UNIVERSITY OF CALIFORNIA SAN DIEGO

Phase Retrieval from Locally Supported Measurements

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Mathematics with Specialization in Computational Science

by

Brian P. Preskitt

Committee in charge:

Professor Rayan Saab, Chair
Professor Kamalika Chaudhuri
Professor Yoav Freund
Professor Todd Kemp
Professor Jiawang Nie

2018

Copyright

Brian P. Preskitt, 2018

All rights reserved.

The dissertation of Brian P. Preskitt is approved:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2018

DEDICATION

This dissertation is dedicated to my brother, Charles Preskitt.
The need to prove you right got me past the finish line.

EPIGRAPH

For in much wisdom is much vexation, and he who increaseth knowledge increaseth sorrow.

—— Ecclesiastes 1:18

For this reason the scientist's life is not an easy one. However, on those rare occasions when his world does conform to the real one, and for this reason does throw light on the world around us, the rewards and the satisfactions are great and more than compensate for the many disappointments.

—— Herbert A. Hauptman

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Rayan Saab for all of his involvement in my academic career. In my first year of graduate school, it was his class on compressed sensing that convinced me to switch into the computational math track, which in turn is where I found the material and the problems that motivated me to continue in academic research. Since then, his support – in our many hours of teamwork at the whiteboard; his connections to problems, authors, conferences, and collaborators; and the simple things such as critiques of my writing and talks – was uniquely indispensable in shaping the environment in which I found my place as a researcher. I am immensely thankful for the inspiration of his talent and energetic leadership.

I would also like to thank Mark Iwen and Aditya Viswanathan, our two co-authors and collaborators on the subjects covered in this thesis. The ground that they broke on this new strain of phase retrieval proved to be incredibly fruitful, and I am thankful for being welcomed as a contributor to their ongoing research efforts. This partnership was foundational to my beginnings in research and much of my subsequent progress, and I am deeply indebted to their generous spirit of academic camaraderie. Chapter 3, Sections 3.1–3.8, in full, is a reprint of material published with Iwen, Saab, and Viswanathan as published in Applied and Computational Harmonic Analysis 2018. Chapter 7 Sections 7.1–7.3, in part, is a reprint of material published with these authors in the Proceedings of SPIE vol. 10394, 2017.

VITA

2013	Bachelor of Science, Texas Christian University
2013-2018	Teaching Assistant, University of California San Diego
2015	Master of Science, University of California San Diego
2016-2017	Associate Instructor, University of California San Diego
2018	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

M. Iwen, B. Preskitt, R. Saab, and A. Viswanathan. *An Eigenvector-Based Angular Synchronization Method for Phase Retrieval from Local Correlation Measurements.* arXiv:1612.01182. Accepted for publication June 2018.

M. Iwen, B. Preskitt, R. Saab, and A. Viswanathan. *Phase retrieval from local measurements in two dimensions.* Proceedings of SPIE 10394, Wavelets and Sparsity XVII, San Diego, CA, 2017.

ABSTRACT

Phase Retrieval from Locally Supported Measurements

by

Brian P. Preskitt

Doctor of Philosophy in Mathematics with Specialization in Computational Science

University of California San Diego, 2018

Professor Rayan Saab, Chair

In this dissertation, we study a new approach to the problem of phase retrieval, which is the task of reconstructing a complex-valued signal from magnitude-only measurements. This problem occurs naturally in several specialized imaging applications such as electron microscopy and X-ray crystallography. Although solutions were first proposed for this problem as early as the 1970s, these algorithms have lacked theoretical guarantees of success, and phase retrieval has suffered from a considerable gap between practice and theory for almost the entire history of its study.

A common technique in fields that use phase retrieval is that of *ptychography*, where measurements are collected by only illuminating small sections of the sample at any time. We refer to measurements designed in this way as *local measurements*, and in this dissertation, we develop and expand the theory for solving phase retrieval in measurement regimes of this

kind. Our first contribution is a basic model for this setup in the case of a one-dimensional signal, along with an algorithm that robustly solves phase retrieval under this model. This work is unique in many ways that represent substantial improvements over previously existing solutions: perhaps most significantly, many of the recovery guarantees in recent work rely on the measurements being generated by a random process, while we devise a class of measurements for which the conditioning of the system is known and quickly checkable (see Section 4.2). These advantages constitute major progress towards producing theoretical results for phase retrieval that are directly usable in laboratory settings.

Chapter 1 conducts a survey of the history of phase retrieval and its applications. Chapter 2 reviews the mathematical literature on the subject, including the first solutions and the theoretical work of the last decade. Chapter 3 presents co-authored results defining and establishing the setting and solution of the base model explored in this dissertation. Chapter 4 expands the theory on what measurement schemes are admissible in our model, including an analysis of conditioning and runtime. Chapter 5 explores results that bring our model nearer to the actual practice of ptychography. Chapter 6 includes a few relevant results that may be used for future expansion on this topic.

Chapter 1

History of Phase Retrieval

1.1 Introduction

Phase retrieval is the problem of solving a system of equations of the form

$$y = |Ax_0|^2 + \eta, \tag{1.1}$$

where $x_0 \in \mathbb{C}^d$ is the objective signal, $A \in \mathbb{C}^{D\times d}$ is a known measurement matrix, $\eta \in \mathbb{R}^D$ is an unknown perturbation vector, and $y \in \mathbb{R}^D$ is the vector of measurement data. $|\cdot|^2$ represents the component-wise magnitude squared operation; i.e. for any $n \in \mathbb{N}$ we have $|\mathbf{v}|_j^2 = |\mathbf{v}_j|^2$ for all $\mathbf{v} \in \mathbb{C}^n$. In phase retrieval, the goal is to recover an estimate of x_0 from knowledge of y and A. We sometimes rephrase the system (1.1) as

$$y_j = |\langle a_j, x_0\rangle|^2 + \eta_j, \tag{1.2}$$

where a_j^* stand for the rows of A and are referred to as the measurement vectors. The name *phase retrieval* comes from viewing the $|\cdot|^2$ operation as erasing the phases of the complex-valued measurements $\langle a_j, x_0\rangle$ and leaving only their magnitudes; solving for x_0 may be considered as a way of retrieving this phase information. We immediately note that this

problem contains an unavoidable phase ambiguity, in the sense that, for any solution x and any $\theta \in [0, 2\pi)$, we will have that $\mathrm{e}^{\mathrm{i}\theta}x$ is also a solution.

The phase retrieval problem appears in a multitude of imaging systems, since most optical sensors – most significantly, charge-coupled devices and photographic film – do not respond to the phase of an incoming light wave. Rather they respond only to the number and energy of photons arriving at its surface, so they indicate only the intensity (absolute value squared), and not the phase, of the electromagnetic waves to which they are exposed. This corresponds to our model in (1.1) by imagining that the i^{th} entry a_i^*x of $Ax \in \mathbb{C}^D$ corresponds to the magnitude and phase of the light arriving at the i^{th} pixel in an array of sensors. When such a sensor responds only to the amount of energy exciting it, we record $|a_i^*x|^2$ at each point and aggregate these data into the vector $|Ax|^2$. Areas of optics that encounter this problem include astronomy [98], diffraction imaging [**?**], and – among the earliest applications, and by far the most celebrated – x-ray crystallography [**?**]. Non-optical disciplines that can benefit from solutions to phase retrieval include speech recognition [4, 78], blind channel estimation [67, 69], and self-calibration [68].

The practice of these disciplines has produced many creative solutions to particular instances of the phase retrieval problem, and throughout the 20^{th} century the field largely evolved by the invention of *ad hoc* solutions that resolved the data at hand. Indeed, in their Nobel prize-winning work in 1915, William and Lawrence Bragg used their intensity-only measurements to deduce the crystal structures of sodium chloride, potassium chloride, and diamonds by largely geometrical analysis of their data based on strong prior knowledge of the atoms present in these materials [15, pp. 88-92, 102-105]. Attempts to systematize this process began in the 1930s with Arthur Lindo Patterson [76], with significant improvements coming with the work of Herbert Hauptman and Jerome Karle in the 1950s [52]. However, these solutions remained primarily non-algorithmic or made strong assumptions about the crystals being solved, slowing the solution process and limiting the progress made in phase retrieval outside x-ray crystallography. The method proposed by R.W. Gerchberg and W.O. Saxton in 1971 [45] shifted this paradigm by providing an algorithm that can be applied to fairly general data, with remarkably minimal assumptions made on the structure of

the object x being detected. This result inspired numerous variants (e.g., [11, 35, 39]), each of which empirically improved performance, but none of which produced a solid mathematical theory to explain why or when they would succeed. Physicists, chemists, and biologists made a great number of astounding scientific achievements in this fashion, but even with all this progress, the community remained largely in want of such a theoretical foundation that could offer reliable solutions in general settings until recent decades.

There are three main questions about phase retrieval problems that the scientific community would wish to answer theoretically: firstly, in an ideal, noiseless case where $\eta = 0$, for what matrices $A \in \mathbb{C}^{D\times d}$ does the system of equations (1.1) possess a unique solution (up to the known phase ambiguity)? Second, given such a case where a unique solution exists, is there an algorithm that can recover it? Third of all, when this recovery process exists, when is it stable in the sense that, in the presence of noise $\eta \neq 0$, the estimate x does not differ too much (or differs to a known degree, as a function of $\|\eta\|$) from x_0?

This dissertation expands upon the theory of phase retrieval by introducing a new class of matrices and an associated recovery algorithm that is proven to solve the system (1.1) with guaranteed stability to noise and with known, competitive computational cost.

1.2 X-ray Crystallography

1.2.1 Historical Preliminaries

The history of phase retrieval cannot be told without making mention of x-ray crystallography, the field that first brought scientific interest to this problem and by many metrics its most decorated and fruitful application. In x-ray crystallography, the goal is to gain an image of the positions of atoms within a molecule by illuminating a crystallized sample with x-rays. The molecular structure is deduced from the pattern of the radiation diffracted by the sample. A rough diagram of this setup is shown in figure 1.1.

Figure 1.1: Experimental setup for x-ray crystallography

This seemingly simple technique has been indispensable for the study of chemistry, biology, and physics, having been used to confirm or identify the arrangements of atoms in a wide variety of important compounds. Over a dozen discoveries made through x-ray crystallography – or made in developing the technique – have been recognized by Nobel Prizes in Physics, Chemistry, and Medicine or Physiology. Indeed, the first Nobel Prize in Physics was awarded to Wilhelm Röntgen in 1901 for his discovery of x-rays. The 1914 Prize in Physics was conferred upon Max von Laue for his discovery the diffraction of x-rays by atomic crystals, and in 1915 William and Lawrence Bragg earned the same distinction for performing the first complete characterizations of atomic crystal structures [44]. Since the time of these highly esteemed pioneering discoveries, x-ray crystallography has been used to produce accurate molecular models of a number of drugs (e.g., [17, 79, 82]), including penicillin in Dorothy Crowfoot Hodgkin's Nobel prize-winning work in 1963 [1]. It has elucidated several human biological compounds, including innumerable proteins [60, 96] and human DNA, for whose analysis in 1953 James Watson, Maurice Wilkins, and Francis Crick were awarded the Nobel prize in 1962, relying on the crystallographic images of Rosalind Franklin [42, 99, 100, 102]. And this technology remains extremely relevant today, playing an active role in material sciences, where crystallography is being used to characterize the degradation of lithium-ion batteries [53, 85] and to study carbon nanostructures such as fullerenes [61, 62], whose analysis earned the 1996 Nobel Prize in Chemistry [44].

