $\tilde{z}$ is orthogonal to the kernel of A, it holds $\langle v_i, \tilde{z}\rangle = 0$ for all i with $\sigma_i = 0$. Consequently, it is $\sum_{i=1}^d \sigma_i^2 \,|\langle v_i, \tilde{z}\rangle|^2 \geq \sum_{i=1}^d \left(\sigma_{\min}^+\right)^2 |\langle v_i, \tilde{z}\rangle|^2$, where $\sigma_{\min}^+$ denotes the smallest singular value of A that is larger than zero. We continue to bound

$$\|Az\|_2^2 \geq \left(\sigma_{\min}^+\right)^2 \sum_{i=1}^d |\langle v_i, \tilde{z}\rangle|^2 = \left(\sigma_{\min}^+\right)^2 \|V^*\tilde{z}\|_2^2 = \left(\sigma_{\min}^+\right)^2 \|\tilde{z}\|_2^2 \geq \left(\sigma_{\min}^+\right)^2 \left(\|z\|_2^2 - \|\hat{z}_0\|_2^2\right),$$

using that V is unitary and the triangle inequality.

Note that $\sigma_{\min}^+$ and $\|\hat{z}_0\|_2^2$ are constant for all $z \in z_0 + \operatorname{range}(A^*)$. Thus, restricted to vectors $z \in z_0 + \operatorname{range}(A^*)$, $\|Az\|_2^2$ is bounded from below by a scaled and shifted version of $\|z\|_2^2$. As the function $z \mapsto \|z\|_2^2$ has compact level sets, f has compact level sets on $z_0 + \operatorname{range}(A^*)$. □

Remark B.4. *According to Lemma B.3,* $\sum_{i=1}^m \ell_i(|\langle a_i, z\rangle|^2)$ *has compact level sets on* $z_0 + \operatorname{range}(A^*)$ *if the functions* $\ell_i : [0,\infty) \to \mathbb{R}$, $i \in [m]$, *are continuous, bounded from below, and satisfy* $\ell_i(t) \to \infty$ *for* $t \to \infty$.

C. Basic properties of Fourier analysis and auxiliary lemmas for Section 8.4

We summarize well-known basic properties of Fourier analysis that are used in the course of this manuscript.

The following properties of the discrete Fourier transform and the circular shift operator, modulation and reflection operator are based on [100, Theorem 3.26].

Lemma C.1. *For all $v \in \mathbb{C}^d$ and $r \in [d]_0$, we have*

$$(i)\ F(S_r v) = M_{-r}(Fv), \quad (ii)\ F\overline{v} = R\overline{Fv}, \quad \text{and} \quad (iii)\ FRv = RFv.$$

Corollary C.2. *The Fourier transform of v is conjugate symmetric, that means, it satisfies $Fv = R\overline{Fv}$, if and only if $v \in \mathbb{R}^d$.*

We further need the following properties, which are all proven in [96].

Lemma C.3. *For all $u, v \in \mathbb{C}^d$ and $r \in [d]_0$, we have*

(i) $|Fv|^2 = F\left(v * \overline{Rv}\right)$,

(ii) $\overline{RS_r v} = S_{-r}\overline{Rv}$, *and*

(iii) $\left(\left[(u \circ S_r v) * \left(\overline{Ru} \circ S_{-r}\overline{Rv}\right)\right]\right)_n = \left(\left[(u \circ S_n \overline{u}) * \left(Rv \circ S_{-n}\overline{Rv}\right)\right]\right)_r, \ n \in [d]_0.$

The following two properties are known as the (discrete) convolution theorem.

Theorem C.4. *For all $u, v \in \mathbb{C}^d$, it holds*

$$(i)\ dF(u \circ v) = Fu * Fv, \quad \text{and} \quad (ii)\ F(u * v) = Fu \circ Fv.$$

Moreover, we work with Plancherel's identity.

Lemma C.5. *For all $v \in \mathbb{C}^d$, it is*

$$\|Fv\|_2^2 = d\,\|v\|_2^2\,.$$

In the following, we present the proofs of the auxiliary lemmas used in the proof of Theorem 8.4.2. They are based on joint work with Oleh Melnyk and have been published in [87].