With such a history, x-ray crystallography is an essential application of phase retrieval, and it is informative to state its mathematical formulation in this dissertation.

1.2.2 Mathematical Model

1.2.3 Diffraction as a Fourier transform

We begin by considering the mechanism of diffraction of waves.

1.3 Notation

In this section, we gather some of the notation that is used throughout the dissertation. Table 1.1 displays some of the most commonly used objects. We remark that, in this table and throughout this work, indices of a vector $x \in \mathbb{C}^n$ or matrix $A \in \mathbb{C}^{m\times n}$ are always taken modulo the appropriate dimension. For example, $x_{n+1} := x_1$ and $A_{00} = A_{mn}$.

Table 1.1: Common operators, objects, and sets. Definitions assume d, i, j, m, n, and k are arbitrary elements of $\mathbb{N}$ unless otherwise stated.

Parameters	Name and Type	Definition	Comments
$n \in \mathbb{Z}$	$[k]_n \subseteq \mathbb{N}$	$[k, k+n) \cap \mathbb{Z}$	We define $[k] = [k]_1$.
	$m \bmod_k n \in [n]_k$	The unique element p of $[n]_k$ satisfying $p \equiv m \bmod n$.	$m \bmod n = m \bmod_0 n$
	$\lvert i-j\rvert \bmod d$	$\min\{k \geq 0 : k \equiv \pm(i-j) \bmod d\}$	
	$I_d \in \mathbb{R}^{d\times d}$	$I_d x = x$	$I := I_d{}^{\ddagger}$
	$S_d \in \mathbb{R}^{d\times d}$	$(S_d x)_i = x_{i-1}$	$S := S_d{}^{\ddagger}$
	$R_d \in \mathbb{R}^{d\times d}$	$(R_d x)_i = x_{2-i}$	$R := R_d{}^{\ddagger}$
	$e_i^n \in \mathbb{R}^n$	$e_i^n = I_n e_i$	Usually infer n and write e_i.
$x \in \mathbb{C}^d, k \in [d]$	$\mathrm{circ}_k(x) \in \mathbb{C}^{d\times k}$	$\mathrm{circ}_k(x) = \begin{bmatrix} S^0 x & \cdots & S^{k-1}x \end{bmatrix}$	$\mathrm{circ}(x) = \mathrm{circ}_d(x)$
	$E_{ij}^{mn} \in \mathbb{R}^{m\times n}$	$E_{ij}^{mn} = e_i^m e_j^{n*}$	$E_{ij} := E_{ij}^{mn\ddagger}$
$A \subseteq B$	$\chi_A : B \to \{0,1\}$	$\chi_A(x) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise} \end{cases}$	

Table 1.1: Common operators, objects, and sets, Continued

Parameters	Name and Type	Definition	Comments
	$\mathbb{1}_d \in \mathbb{C}^d$	$(\mathbb{1}_d)_i = 1$ for $i \in [d]$	$\mathbb{1} := \mathbb{1}_d$[‡]
$A \subseteq [d]$	$\mathbb{1}_A^d \in \mathbb{C}^d$	$(\mathbb{1}_A^d)_i = \chi_A(i)$	$\mathbb{1}_A := \mathbb{1}_A^d$[‡]
	$\omega_d \in \mathbb{C}$	$\omega_d = \mathrm{e}^{\frac{2\pi \mathrm{i}}{d}}$	$\omega := \omega_d$[‡]
	$F_d \in \mathbb{C}^{d\times d}$	$(F_d)_{ij} = \frac{1}{\sqrt{d}}\omega_d^{(i-1)(j-1)}$	Note F_d is unitary. $F := F_d$[‡]
	$f_j^d \in \mathbb{C}^d$	$f_j^d = F_d e_j$	$f_j := f_j^d$[‡]
$x, y \in \mathbb{C}^d$	$x \circ y \in \mathbb{C}^d$	$(x \circ y)_i = x_i y_i$	Hadamard/elementwise product
$\ell \in \mathbb{Z}$, $A \in \mathbb{C}^{m\times n}$	$\mathrm{diag}(A, \ell) \in \mathbb{C}^m$	$\mathrm{diag}(A, \ell)_i = A_{i,i+\ell}$	Notation overloaded with $\mathrm{diag}(\cdot)$.
$x \in \mathbb{C}^d$	$\mathrm{diag}(x) \in \mathbb{C}^{d\times d}$	$\mathrm{diag}(x)e_i = x_i e_i$	Also written D_x or $\mathrm{diag}(x_j)_{j=1}^d$.
	$\mathcal{H}^d \subseteq \mathbb{C}^{d\times d}$	$A \in \mathcal{H}^d$ iff $A = A^*$	Hermitian matrices
	$\mathcal{S}^d \subseteq \mathbb{R}^{d\times d}$	$\mathcal{S}^d = \mathbb{R}^{d\times d} \cap \mathcal{H}^d$	Symmetric matrices

We introduce a few more operators that don't fit well into Table 1.1. Given matrices

[‡]We omit the subscript (or superscript) when it is obvious from context.

$V_j \in \mathbb{C}^{m_j \times n_j}$ for $j \in [n]$, we write

$$\operatorname{diag}(V_j)_{j=1}^n = \begin{bmatrix} V_1 & & \\ & \ddots & \\ & & V_n \end{bmatrix} \in \mathbb{C}^{\sum m_j \times \sum n_j}.$$

To conveniently switch between matrices and vectors of different sizes, $\mathcal{R}_d : \bigcup_{k=1}^{\infty} \mathbb{C}^k \to \mathbb{C}^d$ is a resize mapping, where for $v \in \mathbb{C}^k$ and $i \in [d]$,

$$\mathcal{R}_d(v)_i = \begin{cases} v_i, & i \leq k \\ 0, & \text{otherwise} \end{cases} \quad \text{for } i \in [d].$$

Similarly, $\mathcal{R}_{m \times n} : \bigcup_{k_1,k_2}^{\infty} \mathbb{C}^{k_1 \times k_2} \to \mathbb{C}^{m \times n}$ truncates or zero-pads matrices to size $m \times n$.

We let $\operatorname{vec} : \bigcup_{m,n \in \mathbb{N}} \mathbb{C}^{m \times n} \to \bigcup_{k \in \mathbb{N}} \mathbb{C}^k$ be the columnwise vectorization operator, such that for $A \in \mathbb{C}^{m \times n}$, $\operatorname{vec}(A) \in \mathbb{C}^{mn}$ and $\operatorname{vec}(A)_{(j-1)m+i} = A_{ij}$ for $i, j \in [m] \times [n]$. To invert vec, we use $\operatorname{mat}_{(m,n)} : \mathbb{C}^{mn} \to \mathbb{C}^{m \times n}$, such that $\operatorname{mat}_{(m,n)}(v)_{ij} = v_{(j-1)m+i}$. By a slight overloading of notation, by $\operatorname{vec}(a_j)_{j=1}^d$ or $(a_j)_{j=1}^d$, we intend the vector in $v \in \mathbb{R}^d$ satisfying $v_j = a_j$ – this may come in handy to specify something such as $\begin{bmatrix} 1 & 2 & 4 & 8 \end{bmatrix}^T = \operatorname{vec}(2^{j-1})_{j=1}^4$.

Chapter 2

Applications

One common technique used to generate redundancy in phase retrieval-type measurements is to design a system that illuminates only a small part of the sample at a time. These "partial snapshots" are then positioned along an overlapping grid, which produces the redundancy. The overlap is necessary since, even if you could solve phase retrieval perfectly on each patch, each of these patches would have their own phase ambiguities; these would need to be synchronized to achieve a single, coherent image of the original sample. Usually, ptychography is performed by taking a single *mask* or *illumination function* with small supprt, say $\mathbf{m} \in \mathbb{C}^d$ with $\mathrm{supp}(m) \subseteq [\delta]$ where $\delta \ll d$ and shifting this mask to different positions relative to the sample. These measurements may then be modelled as

$$\mathbf{y}_{\ell,j} = |\mathcal{F}(S^{\ell}\mathbf{m} \circ \mathbf{x})_j|^2 + \eta_{\ell,j} = |\langle f_j \circ \mathbf{m}, \mathbf{x}\rangle|^2 + \eta_{\ell,j}. \tag{2.1}$$

This technique is the inspiration for our model, where we do not require our measurements to take this exact form.

Chapter 3

Phase Retrieval From Local Correlation Measurements

Here we consider the phase retrieval problem as modelled in (2.1).

3.1 Introduction

Consider the problem of recovering a vector $\mathbf{x}_0 \in \mathbb{C}^d$ from measurements $\mathbf{y} \in \mathbb{R}^D$ with entries y_j given by

$$y_j = |\langle \mathbf{a}_j, \mathbf{x}_0 \rangle|^2 + \eta_j, \qquad j = 1, \dots, D. \tag{3.1}$$

Here the measurement vectors $\mathbf{a}_j \in \mathbb{C}^d$ are known and the scalars $\eta_j \in \mathbb{R}$ denote noise terms. This problem is known as the *phase retrieval problem* (see, e.g., [73, 98]), as we may think of the $|\cdot|^2$ in (3.1) as erasing the phases of the measurements $\langle \mathbf{a}_j, \mathbf{x}_0 \rangle$ in an otherwise linear system of equations.

The phase retrieval problem arises in many important signal acquisition schemes, including crystallography and ptychography (e.g., [73], see Figure 3.1), diffraction imaging

[45], and optics [73, 98], among many others. Due to the breadth and importance of the applications, there has been significant interest in developing efficient algorithms to solve this problem. Indeed, one of the first algorithms proposed came in the early 1970's with the work of Gerchberg and Saxton [45]. Since then many variations of their method have been proposed (e.g, [11, 12, 35, 39, 90, 92, 91]) and used widely in practice. On the other hand – until recently – there have not been theoretical guarantees concerning the conditions under which these algorithms recover the underlying signal and the extent to which they can tolerate measurement error. Nevertheless, starting in 2006 a growing body of work (e.g., [4, 5, 7, 14, 20, 34, 57, 66]) has emerged, proposing new methods with theoretical performance guarantees under various assumptions on the signal $\mathbf{x}_0$ and the measurement vectors $\mathbf{a}_j$. Unfortunately, the assumptions (especially on the measurement vectors) often do not correspond to the setups used in practice. In particular, the mathematical analysis often requires that the measurement vectors be random or generic (e.g., [4, 7, 20]) while in practice the measurement vectors are a deterministic aspect of the imaging apparatuses employed. A main contribution of this paper is analyzing a construction that more closely matches practicable and deterministic measurement schemes. We propose a two-stage algorithm for solving the phase retrieval problem in this setting and we analyze our method, providing upper bounds on the associated reconstruction error.

In short, we provide theoretical error guarantees for a numerically efficient reconstruction algorithm in a measurement setting that closely resembles measurements used in practice.

3.1.1 Local Correlation Measurements

Consider the case where the vectors $\mathbf{a}_j$ represent shifts of compactly-supported vectors $\mathbf{m}_j, j = 1, \dots, K$ for some $K \in \mathbb{N}$. Using the notation $[n]_k := \{k, \dots, k+n-1\} \subseteq \mathbb{N}$, and defining $[n] := [n]_1$ we take $\mathbf{x}_0, \mathbf{m}_j \in \mathbb{C}^d$ with $\text{supp}(\mathbf{m}_j) \subseteq [\delta] \subseteq [d]$ for some $\delta \in \mathbb{N}$. We also denote the space of Hermitian matrices in $\mathbb{C}^{k\times k}$ by $\mathcal{H}^k$. Now we have measurements of the

form

$$(\mathbf{y}_\ell)_j = |\langle \mathbf{x}_0, S_\ell^* \mathbf{m}_j \rangle|^2, \quad (j,\ell) \in [K] \times P, \tag{3.2}$$

where $P \subseteq [d]_0$ is arbitrary and $S_\ell : \mathbb{C}^d \to \mathbb{C}^d$ is the discrete circular shift operator, namely

$$(S_\ell \mathbf{x}_0)_j = (\mathbf{x}_0)_{\ell+j}.$$

One can see that (3.2) represents the modulus squared of the correlation between $\mathbf{x}_0$ and locally supported measurement vectors. Therefore, we refer to the entries of $\mathbf{y}$ as local correlation measurements. Following [4, 20, 57], the problem may be lifted to a linear system on the space of $\mathbb{C}^{d\times d}$ matrices. In particular, we observe that

$$\begin{aligned}(\mathbf{y}_\ell)_j &= |\langle S_\ell \mathbf{x}_0, \mathbf{m}_j \rangle|^2 = \mathbf{m}_j^* (S_\ell \mathbf{x}_0)(S_\ell \mathbf{x}_0)^* \mathbf{m}_j \\ &= \langle \mathbf{x}_0 \mathbf{x}_0^*, S_\ell^* \mathbf{m}_j \mathbf{m}_j^* S_\ell \rangle,\end{aligned}$$

where the inner product above is the Hilbert-Schmidt inner product. Restricting to the case $P = [d]_0$, for every matrix $A \in \operatorname{span}\{S_\ell^* \mathbf{m}_j \mathbf{m}_j^* S_\ell\}_{\ell,j}$ we have $A_{ij} = 0$ whenever $|i-j| \bmod d \geq \delta$. Therefore, we introduce the family of operators $T_k : \mathbb{C}^{d\times d} \to \mathbb{C}^{d\times d}$ given by

$$T_k(A)_{ij} = \begin{cases} A_{ij}, & |i-j| \bmod d < k \\ 0, & \text{otherwise.} \end{cases} \tag{3.3}$$

Note that T_δ is simply the orthogonal projection operator onto its range $T_\delta(\mathbb{C}^{d\times d}) \supseteq \operatorname{span}\{S_\ell^* \mathbf{m}_j \mathbf{m}_j^* S_\ell\}_{\ell,j}$; therefore,

$$(\mathbf{y}_\ell)_j = \langle \mathbf{x}_0 \mathbf{x}_0^*, S_\ell^* \mathbf{m}_j \mathbf{m}_j^* S_\ell \rangle = \langle T_\delta(\mathbf{x}_0 \mathbf{x}_0^*), S_\ell^* \mathbf{m}_j \mathbf{m}_j^* S_\ell \rangle, \quad (j,\ell) \in [K] \times P. \tag{3.4}$$

For convenience, we set $D := K|P|$ and define the map $\mathcal{A} : \mathbb{C}^{d\times d} \to \mathbb{C}^D$

$$\mathcal{A}(X) = [\langle X, S_\ell^* \mathbf{m}_j \mathbf{m}_j^* S_\ell \rangle]_{(\ell,j)}. \tag{3.5}$$

Sometimes, we consider $\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})}$, the restriction of $\mathcal{A}$ to the domain $T_\delta(\mathbb{C}^{d\times d})$; indeed, if

this linear system is injective on $T_\delta(\mathbb{C}^{d\times d})$, then we can readily solve for

$$T_\delta(\mathbf{x}_0\mathbf{x}_0^*) =: X_0 \tag{3.6}$$

using our measurements $(\mathbf{y}_\ell)_j = (\mathcal{A}(\mathbf{x}_0\mathbf{x}_0^*))_{(\ell,j)}$. In [57], deterministic masks $\mathbf{m}_j$ were constructed for which (3.4) was indeed invertible for certain choices of K and P. An additional construction is given below in Section 3.2.

Improving on [57], we can further see that $\mathbf{x}_0$ can be deduced from X_0 up to a global phase in the noiseless case as follows: First, X_0 immediately gives the magnitudes of the entries of $\mathbf{x}_0$ since $(X_0)_{ii} = |(x_0)_i|^2$. The only challenge remaining, therefore, is to find $\arg((x_0)_i)$ up to a global phase. We proceed by defining $\tilde{\mathbf{x}}_0$ and $\widetilde{X}_0$ by

$$(\tilde{x}_0)_i = \operatorname{sgn}((x_0)_i)$$

$$(\widetilde{X}_0)_{ij} = \begin{cases} \operatorname{sgn}((X_0)_{ij}), & |i-j| \bmod d < \delta \\ 0, & \text{otherwise} \end{cases},$$

where $\operatorname{sgn} : \mathbb{C} \to \mathbb{C}$ is the usual normalization mapping

$$\operatorname{sgn}(z) = \begin{cases} \dfrac{z}{|z|}, & z \neq 0 \\ 1, & \text{otherwise} \end{cases}.$$

We emphasize that

$$\widetilde{X}_0 = \frac{X_0}{|X_0|} = \frac{T_\delta(\mathbf{x}_0\mathbf{x}_0^*)}{|T_\delta(\mathbf{x}_0\mathbf{x}_0^*)|} \text{ and } \widetilde{\mathbf{x}}_0 = \frac{\mathbf{x}_0}{|\mathbf{x}_0|}, \tag{3.7}$$

where the divisions are taken component-wise. Indeed, in [97], it was shown that the phases of the entries of $\mathbf{x}_0$ (up to a global phase) are given by the leading eigenvector of $\widetilde{X}_0$. Moreover, it was shown that this leading eigenvector is unique. Lemma 2 of this paper improves in these results by giving a lower bound on the gap between the top two eigenvalues of $\widetilde{X}_0$. This better understanding of the spectrum of $\widetilde{X}_0$ is then leveraged to analyze the robustness of this eigenvector-based phase retrieval method to measurement noise.

3.1.2 Contributions

In this paper, we analyze a phase retrieval algorithm (Algorithm 1) for estimating a vector $\mathbf{x}_0$ from noisy localized measurements of the form

$$(\mathbf{y}_\ell)_j = |\langle \mathbf{x}_0, S_\ell^* \mathbf{m}_j \rangle|^2 + n_{j\ell}, \quad (j,\ell) \in [2\delta - 1] \times [d]_0. \tag{3.8}$$

This algorithm is composed of two main stages. First, we apply the inverse of the linear operator

$$\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})} : T_\delta(\mathbb{C}^{d\times d}) \to \mathbb{C}^{(2\delta-1)d}$$

defined immediately after (3.5), to obtain a Hermitian estimate X of $T_\delta(\mathbf{x}_0\mathbf{x}_0^*)$ given by

$$X = \Big((\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})})^{-1}\mathbf{y}\Big)/2 + \Big((\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})})^{-1}\mathbf{y}\Big)^*/2 \in T_\delta(\mathbb{C}^{d\times d}). \tag{3.9}$$

In particular, our choice of $\mathbf{m}_j$ as described in Section 3.2 ensures that $\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})}$ is both invertible and well conditioned. Next, once we have an approximation of $T_\delta(\mathbf{x}_0\mathbf{x}_0^*)$, we estimate the magnitudes and phases of the entries of $\mathbf{x}_0$ separately.

For the magnitudes, we simply use the square-roots of the diagonal entries of X. For the phases, we use the normalized eigenvector corresponding to the top eigenvalue of

$$\widetilde{X} := \frac{X}{|X|}, \tag{3.10}$$

where the operations are considered element-wise. The hope is that the leading eigenvector of $\widetilde{X}$ will serve as a good approximation to the leading eigenvector of $\widetilde{X}_0$, which is seen in Section 3.3 (see also [97]) to indeed be a scaled version of the phase vector $\widetilde{\mathbf{x}}_0$ (up to a global phase ambiguity). The entire method is summarized in Algorithm 1, and its associated recovery guarantees are presented in Theorem 1,[1] while its computational complexity is discussed after the theorem in Section 3.1.3. Here and throughout the paper, $\mathrm{e} = 2.71828\ldots$ refers to the base of the natural logarithm and i refers to the imaginary unit.

[1] Remarks about the usefulness of stating the theorem's main inequality in two ways – with respect to $\sigma_{\min}^{-1}$ and κ, SNR – are included in Section 4.2.