The following results serve an analysis of the linear systems (8.14). We recall that this means the system of equations

$$A^j \begin{pmatrix} \cos(\varphi_j) \\ \sin(\varphi_j) \end{pmatrix} = \frac{d^2}{2|f_0^j|} \begin{pmatrix} 1 - \frac{1}{d^2}\left(|a_0^j|^2 + |f_0^j|^2\right) \\ \vdots \\ 1 - \frac{1}{d^2}\left(|a_{d-1}^j|^2 + |f_0^j|^2\right) \end{pmatrix},$$

where $f_k^j := (F[x \circ S_j \overline{x}])_k$, $\varphi_j := \arg(f_0^j)$,

$$a^j := dF^{-1}\begin{pmatrix} 0 \\ f_1^j \\ \vdots \\ f_{d-1}^j \end{pmatrix} \quad \text{and} \quad A^j := \begin{pmatrix} \operatorname{Re}(a_0^j) & \operatorname{Im}(a_0^j) \\ \vdots & \vdots \\ \operatorname{Re}(a_{d-1}^j) & \operatorname{Im}(a_{d-1}^j) \end{pmatrix}.$$

For any $j \in [d]_0$, the three possible cases $\operatorname{rank}(A^j) = 0$, $\operatorname{rank}(A^j) = 1$, and $\operatorname{rank}(A^j) = 2$ are examined.

Note that we write $F^{-1}f^j$ instead of $x \circ S_j \overline{x}$ since there is not necessarily a unique vector x corresponding to $F^{-1}f^j$.

Lemma C.6. *Let $j \in [d]_0$. The following claims are equivalent:*

(i) The matrix A^j has $\operatorname{rank}(A^j) = 0$.

(ii) $(F^{-1}f^j)_t = e^{i\varphi_j}$ for all $t \in [d]_0$, where $\varphi_j = \arg(f_0^j)$.

(iii) The zeroth Fourier coefficient satisfies $|f_0^j| = d$.

Proof. $(i) \Rightarrow (ii), (iii)$: Let $\operatorname{rank}(A^j) = 0$. Then, according to the definition in (8.15), it must be $a_t^j = 0$ for all $t \in [d]_0$. With this, for all $t \in [d]_0$ we obtain

$$(F^{-1}f^j)_t = \frac{1}{d}\sum_{k=0}^{d-1} e^{\frac{2\pi i k \cdot t}{d}} f_k^j = \frac{1}{d}\left(f_0^j + a_t^j\right) = \frac{1}{d} f_0^j = \frac{1}{d}|f_0^j| e^{i\varphi_j}.$$

Hence, $F^{-1}f^j$ is constant equal to f_0^j/d. Proposition 8.3.1 states that $F^{-1}f^j \in \mathbb{T}^d$, and therefore $(F^{-1}f^j)_t = e^{i\varphi_j}$ for all $t \in [d]_0$ and $|(F^{-1}f^j)_t| = 1$, thus $|f_0^j| = d$.

$(ii) \Rightarrow (iii)$: If $(F^{-1}f^j)_t = e^{i\varphi_j} \in \mathbb{T}$ for all $t \in [d]_0$, it is

$$f_k^j = (F(F^{-1}f^j))_k = \sum_{t=0}^{d-1} e^{-\frac{2\pi i t k}{d}} e^{i\varphi_j} = \begin{cases} d e^{i\varphi_j}, & k = 0, \\ 0, & k \in [d]_0 \setminus \{0\}. \end{cases}$$

Consequently, the magnitude of the lost Fourier coefficient is equal to $|f_0^j| = d|e^{i\varphi_j}| = d$.