Algorithm 1 Fast Phase Retrieval from Local Correlation Measurements

Input: Measurements $\mathbf{y} \in \mathbb{R}^D$ as per (3.8)
Output: $\mathbf{x} \in \mathbb{C}^d$ with $\mathbf{x} \approx \mathbb{e}^{-\mathbb{i}\theta}\mathbf{x}_0$ for some $\theta \in [0, 2\pi]$

1: Compute the Hermitian matrix $X = \left((\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})})^{-1}\mathbf{y}\right)/2 + \left((\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})})^{-1}\mathbf{y}\right)^*/2 \in T_\delta(\mathbb{C}^{d\times d})$ as an estimate of $T_\delta(\mathbf{x}_0\mathbf{x}_0^*)$
2: Form the banded matrix of phases, $\tilde{X} \in T_\delta(\mathbb{C}^{d\times d})$, by normalizing the non-zero entries of X
3: Compute the top eigenvector $u \in \mathbb{C}^d$ of $\tilde{X}$ and set $\widetilde{\mathbf{x}} := \operatorname{sgn}(u)$.
4: Set $x_j = \sqrt{X_{j,j}} \cdot (\tilde{x})_j$ for all $j \in [d]$ to form $\mathbf{x} \in \mathbb{C}^d$

Theorem 1. *Suppose $\delta > 2$ and $d \geq 4\delta$. Let $(x_0)_{\min} := \min_j |(x_0)_j|$ be the smallest magnitude of any entry in $\mathbf{x}_0 \in \mathbb{C}^d$. Then, the estimate $\mathbf{x}$ produced in Algorithm 1 satisfies*

$$\min_{\theta\in[0,2\pi]} \left\|\mathbf{x}_0 - \mathbb{e}^{\mathbb{i}\theta}\mathbf{x}\right\|_2 \leq 24\left(\frac{\|\mathbf{x}_0\|_\infty}{(x_0)_{\min}^2}\right)\frac{d^2}{\delta^{5/2}}\sigma_{\min}^{-1}\|\mathbf{n}\|_2 + d^{\frac{1}{4}}\sqrt{\sigma_{\min}^{-1}\|\mathbf{n}\|_2}$$

$$\min_{\theta\in[0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 \leq 24\left(\frac{\|\mathbf{x}_0\|_\infty}{(x_0)_{\min}^2}\right)\frac{d^2}{\delta^{5/2}}\kappa\frac{\|X_0\|_F}{\mathrm{SNR}} + d^{\frac{1}{4}}\sqrt{\kappa\frac{\|X_0\|_F}{\mathrm{SNR}}}$$

where $\sigma_{\min}$ is the smallest singular value of the system (3.9), $\kappa = \sigma_{\max}/\sigma_{\min}$ is its condition number, $X_0 = T_\delta(\mathbf{x}_0\mathbf{x}_0^*)$, and $\mathrm{SNR} = \|\mathcal{A}(X_0)\|_2/\|\mathbf{n}\|_2$ is the signal to noise ratio.

Theorem 1, which deterministically depends on both the masks and the signal, provides improvements over the first deterministic theoretical robust recovery guarantees proven in [57] for a wide class of non-vanishing signals. Momentarily consider, e.g., the class of "flat" vectors $\mathbf{x}_0 \in \mathbb{C}^d$ for which both (i) $(x_0)_{\min} \geq \frac{\|\mathbf{x}_0\|_2}{2\sqrt{d}}$, and (ii) $\left(\frac{\|\mathbf{x}_0\|_\infty}{(x_0)_{\min}^2}\right) \leq \tilde{C}$ for some absolute constant $\tilde{C} \in \mathbb{R}^+$, hold. The main deterministic result of [57] also applies to this class of vectors and states that an algorithm exists which can achieve the following robust recovery guarantee.

Theorem 2 (See Theorem 5 in [57])**.** *There exist fixed universal constants $C, C' \in \mathbb{R}^+$ such that the following holds for all $\mathbf{x}_0 \in \mathbb{C}^d$ of the class mentioned above: Let $\|\mathbf{x}_0\|_2 \geq Cd\sqrt{(\delta-1)\|\mathbf{n}\|_2}$. Then, the algorithm in [57], when provided with noisy measurements of $\mathbf{x}_0 \in \mathbb{C}^d$ (3.8) resulting from the masks discussed in Example 1 of Section 3.2, will output a*

vector $\mathbf{x} \in \mathbb{C}^d$ *satisfying*

$$\min_{\theta \in [0,2\pi]} \left\| \mathbf{x}_0 - \mathbb{e}^{\mathbb{i}\theta} \mathbf{x} \right\|_2 \leq C' d \sqrt{(\delta - 1) \left\| \mathbf{n} \right\|_2}.$$

Comparing the error bounds provided by Theorems 1 and 2 for the class of flat vectors $\mathbf{x}_0$ mentioned above when using measurements resulting from the masks discussed in Example 1 of Section 3.2,[2] we can see that Theorem 1 makes the following improvements over Theorem 2:

- Theorem 1 improves on the error bound of Theorem 2 for arbitrary small-norm noise $\mathbf{n}$ having $\|\mathbf{n}\|_2 = \mathcal{O}\left(\delta/d^2\right)$.
- Theorem 2's error bound breaks down entirely for noise $\mathbf{n}$ with ℓ^2-norm on the order of $\|\mathbf{n}\|_2 = \Theta\left(\frac{\|\mathbf{x}_0\|_2^2}{\delta d^2}\right)$. Theorem 1's error bound, on the other hand, still provides non-trivial error guarantees for such noise levels as long as $\|\mathbf{x}_0\|_2 = \mathcal{O}\left(\delta\right)$.[3]

In addition, Theorem 1 also applies to a more general set of masks and a larger class of signals $\mathbf{x}_0$ than Theorem 2 does. And, perhaps most importantly, Algorithm 1 generally outperforms the algorithm referred to by Theorem 2 numerically for all noise levels (see, e.g., Figures 3.2a and 3.6a in Section 3.6). Theorem 1 provides theoretical error guarantees for this numerically improved method.

We note that the $\mathcal{O}\left(\left(\frac{d}{\delta}\right)^2\right)$-factor in the first term of the error bound provided by Theorem 1 is probably suboptimal, especially in practice. Indeed, Theorem 1 provides a worst-case error guarantee that holds for any arbitrary (including worst-case/adversarial) perturbation $\mathbf{n}$ of the measurements (3.8). However, while the quadratic dependence on d/δ is probably suboptimal, *some* dependence on d/δ likely exists for worst-case additive-noise in our local measurement setting. There is, e.g., numerical evidence that the empirical noise

[2]Note that the inverse singular value $\sigma_{\min}^{-1}$ mentioned in Theorem 1 for the masks discussed in Example 1 of Section 3.2 is $\mathcal{O}(\delta)$, and $\kappa = \mathcal{O}(\delta^2)$. See Theorem 3 below for a more exact statement. All asymptotic notation is with respect to $d \to \infty$. Below δ is always assumed to be independent of (and less than) d unless otherwise noted.

[3]This allows Theorem 1 to cover, e.g., the case of larger $\delta = \Omega\left(\sqrt{d}\right)$.

robustness of many phase retrieval methods deteriorates as d grows for local measurements whose support size δ is held fixed. We will leave a rigorous theoretical investigation of the optimal scaling of such noise robustness guarantees with d/δ to future work. For the time being we will simply note here that the $\mathcal{O}\left(\left(\frac{d}{\delta}\right)^2\right)$-factor is mainly a product of the relatively small eigenvalue gap of the matrix $\widetilde{X}_0$ defined in (3.7) above. See Section 3.3 below for more details.

3.1.3 The Runtime Complexity of Algorithm 1

Consider now the computational complexity of Algorithm 1 (assuming, of course, that $\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})}$ is actually invertible). One can see that line 1 can always be done in at most $\mathcal{O}(d\cdot\delta^3+\delta\cdot d\log d)$ flops using a block circulant matrix factorization approach (see Section 3.1 in [57]). In certain cases one can improve on this; for example, the second (new) mask construction of Section 3.2 allows line 1 to be performed in only $\mathcal{O}(d\cdot\delta)$ flops. Even in the worst case, however, if one precomputes this block circulant matrix factorization in advance given the masks $\mathbf{m}_j$ then line 1 can always be done in $\mathcal{O}(d\cdot\delta^2+\delta\cdot d\log d)$ flops thereafter.

The top eigenvector $\tilde{\mathbf{x}}$ of $\tilde{X}$ is guaranteed to be found in line 3 of Algorithm 1 in the low-noise (e.g., noiseless) setting via the shifted inverse power method with shift $\mu := 2\delta - 1$ and initial vector $\mathbf{e}_1$ (the first standard basis vector). More generally, one may utilize the Rayleigh quotient iteration with the initial eigenvalue estimate fixed to $2\delta-1$ for the first few iterations. In either case, each iteration can be accomplished with $\mathcal{O}(d\cdot\delta^2)$ flops due to the banded structure of $\tilde{X}$ (see, e.g., [94]). In the low-noise setting the top eigenvector $\tilde{\mathbf{x}}$ can be computed to machine precision in $\mathcal{O}(\log d)$ such iterations,[4] for a total flop count of

[4]To see why $\mathcal{O}(\log d)$ iterations suffice one can appeal to lemmas 1 and 2 below. Let $|\lambda_1|>|\lambda_2|\geq\cdots\geq|\lambda_d|$ be the eigenvalues of $\tilde{X}$ with associated orthonormal eigenvectors $\mathbf{u}_j\in\mathbb{C}^d$. Let $\delta:=|\lambda_1|-|\lambda_2|>0$. When the noise level is sufficiently low (so that $\tilde{X}\approx\tilde{X}_0$) one will have both (i) $|\mathbf{e}_1^*\mathbf{u}_j|=\Theta(1/\sqrt{d})\ \forall j\in[d]$, and (ii) $\mu\in(\lambda_1-\delta/4,\lambda_1+\delta/4)$ be true. Thus, we will have that there exists some unit norm $\mathbf{r}\in\mathbb{C}^d$ such that

$$\frac{\left(\tilde{X}-\mu I\right)^{-k}\mathbf{e}_1}{\left\|\left(\tilde{X}-\mu I\right)^{-k}\mathbf{e}_1\right\|_2}=\frac{\mathbf{u}_1+\sum_{j=2}^d\mathcal{O}\left(\left|\frac{\lambda_1-\mu}{\lambda_j-\mu}\right|^k\right)\mathbf{u}_j}{1+\mathcal{O}\left(\frac{d}{9^k}\right)}=\mathbf{u}_1+\mathcal{O}\left(\frac{d}{3^k}\right)\mathbf{r}$$

holds for any given integer $k=\Omega\left(\log_3 d\right)$.

$\mathcal{O}(\delta^2 \cdot d \log d)$ for line 3 in that case. In total, then, one can see that Algorithm 1 will always require just $\mathcal{O}(\delta^2 \cdot d \log d + d \cdot \delta^3)$ total flops in low-noise settings. Furthermore, in all such settings a measurement mask support of size $\delta = \mathcal{O}(\log d)$ appears to suffice.

3.1.4 Connection to Ptychography

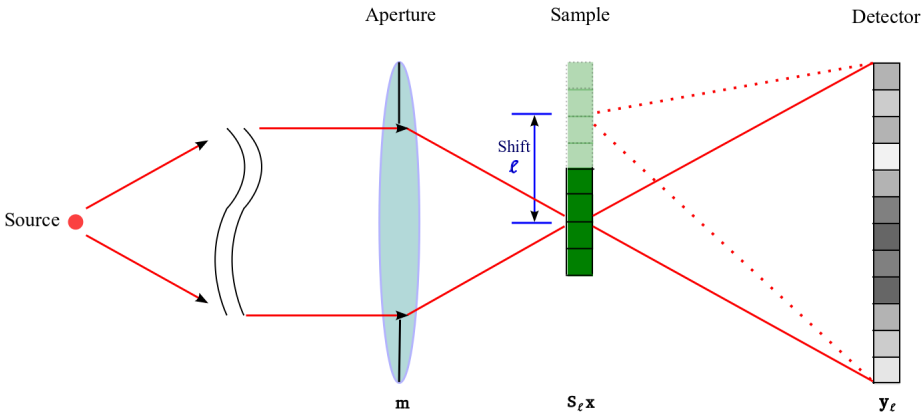


Figure 3.1: Illustration of one-dimensional ptychographic imaging (Adapted from "Fly-scan ptychography", Huang et al., Scientific Reports 5 (9074), 2015.)

In ptychographic imaging (see Fig. 3.1), small regions of a specimen are illuminated one at a time and an intensity[5] detector captures each of the resulting diffraction patterns. Thus each of the ptychographic measurements is a local measurement, which under certain assumptions (e.g., appropriate wavelength of incident radiation, far-field Fraunhofer approximation), can be modeled as [29, 46]

$$y(t, \omega) = \left| \mathcal{F}[\widetilde{h} \cdot S_t f](\omega) \right|^2 + \eta(t, \omega). \tag{3.11}$$

Here, $\mathcal{F}$ denotes the Fourier transform, $f : [0, 1] \to \mathbb{C}$ represents the unknown test specimen, S_t is the shift operator defined via

$$(S_t f)(s) := f(s + t),$$

and $\widetilde{h} : [0, 1] \to \mathbb{C}$ is the so-called illumination function [105] of the imaging system. To

[5] By intensity, we mean magnitude squared.

account for the local nature of the measurements in (3.11), we assume that $\mathrm{supp}(\widetilde{h}) \subseteq \mathrm{supp}(f)$.

As the phase retrieval problem is inherently non-linear and requires sophisticated computer algorithms to solve, consider the discrete version of (3.11), with $\widetilde{\mathbf{m}}, \mathbf{x}_0 \in \mathbb{C}^d$ discretizing $\widetilde{h}$ and f. Thus (3.11), in the absence of noise, becomes

$$(\mathbf{y}_\ell)_j = \left| \sum_{n=1}^{d} \widetilde{m}_n \, (x_0)_{n+\ell} \, \mathrm{e}^{-\frac{2\pi \mathrm{i}(j-1)(n-1)}{d}} \right|^2, \quad (j,\ell) \in [d] \times [d]_0, \tag{3.12}$$

where indexing is considered modulo-d, so $(\mathbf{y}_\ell)_j$ is a diffraction measurement corresponding to the j^{th} Fourier mode of a circular ℓ-shift of the specimen. We use circular shifts for convenience and we remark that this is appropriate as one can zero-pad $\mathbf{x}_0$ and $\widetilde{\mathbf{m}}$ in (3.12) and obtain the same $(\mathbf{y}_\ell)_j$ as one would with non-circular shifts. In practice, one may not need to use all the shifts $\ell \in [d]_0$ as a subset may suffice. Defining $\mathbf{m}_j \in \mathbb{C}^d$ by

$$(\mathbf{m}_j)_n = \overline{\widetilde{m}_n} \, \mathrm{e}^{\frac{2\pi \mathrm{i}(j-1)(n-1)}{d}} \tag{3.13}$$

and rearranging (3.12), we obtain

$$\begin{aligned} (\mathbf{y}_\ell)_j &= \left| \sum_{n=1}^{d} (x_0)_{n+\ell} \, \overline{(\mathbf{m}_j)_n} \right|^2 = \left| \sum_{n=1}^{\delta} (x_0)_{n+\ell} \, \overline{(\mathbf{m}_j)_n} \right|^2 \\ &= |\langle S_\ell \mathbf{x}_0, \mathbf{m}_j \rangle|^2 = \langle S_\ell \mathbf{x}_0 \mathbf{x}_0^* S_\ell^*, \mathbf{m}_j \mathbf{m}_j^* \rangle \\ &= \langle T_\delta(\mathbf{x}_0 \mathbf{x}_0^*), S_\ell^* \mathbf{m}_j \mathbf{m}_j^* S_\ell \rangle, \quad (j,\ell) \in [d] \times [d]_0 \end{aligned} \tag{3.14}$$

where the second and last equalities follow from the fact that $\widetilde{\mathbf{m}}$ (and hence each $\mathbf{m}_j$) is locally supported. We note that (3.14) defines a correlation with local masks or window functions $\mathbf{m}_j$. More importantly, (3.14) shows that ptychography (with ℓ ranging over any subset of $[d]_0$) represents a case of the general system seen in (3.4).

3.1.5 Connections to Masked Fourier Measurements

Often, in imaging applications involving phase retrieval, a mask is placed either between the illumination source and the sample or between the sample and the sensor. Here, we will see that the mathematical setup that we consider is applicable in this scenario, albeit when the masks are band-limited. As before, let $\mathbf{x}_0, \mathbf{m} \in \mathbb{C}^d$ denote the unknown signal of interest, and a known mask (or window), respectively. Moreover, for a vector $\mathbf{x}_0 \in \mathbb{C}^d$ we denote its discrete Fourier transform $\widehat{\mathbf{x}_0} \in \mathbb{C}^d$ by

$$(\widehat{x_0})_k := \sum_{n=1}^{d} (x_0)_n e^{-2\pi \mathbb{i}(n-1)(k-1)/d}.$$

Here, we consider squared magnitude *windowed Fourier transform* measurements of the form

$$(\mathbf{y}_\ell)_k = \left| \sum_{n=1}^{d} (x_0)_n \, m_{n-\ell} \, \mathbb{e}^{-\frac{2\pi \mathbb{i}(k-1)(n-1)}{d}} \right|^2, \; k \in [d], \; \ell \in \{\ell_1, \ldots, \ell_L\} \subseteq [d]_0. \tag{3.15}$$

As before, ℓ denotes a shift or translation of the mask/window, so $(\mathbf{y}_\ell)_k$ corresponds to the (squared magnitude of) the k^{th} Fourier mode associated with an ℓ-shift[6] of the mask $\mathbf{m}$. Defining the modulation operator, $W_k : \mathbb{C}^d \mapsto \mathbb{C}^d$, by its action $(W_k \mathbf{x}_0)_n = e^{2\pi \mathbb{i}(k-1)(n-1)/d} \, (x_0)_n$ and applying elementary Fourier transform properties[7] one has

$$\begin{aligned}
(\mathbf{y}_\ell)_k &= |\langle \mathbf{x}_0, S_{-\ell}(e^{2\pi \mathbb{i}(k-1)\ell/d} W_k \overline{\mathbf{m}}) \rangle|^2 \\
&= |\langle \mathbf{x}_0, S_{-\ell}(W_k \overline{\mathbf{m}}) \rangle|^2 = |\langle \widehat{\mathbf{x}_0}, \widehat{S_{-\ell}(W_k \overline{\mathbf{m}})} \rangle|^2 \\
&= |\langle \widehat{\mathbf{x}_0}, W_{-\ell+1}(S_{-k+1} \widehat{\overline{\mathbf{m}}}) \rangle|^2 \\
&= |\langle \widehat{\mathbf{x}_0}, S_{-k+1}(W_{-\ell+1} \widehat{\overline{\mathbf{m}}}) \rangle|^2.
\end{aligned} \tag{3.16}$$

Defining $\widehat{\mathbf{m}}_\ell := W_{-\ell+1} \widehat{\overline{\mathbf{m}}}$ and assuming that $\mathrm{supp}(\widehat{\overline{\mathbf{m}}}) \subseteq [\delta]$ (e.g., assuming that $\mathbf{m}$ is

[6] As above, all indexing and shifts are considered modulo-d.

[7] $\widehat{S_\ell \mathbf{x}_0} = W_{\ell+1} \widehat{\mathbf{x}_0}$, $\widehat{W_k \mathbf{x}_0} = S_{-k+1} \widehat{\mathbf{x}_0}$, and $W_k S_\ell \mathbf{x}_0 = e^{-2\pi \mathbb{i}(k-1)\ell/d} S_\ell W_k \mathbf{x}_0$.

real-valued and band-limited), we now have that

$$\begin{aligned}(\mathbf{y}_\ell)_k &= \langle \widehat{\mathbf{x}}_0 \widehat{\mathbf{x}}_0^*, S_{-k+1} \widehat{\mathbf{m}}_\ell \widehat{\mathbf{m}}_\ell^* S_{-k+1}^* \rangle \\ &= \langle T_\delta(\widehat{\mathbf{x}}_0 \widehat{\mathbf{x}}_0^*), S_{-k+1} \widehat{\mathbf{m}}_\ell \widehat{\mathbf{m}}_\ell^* S_{-k+1}^* \rangle,\end{aligned}$$

which again represents a case of the general system seen in (3.4). Moreover, our results all hold for this setting, albeit with the Fourier transforms of signals and conjugated masks.

3.1.6 Related Work

The first approach to the phase retrieval problem was proposed in the 1970's in [45] by Gerchberg and Saxton, where the measurement data corresponded to knowing the magnitude of both the image $\mathbf{x}_0$ and its Fourier transform. This result was famously expanded upon by Fienup [39] later that decade, one significant improvement being that only the magnitude of the Fourier transform of $\mathbf{x}_0$ must be known in the case of a signal $\mathbf{x}_0$ belonging to some fixed convex set $\mathcal{C}$ (typically, $\mathcal{C}$ is the set of non-negative, real-valued signals restricted to a known domain). Though these techniques work well in practice and have been popular for decades, they are notoriously difficult to analyze (see, e.g., [11, 12, 35, 90, 92, 91]). These are iterative methods that work by improving an initial guess until they stagnate. Recently Marchesini et al. proved that alternating projection schemes using generic measurements are guaranteed to converge to the correct solution *if provided with a sufficiently accurate initial guess* and algorithms for ptychography were explored in particular [70]. However, no global recovery guarantees currently exist for alternating projection techniques using local measurements (i.e., finding a sufficiently accurate initial guess is not generally easy).

Other authors have taken to proving probabilistic recovery guarantees when provided with globally supported Gaussian measurements. Methods for which such results exist vary in their approach, and include convex relaxations [19, 20], gradient descent strategies [22], graph-theoretic [2] and frame-based approaches [6, 14], and variants on the alternating minimization (e.g., with resampling) [75].

Several recovery algorithms achieve theoretical recovery guarantees while using at most $D = \mathcal{O}(d \log^4 d)$ masked Fourier coded diffraction pattern measurements, including both *PhaseLift* [21, 51], and *Wirtinger Flow* [22]. However, these measurements are both randomized (which is crucial to the probabilistic recovery guarantees developed for both PhaseLift and Wirtinger Flow – deterministic recovery guarantees do not exist for either method in the noisy setting), and provide global information about $\mathbf{x}_0$ from each measurement (i.e., the measurements are not locally supported).