$(iii) \Rightarrow (i)$: Let $|f_0^j| = d$. Using Plancherel's identity in Lemma C.5, the definition in (8.15), and Proposition 8.3.1, we obtain

$$\left\|a^j\right\|_2^2 = \tfrac{1}{d}\left\|Fa^j\right\|_2^2 = d\left\|(0, f_1^j, \ldots, f_{d-1}^j)\right\|_2^2 = d\left(\left\|f^j\right\|_2^2 - |f_0^j|^2\right) = d^3 - d^3 = 0,$$

and conclude that $a_t^j = 0$ for all $t \in [d]_0$, which is equivalent to $\operatorname{rank}(A^j) = 0$. $\square$

Lemma C.7. *Let $j \in [d]_0$. If $\operatorname{rank}(A^j) = 0$ and j is coprime with d, the argument of f_0^j satisfies $\varphi_j \in \frac{2\pi}{d}\mathbb{Z}$ and the ground-truth object is $x \sim \left(e^{-\frac{2\pi i t j' m}{d}}\right)_{t \in [d]_0}$, for some $m \in [d]_0$, where j' denotes the multiplicative inverse of j modulo d.*

Proof. Assume that $\operatorname{rank}(A^j) = 0$ and that j is coprime with d. Let $x \in \mathbb{T}^d$ satisfy $x \circ S_j \overline{x} = F^{-1} f^j$. The equivalence of (i) and (ii) in Lemma C.6 gives $x_t \overline{x}_{t-j} = (F^{-1} f^j)_t = e^{i\varphi_j}$ for all $t \in [d]_0$. Thus, for every $t \in [d]_0$, there exists a $\zeta \in \mathbb{Z}$ so that

$$\arg(x_t) - \arg(x_{t-j}) = \varphi_j + 2\pi\zeta. \tag{C.1}$$

From this, we conclude that

$$\arg(x_0) - \arg(x_{-\eta j}) = \sum_{t=0}^{\eta-1} \big(\arg(x_{-tj}) - \arg(x_{-(t+1)j})\big) \in \eta\varphi_j + 2\pi\mathbb{Z}$$

for any $\eta \in \mathbb{Z}$. Under the assumption that j and d are coprime, the smallest integer that satisfies $\eta j = 0 \bmod d$ is $\eta = d$. It follows that

$$0 = \arg(x_0) - \arg(x_{-dj}) \in d\varphi_j + 2\pi\mathbb{Z},$$

implying that $\varphi_j \in \frac{2\pi}{d}\mathbb{Z}$.

To describe the corresponding object more closely, rewrite

$$x_0 \overline{x}_{-t} = \prod_{\ell=0}^{\tau-1} x_{-\ell j} \overline{x}_{-(\ell+1)j},$$

where τ needs to satisfy $\tau j = t \bmod d$. Since j and d are coprime, $\tau j = t \bmod d$ is equivalent to $\tau = tj' \bmod d$, where j' denotes the multiplicative inverse of j modulo d.

Then we use (C.1) to conclude that

$$\arg(x_0) - \arg(x_{-t}) = \sum_{\ell=0}^{tj'-1} \big(\arg(x_{-\ell j}) - \arg(x_{-(\ell+1)j}\big) \in tj'\varphi_j + 2\pi\mathbb{Z},$$

which, together with $\varphi_j \in \frac{2\pi}{d}\mathbb{Z}$, yields

$$\arg(x_{-t}) \in \arg(x_0) + 2\pi \left(\frac{tj'}{d} + 1\right) \mathbb{Z}.$$

In summary, the object is of type $x \sim \left(e^{-\frac{2\pi i t j' m}{d}}\right)_{t \in [d]_0}$, using the notion of equivalence up to a global phase. □

Lemma C.8. *Let $j \in [d]_0$ and $\operatorname{rank}(A^j) = 1$. There exist two solutions $\varphi_1^j \neq \varphi_2^j$ to (8.14) given by*

$$\varphi_q^j := \arg(a_0^j) + (-1)^q \arccos\left(\frac{d^2 - |a_0^j|^2 - |f_0^j|^2}{2|a_0^j||f_0^j|}\right), \quad q = 1, 2.$$