Among the first treatments of local measurements are [13, 33] and [59], in which it is shown that STFT measurements with specific properties can allow (sparse) phase retrieval in the noiseless setting, and several recovery methods are proposed. Similarly, the phase retrieval approach from [2] was extended to STFT measurements in [81] in order to produce recovery guarantees in the noiseless setting. More recently, randomized robustness guarantees were developed for time-frequency measurements in [77]. However, no *deterministic* robust recovery guarantees have been proven in the noisy setting for any of these approaches. Furthermore, none of the algorithms developed in these papers are empirically demonstrated to be competitive numerically with standard alternating projection techniques for large signals when utilizing windowed Fourier and/or correlation-based measurements. In [57], the authors propose the measurement scheme developed in the current paper and prove the first deterministic robustness results for a different greedy recovery algorithm.

3.1.7 Organization

Section 3.2 discusses two collections of local correlation masks $\mathbf{m}_j$, one of which is novel and the other of which was originally studied in [57]. Most importantly, Section 3.2 shows that the recovery of $T_\delta(\mathbf{x}_0\mathbf{x}_0^*)$ from measurements associated with the proposed masks can be done stably in the presence of measurement noise. Moreover, since in the noisy regime, the leading eigenvector $\widetilde{\mathbf{x}}$ of $\widetilde{X}$ (associated with line 3 of Algorithm 1) will no longer correspond exactly to the true phases $\widetilde{\mathbf{x}}_0$, we are interested in a perturbation theory for the eigenvectors of $\widetilde{X}_0$. Intuitively, $\widetilde{\mathbf{x}}$ will be most accurate when the eigenvalue of $\widetilde{X}_0$ associated

with $\widetilde{\mathbf{x}}_0$ is well separated from the rest of the eigenvalues and so, accordingly, Section 3.3 studies the spectrum of $\widetilde{X}_0$. Indeed, this eigenvalue is rigorously shown to control the stability of the top eigenvector of $\widetilde{X}_0$ with respect to noise, and Section 3.4 develops perturbation results concerning their top eigenvectors by adapting the spectral graph techniques used in [2]. Recovery guarantees for the proposed phase retrieval method are then compiled in Section 3.5. Numerical results demonstrating the accuracy, efficiency, and robustness of the proposed methods are finally provided in Section 3.6[8], while Section 3.7 contains some concluding remarks and avenues for further research. In Appendix 3.8, we provide an alternate, weaker but easier to derive eigenvector perturbation result analogous to the one in Section 3.4 which may be of independent interest.

3.2 Well-conditioned measurement maps

Here, we present two example constructions for which the linear operator $\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})}$ used in Step 1 of Algorithm 1 is well conditioned. Such constructions are crucial for the stability of the method to additive noise.

Example 1:

In [57], a construction was proposed for the masks $\mathbf{m}_\ell$ in (3.2) that guarantees the stable invertibility of $\mathcal{A}$. This construction comprises windowed Fourier measurements with parameters $\delta \in \mathbb{Z}^+$ and $a \in [4,\infty)$ corresponding to the $2\delta-1$ masks $\mathbf{m}_j \in \mathbb{C}^d$, $j = 1, ..., 2\delta-1$ with entries given by

$$(\mathbf{m}_j)_n = \begin{cases} \frac{\mathrm{e}^{-n/a}}{\sqrt[4]{2\delta-1}} \cdot \mathrm{e}^{\frac{2\pi \mathrm{i}\cdot(n-1)\cdot(j-1)}{2\delta-1}} & \text{if } n \le \delta \\ 0 & \text{if } n > \delta \end{cases}. \tag{3.17}$$

[8]MATLAB code to run the BlockPR algorithm is available online at [58].

Here, measurements using all shifts $\ell = 1, ..., d$ of each mask are taken. In the notation of (3.4), this corresponds to $K = 2\delta - 1$ and $P = [d]_0$, which yields $D = (2\delta - 1)d$ total measurements. By considering the basis $\{E_{ij}\}$ for $T_\delta(\mathbb{C}^{d\times d})$ given by

$$E_{i,j}(s,t) = \begin{cases} 1, & (i,j) = (s,t) \\ 0, & \text{otherwise} \end{cases}$$

it was shown in [57] that this system is both well conditioned and rapidly invertible. In particular, if M' is the matrix representing the measurement mapping $\mathcal{A} : T_\delta(\mathbb{C}^{d\times d}) \to T_\delta(\mathbb{C}^{d\times d})$ with respect to the basis $\{E_{ij}\}$, the following estimates of the condition number and cost of inversion hold.

Theorem 3 ([57]). *Consider measurements of the form* (3.17) *with* $a := \max\left\{4, \ \frac{\delta-1}{2}\right\}$. *Let* $M' \in \mathbb{C}^{D\times D}$ *be the matrix representing the measurement mapping* $\mathcal{A} : T_\delta(\mathbb{C}^{d\times d}) \to T_\delta(\mathbb{C}^{d\times d})$ *with respect to the basis* $\{E_{ij}\}$. *Then, the condition number of* M' *satisfies*

$$\kappa\left(M'\right) \ < \ \max\left\{144\mathrm{e}^2, \ \frac{9\mathrm{e}^2}{4}\cdot(\delta-1)^2\right\},$$

and the smallest singular value of M' *satisfies*

$$\sigma_{\min}\left(M'\right) > \frac{7}{20a}\cdot \mathrm{e}^{-(\delta+1)/a} > \frac{C}{\delta}$$

for an absolute constant $C \in \mathbb{R}^+$. *Furthermore,* M' *can be inverted in* $\mathcal{O}\left(\delta \cdot d \log d\right)$*-time.*

This theorem indicates that one can both efficiently and stably solve for $\mathbf{x}_0\mathbf{x}_0^*$ using (3.4) with the measurements given in (3.17). This measurement scheme is also interesting because it corresponds to a ptychography system if we take the illumination function (i.e. the physical mask) in (3.12) to be $\widetilde{m}_n = \frac{\mathrm{e}^{-n/a}}{\sqrt[4]{2\delta-1}}$ and assume that $d = k(2\delta-1)$ for some $k \in \mathbb{N}$; in practice, this may be achieved by zero-padding the specimen. Then we may take the subset of the measurements (3.13) given by $j = (p-1)k + 1, \ p \in [2\delta - 1]$ to obtain the masks specified in (3.17). We also remark that in this setup, only one physical mask is required, as the index j in (3.17) denotes the different frequencies observed in the Fourier domain at

the sensor array.

Example 2:

We provide a second deterministic construction that improves on the condition number of the previous collection of measurement vectors. We merely set

$$\mathbf{m}_j = \begin{cases} e_1, & j = 1 \\ e_1 + e_{p+1}, & j = 2p, p \in [\delta - 1] \\ e_1 + \mathrm{i}e_{p+1}, & j = 2p+1, p \in [\delta - 1] \end{cases} \tag{3.18}$$

A simple induction on k shows that $\{S_\ell \mathbf{m}_j \mathbf{m}_j^* S_\ell^*\}_{\ell \in [d]_0, j \in [2k-1]}$ is a basis for $T_k(\mathbb{C}^{d\times d})$, so if we take $\mathbf{m}_1, \ldots, \mathbf{m}_{2\delta-1}$ for our masks we'll have a basis for $T_\delta(\mathbb{C}^{d\times d})$. Indeed, if we let

$$\mathcal{B} : T_k(\mathbb{C}^{d\times d}) \to \mathbb{C}^{\delta \times d}$$

be the measurement operator defined via

$$\big(\mathcal{B}(X)\big)_{\ell,j} = \langle S_\ell \mathbf{m}_j \mathbf{m}_j^* S_\ell^*, X \rangle, \quad (\ell, j) \in [d]_0 \times [2k-1]$$

we can immediately solve for the entries of $X \in T_k(\mathcal{H}^{d\times d})$ from $\mathcal{B}(X) =: B$ by observing that

$$\begin{aligned} X_{i,i} &= B_{i-1,1} \\ X_{i,i+k} &= \tfrac{1}{2} B_{i-1,2k} + \tfrac{i}{2} B_{i-1,2k+1} - \tfrac{1+i}{2}(B_{i-1,1} + B_{i+k-1,1}), \end{aligned}$$

where we naturally take the indices of B mod d. This leads to an upper triangular system if we enumerate X by its diagonals; namely we regard $T_\delta(\mathcal{H}^{d\times d})$ as a $d(2\delta - 1)$ dimensional

vector space over $\mathbb{R}$ and set, for $i \in [d]$

$$z_{kd+i} = \begin{cases} \mathrm{Re}(X_{i,i+k}), & 0 \le k < \delta \\ \mathrm{Im}(X_{i,i+k-\delta+1}), & \delta \le k < 2\delta - 1 \end{cases},$$

$$y_{kd+i} = \begin{cases} \mathcal{B}(X)_{i,1}, & k = 0 \\ \mathcal{B}(X)_{i,2k}, & 1 \le k < \delta \\ \mathcal{B}(X)_{i,2(k-\delta+1)+1}, & \delta \le k < 2\delta - 1 \end{cases}$$

Then with $S = S_1 \in \mathbb{R}^{d\times d}$ representing the circular shift operator as before, we have

$$y = \begin{bmatrix} I_d & 0 & 0 \\ D & 2I_{d(\delta-1)} & 0 \\ D & 0 & 2I_{d(\delta-1)} \end{bmatrix} z =: Cz, \text{ where } D = \begin{bmatrix} I_d + S \\ I_d + S^2 \\ \vdots \\ I_d + S^{\delta-1} \end{bmatrix}.$$

Since the matrix C is upper triangular, its inverse is immediate:

$$C^{-1} = \begin{bmatrix} I_d & 0 & 0 \\ -D/2 & I_{d(\delta-1)}/2 & 0 \\ -D/2 & 0 & I_{d(\delta-1)}/2 \end{bmatrix}.$$

To ascertain the condition number of $\mathcal{B}$, then, all we need is the extremal singular values of C. We bound the top singular value by considering

$$\begin{aligned} \sigma_{\max}(C) = \max_{||w||^2+||v||^2=1} \left\| C \begin{bmatrix} w \\ v \end{bmatrix} \right\| &= \left\| \begin{bmatrix} w \\ Dw \\ Dw \end{bmatrix} + 2 \begin{bmatrix} 0 \\ v \end{bmatrix} \right\| \\ &\le \sqrt{||w||^2 + 2||w + Sw||^2 + \cdots + 2||w + S^{\delta-1}w||^2} + ||2v|| \\ &\le \sqrt{8(\delta-1)+1}||w|| + 2||v|| \le \sqrt{8(\delta-1)+5} \le 2\sqrt{2\delta}, \end{aligned}$$

where in the last line we have used $||w||^2 + ||v||^2 = 1$. By a nearly identical argument, we find

$$\frac{1}{\sigma_{\min}(C)} = \sigma_{\max}(C^{-1}) \leq \sqrt{2\delta}$$

so that the condition number is bounded by $\kappa(C) \leq 4\delta$. We collect these results in Proposition 1.

Proposition 1. *Fix $d, \delta \in \mathbb{N}$ with $2\delta - 1 \leq d$ and let $\mathbf{m}_j$ be as in (3.18). Then the condition number of $\mathcal{A}$ satisfies $\kappa \leq 4\delta$ and its minimum singular value satisfies $\sigma_{\min}^{-1} \leq \sqrt{2\delta}$.*

3.3 The Spectrum of $\tilde{X}_0$

Consider line 3 of Algorithm 1, which shows that we are trying to recover $\widetilde{\mathbf{x}}_0 := \frac{\mathbf{x}_0}{|\mathbf{x}_0|}$ via an eigenvector method. Here, we show that $\widetilde{X}_0$ has $\widetilde{\mathbf{x}}_0$ as its top eigenvector and we investigate the spectral properties of $\widetilde{X}_0$ in this section, following the intuition that the eigenvalue gap $|\lambda_1 - \lambda_2|$ will affect the robustness of the spectral step in the algorithm.

For the remainder of the paper, we let $\mathbb{1}$ refer to a constant vector of all ones; its size will always be determined by context. To begin, consider $U = T_\delta(\mathbb{1}\mathbb{1}^*)$, i.e.,

$$U_{j,k} = \begin{cases} 1 & \text{if } |j-k| \bmod d < \delta \\ 0 & \text{otherwise} \end{cases}. \tag{3.19}$$

Observe that U is circulant for all δ, so its eigenvectors are always discrete Fourier vectors. Setting $\omega_j = \mathrm{e}^{2\pi \mathrm{i} \frac{j-1}{d}}$ for $j = 1, 2, \ldots, d$, one can also see that the eigenvalues of U are given by

$$\nu_j = \sum_{k=1}^{d} (U)_{1,k} \omega_j^{k-1} = 1 + \sum_{k=1}^{\delta-1} \omega_j^k + \omega_j^{-k} \;=\; 1 + 2\sum_{k=1}^{\delta-1} \cos\left(\frac{2\pi(j-1)k}{d}\right), \tag{3.20}$$

for all $j = 1, \ldots, d$. In particular, $\nu_1 = 2\delta - 1$. Set $\Lambda = \mathrm{diag}\{\nu_1, \ldots, \nu_d\}$ and let F denote

the unitary $d \times d$ discrete Fourier matrix with entries

$$F_{j,k} := \frac{1}{\sqrt{d}} \mathrm{e}^{2\pi \mathrm{i} \frac{(j-1)(k-1)}{d}},$$

then $U = F\Lambda F^*$.

We consider that $\widetilde{X}_0$ and U are similar; indeed $\widetilde{X}_0 = \widetilde{D}_0 U \widetilde{D}_0^*$, where $\widetilde{D}_0 = \mathrm{diag}\{(\widetilde{x}_0)_1, \ldots, (\widetilde{x}_0)_d\}$. Since $|(\widetilde{\mathbf{x}}_0)_j| = 1$ for each j, we have that $\widetilde{D}_0$ is unitary. Thus the eigenvalues of $\widetilde{X}_0$ are given by (3.20), and its eigenvectors are simply the discrete Fourier vectors modulated by the entries of $\widetilde{\mathbf{x}}_0$. We now have the following lemma.

Lemma 1. *Let $\widetilde{X}_0$ be defined as in* (3.7)*. Then*

$$\widetilde{X}_0 = \widetilde{D}_0 F \Lambda F^* \widetilde{D}_0^*$$

where F is the unitary $d \times d$ discrete Fourier transform matrix, $\widetilde{D}_0$ is the $d \times d$ diagonal matrix $\mathrm{diag}\{(\widetilde{x}_0)_1, \ldots, (\widetilde{x}_0)_d\}$*, and Λ is the $d \times d$ diagonal matrix* $\mathrm{diag}\{\nu_1, \ldots, \nu_d\}$ *where*

$$\nu_j := 1 + 2 \sum_{k=1}^{\delta-1} \cos\left(\frac{2\pi(j-1)k}{d}\right)$$

for $j = 1, \ldots, d$.

We next estimate the principal eigenvalue gap of $\widetilde{X}_0$. This information will be crucial to our understanding of the stability and robustness of Algorithm 1.

3.3.1 The Spectral Gap of $\widetilde{X}_0$

Set $\theta_j = \frac{2\pi j}{d}$ and begin by observing that (by, for example, Lemma 27), for any $\theta \in \mathbb{R}$,

$$\sum_{k=1}^{\delta-1} \cos(\theta k) \;=\; \frac{1}{2}\left(\frac{\sin(\theta(\delta - 1/2))}{\sin(\theta/2)} - 1\right).$$

Accordingly, defining $l_\delta : \mathbb{R} \to \mathbb{R}$ by $l_\delta(\theta) := 1 + 2\sum_{k=1}^{\delta-1} \cos(\theta k)$ we have that

$$\nu_{j+1} \;=\; l_\delta(\theta_j) \;=\; \frac{\sin(\theta_j(\delta - 1/2))}{\sin(\theta_j/2)}. \tag{3.21}$$

Thus, the eigenvalues of $\widetilde{X}_0$ are sampled from the $(\delta-1)^{\text{st}}$ Dirichlet kernel. Of course, $\nu_1 = 2\delta - 1$ is the largest of these in magnitude, so the eigenvalue gap $\min_j \nu_1 - |\nu_j|$ is at most equal to

$$\begin{aligned}
\nu_1 - \nu_2 &= (2\delta - 1) - \frac{\sin(\pi/d(2\delta-1))}{\sin(\pi/d)} \\
&\leq (2\delta - 1) - \frac{\pi/d(2\delta-1) - \frac{1}{6}(\pi/d(2\delta-1))^3}{\pi/d} \\
&= \frac{1}{6}\left(\frac{\pi}{d}\right)^2 (2\delta-1)^3 \leq \frac{4\pi^2}{3}\frac{\delta^3}{d^2}.
\end{aligned}$$

Thus, $\nu_1 - |\nu_2| \lesssim \frac{\delta^3}{d^2}$. However, a lower bound on the spectral gap is more useful. The following lemma establishes that the spectral gap is indeed $\sim \frac{\delta^3}{d^2}$ for most reasonable choices of $\delta < d$.

Lemma 2. *Let $\nu_1 = 2\delta - 1, \nu_2, \ldots, \nu_d$ be the eigenvalues of $\widetilde{X}_0$. Then*

$$\min_{j \in \{2,3,\ldots,d\}} (\nu_1 - |\nu_j|) \geq \frac{\pi^2}{3}\frac{\delta^3}{d^2}$$

whenever $d \geq 4\delta$ and $\delta \geq 3$.

Proof. Let $\theta_j = \frac{2\pi j}{d}$. We find the lower bound by considering that $\theta_j \in [\pi/d, 2\pi - \pi/d]$ for every $j > 0$, so

$$\nu_1 - \max|\nu_j| \geq \nu_1 - \max_{\theta \in [\pi/d, 2\pi - \pi/d]} |l_\delta(\theta)| = (2\delta - 1) - \max_{\theta \in [\pi/d, \pi]} |l_\delta(\theta)|,$$

where we have used our eigenvalue formula from (3.21), and the symmetry of l_δ about $\theta = \pi$.

We now show that l_δ is decreasing towards its first zero at $\theta = \frac{2\pi}{2\delta - 1}$ by considering

the derivative

$$l'_\delta(\theta) = \frac{(\delta - 1/2)\cos((\delta-1/2)\theta)\sin(\theta/2) - 1/2\sin((\delta-1/2)\theta)\cos(\theta/2)}{\sin(\theta/2)^2},$$

which is non-positive if and only if

$$(2\delta - 1)\sin(\theta/2)\cos((\delta-1/2)\theta) \leq \sin((\delta-1/2)\theta)\cos(\theta/2).$$

Since $\tan(\cdot)$ is convex on $[0, \pi/2)$, this last inequality will hold for $\theta \in [0, \frac{\pi}{2\delta-1})$. For $\theta \in [\frac{\pi}{2\delta-1}, \frac{2\pi}{2\delta-1})$, $\cos((\delta-1/2)\theta) \leq 0$ while the remainder of the terms are non-negative, so the inequality also holds. Therefore,

$$\nu_1 - \max_{j>1}|\nu_j| \geq (2\delta-1) - \max\left\{\nu_2, \max_{\theta\in[\frac{2\pi}{2\delta-1},\pi]}|l_\delta(\theta)|\right\},$$

which permits us to bound $(2\delta-1)-\nu_2$ and $(2\delta-1) - \max_{\theta\in[\frac{2\pi}{2\delta-1},\pi]}|l_\delta(\theta)|$ separately.

For $\max_{\theta\in[\frac{2\pi}{2\delta-1},\pi]}|l_\delta(\theta)|$, we simply observe that

$$\max_{\theta\in[\frac{2\pi}{2\delta-1},\pi]}|l_\delta(\theta)| \leq \max_{\theta\in[\frac{2\pi}{2\delta-1},\pi]}\frac{1}{\sin(\theta/2)} = \left(\sin\left(\frac{\pi}{2\delta-1}\right)\right)^{-1} \leq \frac{2\delta-1}{2},$$

where the last line uses that $\frac{\pi}{2\delta-1} \leq \pi/2$ (since $\delta \geq 3$). This yields $\nu_1 - \max_{\theta\in[\frac{2\pi}{2\delta-1},\pi]}|l_\delta(\theta)| \geq \frac{1}{2}(2\delta-1)$.

As for ν_2, we have $\theta_1\cdot(\delta-1) \leq \pi/2$ (since $4(\delta-1)\leq d$). Thus, $\cos(\cdot)$ will be concave on $[0, \theta_1(\delta-1)]$. Considering (3.20), this will give $\sum_{k=1}^{\delta-1}\cos(k\theta_1) \leq (\delta-1)\cos\left(\theta_1\frac{\delta}{2}\right)$, so

$$\begin{aligned}
\nu_1 - \nu_2 &\geq 2(\delta-1)\left(1-\cos\left(\pi\tfrac{\delta}{d}\right)\right) \\
&\geq 2(\delta-1)\left(\frac{(\pi\frac{\delta}{d})^2}{4}\right) \\
&\geq \frac{\pi^2}{3}\cdot\frac{\delta^3}{d^2}.
\end{aligned}$$

The stated result follows, since $d \geq 4\delta$ gives

$$\nu_1 - \max_{\theta \in [\frac{2\pi}{2\delta-1}, \pi]} |l_\delta(\theta)| \geq \frac{1}{2}(2\delta - 1) \geq \frac{1}{2}\delta \geq \frac{1}{3}\left(\frac{\pi\delta}{d}\right)^2 \delta = \frac{\pi^2}{3} \cdot \frac{\delta^3}{d^2}.$$

□

We are now sufficiently well informed about $\widetilde{X}_0$ to consider perturbation results for its leading eigenvector.