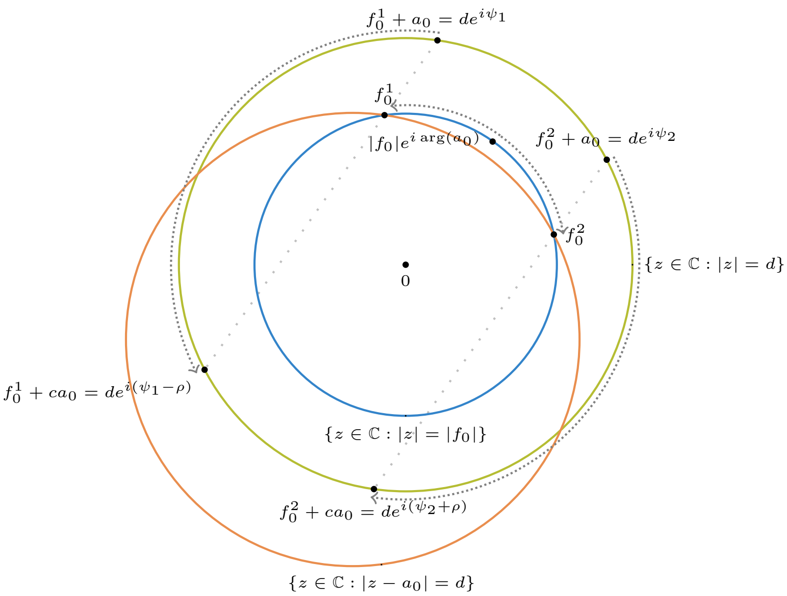


Figure 1.: Geometric visualization of Lemma C.8. By Proposition 8.3.1, the magnitude $|f_0^q|$ is known. This guarantees that all possible Fourier coefficients f_0^q are elements of the blue circle. By definition, $f_0^q + a_t = d(F^{-1}f^q)_t$ and $|(F^{-1}f^q)_t| = 1$, so that $f_0^q + a_t$ has to lie on the green circle for all $t \in [d]_0$. Consequently, the possible solutions f_0^q are also elements of the orange circle. Combining these facts, the possible solutions f_0^q correspond to the intersections of the orange circle with the blue circle. It can be proven that neither the blue and the orange circle coincide nor the orange circle is tangent to the blue circle, hence there are precisely two intersection points, corresponding to two solutions. The light-gray dotted lines in the direction of a_0 intersect with the blue circle at the solutions f_0^q. Since they have two distinct intersections with the green circle, the entries of the diagonal, $\frac{1}{d}(f_0^q + a_t) = \frac{1}{d}(f_0^q + c_t a_0)$, can take two distinct values.

Furthermore, there exists a disjoint partition $\mathcal{S}_j, \mathcal{S}_j^c$ of $[d]_0$ such that $0 \in \mathcal{S}_j$, $\mathcal{S}_j^c \neq \varnothing$, and for the corresponding vector of Fourier coefficients $f^{j,q}$ it is

$$(F^{-1}f^{j,q})_t = \begin{cases} e^{i\psi_q^j}, & t \in \mathcal{S}_j, \\ e^{i(\psi_q^j + (-1)^q \rho^j)}, & t \in \mathcal{S}_j^c, \end{cases} \qquad q = 1, 2,$$

with $\rho^j \in (-\pi, \pi) \setminus \{0\}$, $\psi_1^j \neq \psi_2^j$, and

$$\psi_2^j \in \psi_1^j - \rho^j + \pi + 2\pi\mathbb{Z}. \tag{C.2}$$

Proof. The statement in Lemma C.8 can be interpreted geometrically as a problem of intersecting circles. A visualization of the problem is displayed in Figure 1.

Let $j \in [d]_0$ be fixed. Throughout this proof, we leave out the index j to simplify the notation.

Since $\mathrm{rank}(A) = 1$, all equations in the linear system (8.14) are equal up to a multiplicative constant $c_t \in \mathbb{R}$, i.e.,

$$a_t = c_t \cdot a_0 \tag{C.3}$$

for all $t \in [d]_0$. If $|a_t| = 0$ for some $t \in [d]_0$, the corresponding equation

$$\mathrm{Re}(a_t)\cos(\varphi) + \mathrm{Im}(a_t)\sin(\varphi) = \frac{1}{2|f_0|}\left(d^2 - |a_t|^2 - |f_0|^2\right)$$

in (8.14) reduces to $d = |f_0|$. According to Lemma C.6, this is equivalent to $\mathrm{rank}(A) = 0$, contradicting $\mathrm{rank}(A) = 1$. We conclude that $|a_t| > 0$ and $c_t \neq 0$ for all $t \in [d]_0$.

The equations of (8.14) can also be written as

$$\begin{aligned} d^2 &= |a_t|^2 + |f_0|^2 + 2|f_0|\,\mathrm{Re}\left(a_t e^{-i\varphi}\right) \\ &= |a_t|^2 + |f_0|^2 + 2|f_0|\,\mathrm{Re}\left(|a_t| e^{i(\arg(a_t)-\varphi)}\right) \\ &= |a_t|^2 + |f_0|^2 + 2|f_0||a_t|\cos(\arg(a_t) - \varphi). \end{aligned}$$

As (8.14) is only of interest for $|f_0| > 0$ and, as just proved, $|a_t| > 0$ for all $t \in [d]_0$, this is equivalent to

$$\cos(\arg(a_t) - \varphi) = \frac{d^2 - |a_t|^2 - |f_0|^2}{2|a_t||f_0|},$$

which may have the two solutions

$$\varphi_q = \arg(a_t) + (-1)^q \arccos\left(\frac{d^2 - |a_t|^2 - |f_0|^2}{2|a_t||f_0|}\right), \quad q = 1, 2.$$

The argument φ_q must be the solution of all equations in (8.14) and is the same for all $t \in [d]_0$. By setting, e.g., $t = 0$ we obtain a formula for φ_q, $q = 1, 2$.

The system has exactly one solution, meaning that $\varphi_1 = \varphi_2$, if and only if, for all $t \in [d]_0$,

$$\frac{d^2 - |a_t|^2 - |f_0|^2}{2|a_t||f_0|} = \delta$$

with $\delta \in \{-1, 1\}$. Since this has to hold for all $t \in [d]_0$, also

$$d^3 - \sum_{t=0}^{d-1} |a_t|^2 - d|f_0|^2 = 2\delta|f_0| \sum_{t=0}^{d-1} |a_t|, \tag{C.4}$$

resulting as the sum of all these equations, must be true.

Using Plancherel's identity (Lemma C.5) together with (8.15) and Proposition 8.3.1, we obtain

$$\|a\|_2^2 = \tfrac{1}{d}\|Fa\|_2^2 = d\,\|(0, f_1, \ldots, f_{d-1})\|_2^2 = d\,\|f\|_2^2 - d|f_0|^2 = d^3 - d|f_0|^2,$$

and inserting this into (C.4) yields

$$0 = 2|f_0| \sum_{t=0}^{d-1} |a_t|. \tag{C.5}$$

Recall that the case $|f_0| = 0$ is not considered since it does not require to solve the linear system (8.14). Moreover, we proved before that $a_t \neq 0$ for all $t \in [d]_0$. Consequently, if $\operatorname{rank}(A) = 1$, (C.5) cannot be true, and we conclude that there always exist two solutions $\varphi_1 \neq \varphi_2$ that satisfy the equations (8.14). We denote the two corresponding vectors of Fourier coefficients by f^1 and f^2, respectively.

We aim to describe the corresponding diagonals $F^{-1} f^q$ more precisely.

Due to the existence of the ground-truth object x, the system (8.14) has at least one solution. Consequently, the right-hand side of (8.14) is an element of

$$\operatorname{im}(A) := \{v \in \mathbb{R}^d : Au = v \text{ for some } u \in \mathbb{R}^2\}.$$

Every $v \in \operatorname{im}(A)$ satisfies

$$v_t = \operatorname{Re}(a_t)u_1 + \operatorname{Im}(a_t)u_2 = c_t \operatorname{Re}(a_0)u_1 + c_t \operatorname{Im}(a_0)u_2 = c_t v_0,$$

and, thus, all entries of the right-hand side of (8.14) are multiples of the first entry of the right-hand side of (8.14), i.e.,

$$1 - \frac{1}{d^2}\left(|a_t|^2 + |f_0|^2\right) = c_t \left[1 - \frac{1}{d^2}\left(|a_0|^2 + |f_0|^2\right)\right]. \tag{C.6}$$

Substituting (C.3) into (C.6), we obtain

$$d^2 - c_t^2|a_0|^2 - |f_0|^2 = c_t \cdot \left(d^2 - |a_0|^2 - |f_0|^2\right).$$

With respect to c_t, this is a quadratic equation with two roots, $c := (|f_0|^2 - d^2)/|a_0|^2 < 0$ and 1, both of which are independent of t.

Let the two index sets $\mathcal{S}$ and $\mathcal{S}^c$ be defined via

$$\mathcal{S} := \{t \in [d]_0 : c_t = 1\} \quad \text{and} \quad \mathcal{S}^c := \{t \in [d]_0 : c_t = c\}.$$

By definition, $0 \in \mathcal{S}$, $\mathcal{S} \cap \mathcal{S}^c = \varnothing$, and $\mathcal{S} \cup \mathcal{S}^c = [d]_0$.

If $\mathcal{S}^c$ is empty, all entries of $(F^{-1}f^q)$ are equal to

$$(F^{-1}f^q)_t = \frac{1}{d}\sum_{j=0}^{d-1} e^{\frac{2\pi ijt}{d}} f_j^q = \frac{1}{d}\left(f_0^q + a_t\right) = \frac{1}{d}\left(f_0^q + 1\cdot a_0\right).$$

By Lemma C.6, this is equivalent to $\operatorname{rank}(A^j) = 0$, contradicting the assumption that $\operatorname{rank}(A) = 1$. Hence, it must be $\mathcal{S}^c \neq \varnothing$.

For $q \in \{1,2\}$, let ψ_q be the value satisfying

$$\left(F^{-1}f^q\right)_0 = e^{i\psi_q}.$$

Then, for both $q=1$ and $q=2$ the entries of $F^{-1}f^q$ are given by

$$(F^{-1}f^q)_t = \frac{1}{d}(f_0^q + a_0) = e^{i\psi_q}, \quad t\in\mathcal{S},$$

and

$$(F^{-1}f^q)_t = \frac{1}{d}(f_0^q + ca_0) = e^{i(\psi_q+\rho_q)}, \quad t\in\mathcal{S}^c,$$

for some $\rho_q \in (-\pi,\pi]\setminus\{0\}$.

We continue by showing that $\rho_2 = -\rho_1$. By definition, it is

$$e^{i(\psi_q+\rho_q)} = \tfrac{1}{d}(f_0^q + ca_0 + a_0 - a_0) = e^{i\psi_q} + (c-1)\cdot\frac{1}{d}a_0, \quad q=1,2.$$

For both $q=1$ and $q=2$, it holds

$$e^{i\psi_1}\left(1-e^{i\rho_1}\right) = (1-c)\cdot\frac{1}{d}a_0 = e^{i\psi_2}\left(1-e^{i\rho_2}\right). \tag{C.7}$$

Since $|e^{i\psi_1}| = |e^{i\psi_2}| = 1$, this yields

$$|1-e^{i\rho_1}|^2 = |1-e^{i\rho_2}|^2.$$

Writing this out, we obtain

$$\left(\operatorname{Re}(1-e^{i\rho_1})\right)^2 + \left(\operatorname{Im}(1-e^{i\rho_1})\right)^2 = \left(\operatorname{Re}(1-e^{i\rho_2})\right)^2 + \left(\operatorname{Im}(1-e^{i\rho_2})\right)^2,$$

hence

$$\left(1-\cos\rho_1\right)^2 + \left(-\sin\rho_1\right)^2 = \left(1-\cos\rho_2\right)^2 + \left(-\sin\rho_2\right)^2,$$

and, equivalently,

$$\cos\rho_1 = \cos\rho_2.$$

Since $\rho_1,\rho_2 \in (-\pi,\pi]\setminus\{0\}$, it must be either $\rho_2=\rho_1$ or $\rho_2=-\rho_1$. If $\rho_2=\rho_1$, then

(C.7) yields $\psi_1 = \psi_2$, hence $f_0^1 = f_0^2$ and also $\varphi_1 = \varphi_2$, contradicting $\varphi_1 \neq \varphi_2$. We conclude that $\rho_2 = -\rho_1$.

In particular, for $\rho_1 = \pi$, the only $\rho_2 \in (-\pi, \pi] \setminus \{0\}$ that satisfies $\cos(\rho_2) = \cos(\rho_1) = -1$ is $\rho_2 = \pi = \rho_1$, which contradicts $\rho_2 \neq \rho_1$. Therefore, we get $\rho_1 \in (-\pi, \pi) \setminus \{0\}$.

Finally, by (C.7), we find that

$$e^{i\psi_1} = e^{i\psi_2} \cdot \frac{e^{-i\rho_1} - 1}{e^{i\rho_1} - 1} = e^{i\psi_2} \cdot e^{-i\rho_1} \cdot \frac{1 - e^{i\rho_1}}{e^{i\rho_1} - 1} = e^{i\psi_2} \cdot (-e^{-i\rho_1}) = e^{i\psi_2} \cdot e^{-i\rho_1} \cdot e^{i\pi},$$

so that

$$\psi_1 = \psi_2 - \rho_1 + \pi + 2\pi\zeta$$

for some $\zeta \in \mathbb{Z}$. We set $\rho = \rho_2 = -\rho_1$ and obtain the statement of the lemma. $\square$

Lemma C.9. *(i) If* $\operatorname{rank}(A^j) = 2$, $j \in [d]_0$, *system* (8.14) *is uniquely solvable, and hence φ_j is uniquely recoverable from* (8.14)*, therefore $F^{-1}f^j$ is uniquely recoverable from measurements* (8.1)*.*

(ii) If $j \in [d]_0$ is coprime with d, x can be uniquely recovered from $F^{-1}f^j = x \circ S_j\overline{x}$.

Proof. (i) : If $\operatorname{rank}(A^j) = 2$, the linear system (8.14) has one solution. Consequently, we can recover φ_j uniquely, and thus also obtain back the lost Fourier coefficient $(F\left[x \circ S_j\overline{x}\right])_0$. Knowing $f^j = F\left[x \circ S_j\overline{x}\right]$, we can compute $F^{-1}f^j = x \circ S_j\overline{x}$.

(ii) : Let j and d be coprime. Then, j is a generator of the additive group of integers modulo d, denoted by $\mathbb{Z}_d$, which means that, for every $t \in [d]_0$, there exists $\zeta \in \mathbb{Z}$ with $\zeta j = t \bmod d$. Hence, for every $t \in [d]_0$, there exists $\zeta \in \mathbb{Z}$ with

$$x_0\overline{x}_t = \prod_{n=1}^{\zeta} x_{(n-1)j}\overline{x}_{nj} \Bigg/ \prod_{n=1}^{\zeta-1} |x_{nj}|^2.$$

Consequently, for all $t \in [d]_0$, the product $x_0\overline{x}_t$ can be constructed by the entries of $x \circ S_j\overline{x}$ together with the magnitudes of x, which, for a phase object, are a priori known to be equal to 1. The phase of the first coordinate $x_0 \in \mathbb{T}$ can be arbitrarily chosen since we consider uniqueness only up to a global phase. Subsequently, the phases of $x_t \in \mathbb{T}$ for all $t > 0$ can be recovered using the relations $x_0\overline{x}_t$. $\square$

Remark C.10. *If j is not coprime with d, j is not a generator of the group $\mathbb{Z}_d$. That means, $\mathcal{C}_0 := \{jt \bmod d,\ t \in [d]_0\} \neq \mathbb{Z}_d$, and there exists at least one $n \in \mathbb{Z}_d$ such that the set $\mathcal{C} := \{(jt + n) \bmod d,\ t \in [d]\}$ satisfies $\mathcal{C}_0 \cap \mathcal{C} = \varnothing$. Hence, for every object $x \in \mathbb{C}^d$ we can construct another object $\tilde{x}$ as*

$$\tilde{x}_t = \begin{cases} \alpha x_t, & t \in \mathcal{C}, \\ x_t, & t \notin \mathcal{C}, \end{cases}$$

with $\alpha \in \mathbb{T} \setminus \{1\}$, such that $\tilde{x} \not\sim x$ but $x \circ S_j\overline{x} = \tilde{x} \circ S_j\overline{\tilde{x}}$.