3.4 Perturbation Theory for $\tilde{X}_0$

In this section we will use spectral graph theoretic techniques to obtain a bound on the error associated with recovering phase information using our method. In particular, we will adapt the proof of Theorem 6.3 from [2] to develop a bound for $\min_{\theta \in [0,2\pi]} \|\widetilde{\mathbf{x}}_0 - \mathbb{e}^{\mathbb{i}\theta}\widetilde{\mathbf{x}}\|_2$. This approach involves considering both $\widetilde{X}$ from Algorithm 1 and $\widetilde{X}_0$ from (3.7) in the context of spectral graph theory, so we begin by defining essential terms. The idea is to consider a graph whose vertices correspond to the entries of $\widetilde{\mathbf{x}}_0$ from (3.7), and whose edges carry the relative phase data.[9]

We begin with an undirected graph $G = (V, E)$ with vertex set $V = \{1, 2, \ldots, d\}$ and weight mapping $w : V \times V \to \mathbb{R}^+$, where $w_{ij} = w_{ji}$ and $w_{ij} = 0$ iff $\{i, j\} \notin E$. The *degree* of a vertex i is

$$\deg(i) := \sum_{j \text{ s.t. } (i,j) \in E} w_{ij},$$

and we define the *degree matrix* and *weighted adjacency matrix* of G by

$$D := \operatorname{diag}(\deg(i)) \text{ and } W_{ij} := w_{ij},$$

[9]The interested reader is also referred to Appendix 3.8 where more standard perturbation theoretic techniques are utilized in order to obtain a weaker bound on the error associated with recovering phase information via the proposed approach.

respectively. The *volume* of G is

$$\mathrm{vol}(G) := \sum_{i \in V} \deg(i).$$

Finally, the *Laplacian* of G is the $d \times d$ real symmetric matrix

$$L := I - D^{-1/2} W D^{-1/2} = D^{-1/2}(D - W)D^{-1/2}, \tag{3.22}$$

where $I \in \{0,1\}^{d \times d}$ is the identity matrix.

When G is connected, Lemma 1.7 of [26] shows that the nullspace of $(D - W)$ is $\mathrm{span}(\mathbb{1})$, and the nullspace of L is $\mathrm{span}(D^{1/2}\mathbb{1})$. Observing that $D - W$ is diagonally semi-dominant, it follows from Gershgorin's disc theorem that $(D - W)$ and L are both positive semidefinite. Alternatively, one may also note that

$$\mathbf{v}^*(D - W)\mathbf{v} = \sum_{i \in V} \left(v_i^2 \deg(i) - \sum_{j \in V} v_i v_j w_{ij} \right) = \frac{1}{2} \sum_{i,j \in V} w_{ij}(v_i - v_j)^2 \geq 0$$

holds for all $\mathbf{v} \in \mathbb{R}^d$. Thus, we may order the eigenvalues of L in increasing order so that $0 = \lambda_1' < \lambda_2' \leq \cdots \leq \lambda_n'$. We then define the *spectral gap* of G to be $\tau = \lambda_2'$.

Herein, though we will state the main theorem of this section more generally, we will mainly be interested in the case where the graph $G = (V, E)$ is the simple unweighted graph whose adjacency matrix is $U = T_\delta(\mathbb{1}\mathbb{1}^*)$ as in (3.19). In this case we will have $W = U - I_d$ and $D = (2\delta - 2)I_d$. We also immediately obtain the following corollary of Lemmas 1 and 2.

Corollary 1. *Let G be the simple unweighted graph whose adjacency matrix is U from* (3.19)*. Let L be the Laplacian of G. Then, there exists a bijection $\sigma : [d] \to [d]$ such that*

$$\lambda'_{\sigma(j)} = \frac{2\delta - 1}{2\delta - 2} - \frac{1 + 2\sum_{k=1}^{\delta - 1} \cos\left(\frac{2\pi(j-1)k}{d}\right)}{2\delta - 2}$$

for $j = 1, \ldots, d$. In particular, if $d \geq 4\delta$ and $\delta \geq 3$ then

$$\tau = \lambda_2' > \frac{\pi^2}{6}\frac{\delta^2}{d^2}.$$

Using this graph G as a scaffold we can now represent our computed relative phase matrix $\widetilde{X}$ from Algorithm 1 by noting that for some (Hermitian) perturbations η_{ij} we will have

$$\widetilde{X}_{ij} = \frac{(x_0)_i(x_0)_j^* + \eta_{ij}}{|(x_0)_i(x_0)_j^* + \eta_{ij}|} \cdot w_{ij} = \frac{(x_0)_i(x_0)_j^* + \eta_{ij}}{|(x_0)_i(x_0)_j^* + \eta_{ij}|} \cdot \chi_{E(i,j)}. \tag{3.23}$$

Using this same notation we may also represent our original phase matrix $\widetilde{X}_0$ via G by noting that

$$(\widetilde{X}_0)_{ij} = \frac{(x_0)_i(x_0)_j^*}{|(x_0)_i(x_0)_j^*|} \cdot w_{ij} = \operatorname{sgn}\left((x_0)_i(x_0)_j^*\right) \cdot \chi_{E(i,j)}. \tag{3.24}$$

We may now define the *connection Laplacian* of the graph G associated with the Hermitian and entrywise normalized data given by $\widetilde{X}$ to be the matrix

$$L_1 = I - D^{-1/2}(\widetilde{X} \circ W)D^{-1/2}, \tag{3.25}$$

where $\circ$ denotes entrywise (Hadamard) multiplication. Following [8], given $\widetilde{X}$ and a vector $\mathbf{y} \in \mathbb{C}^d$, we define the *frustration of* $\mathbf{y}$ *with respect to* $\widetilde{X}$ by

$$\eta_{\widetilde{X}}(\mathbf{y}) := \frac{1}{2}\frac{\sum_{(i,j)\in E} w_{ij}|y_i - \widetilde{X}_{ij}y_j|^2}{\sum_{i\in V}\deg(i)|y_i|^2} = \frac{\mathbf{y}^*(D - (\widetilde{X} \circ W))\mathbf{y}}{\mathbf{y}^* D\mathbf{y}}. \tag{3.26}$$

We may consider $\eta_{\widetilde{X}}(\mathbf{y})$ to measure how well $\mathbf{y}$ (viewed as a map from V to $\mathbb{C}$) conforms to the computed relative phase differences $\widetilde{X}$ across the graph G.

In addition, we adapt a result from [8]:

Lemma 3 (Cheeger inequality for the connection Laplacian)**.** *Suppose that $G = (V = [d], E)$ is a connected graph with degree matrix $D \in [0,\infty)^{d\times d}$, weighted adjacency matrix $W \in [0,\infty)^{d\times d}$, and spectral gap $\tau > 0$, and that $\widetilde{X} \in \mathbb{C}^{d\times d}$ is Hermitian and entrywise*

normalized. Let $\mathbf{u} \in \mathbb{C}^d$ *be an eigenvector of* L_1 *from* (3.25) *corresponding to its smallest eigenvalue. Then,* $\mathbf{w} = \operatorname{sgn}(\mathbf{u}) = \operatorname{sgn}\left(D^{-1/2}\mathbf{u}\right)$ *satisfies*

$$\eta_{\widetilde{X}}(\mathbf{w}) \leq \frac{44}{\tau} \cdot \min_{\mathbf{y}\in\mathbb{C}^d} \eta_{\widetilde{X}}(\operatorname{sgn}(\mathbf{y})).$$

Proof of Lemma 3. One can see that

$$\begin{aligned}\inf_{\mathbf{v}\in\mathbb{C}^d\setminus\{\mathbf{0}\}} \frac{\mathbf{v}^* L_1 \mathbf{v}}{\mathbf{v}^*\mathbf{v}} &= \inf_{\mathbf{y}\in\mathbb{C}^d\setminus\{\mathbf{0}\}} \frac{(D^{1/2}\mathbf{y})^* L_1 (D^{1/2}\mathbf{y})}{(D^{1/2}\mathbf{y})^*(D^{1/2}\mathbf{y})} = \inf_{\mathbf{y}\in\mathbb{C}^d\setminus\{\mathbf{0}\}} \frac{\mathbf{y}^*(D - (\widetilde{X}\circ W))\mathbf{y}}{\mathbf{y}^* D\mathbf{y}} \\ &= \inf_{y\in\mathbb{C}^d\setminus\{\mathbf{0}\}} \eta_{\widetilde{X}}(\mathbf{y}) \leq \min_{\mathbf{y}\in\mathbb{C}^d} \eta_{\widetilde{X}}(\operatorname{sgn}(\mathbf{y})).\end{aligned}$$

From here, Lemma 3.6 in [8] gives

$$\eta_{\widetilde{X}}(\mathbf{w}) \leq \frac{44}{\tau}\eta_{\widetilde{X}}\left(D^{-1/2}\mathbf{u}\right) = \frac{44}{\tau} \cdot \inf_{\mathbf{v}\in\mathbb{C}^d\setminus\{\mathbf{0}\}} \frac{\mathbf{v}^* L_1 \mathbf{v}}{\mathbf{v}^*\mathbf{v}} \leq \frac{44}{\tau} \cdot \min_{\mathbf{y}\in\mathbb{C}^d} \eta_{\widetilde{X}}(\operatorname{sgn}(\mathbf{y})).$$

□

We now state the main result of this section:

Theorem 4. *Suppose that* $G = (V = [d], E)$ *is an undirected, connected, and unweighted graph (so that* $W_{ij} = \chi_{E(i,j)}$*) with spectral gap* $\tau > 0$*. Let* $\mathbf{u} \in \mathbb{C}^d$ *be an eigenvector of* L_1 *from* (3.25) *corresponding to its smallest eigenvalue, and let*

$$\widetilde{\mathbf{x}} = \operatorname{sgn}(\mathbf{u}) \text{ and } \widetilde{\mathbf{x}}_0 = \operatorname{sgn}(\mathbf{x}_0).$$

Then

$$\min_{\theta\in[0,2\pi]} ||\widetilde{\mathbf{x}} - \mathrm{e}^{\mathrm{i}\theta}\widetilde{\mathbf{x}}_0||_2 \leq 19\frac{\|\widetilde{X} - \widetilde{X}_0\|_F}{\tau \cdot \sqrt{\min_{i\in V}(\deg(i))}},$$

where $\widetilde{X}$ *and* $\widetilde{X}_0$ *are defined as per* (3.23) *and* (3.24)*, respectively.*

The proof follows by combining the two following lemmas, which share the hypotheses of the theorem. Additionally, we introduce the notation $\mathbf{g} \in \mathbb{C}^d$ and $\Lambda \in \mathbb{C}^{d\times d}$, where

$$g_i = (\widetilde{\mathbf{x}}_0)_i^* \widetilde{\mathbf{x}}_i \quad \text{and} \quad \Lambda_{ij} = (\widetilde{X}_0)_{ij}^* \widetilde{X}_{ij},$$

and observe that $|g_i| = |\Lambda_{ij}| = 1$ for each $(i,j) \in E$.

Lemma 4. *Under the hypotheses of Theorem 4, there exists an angle $\theta \in [0, 2\pi]$ such that*

$$\tau \sum_{i \in V} \deg(i) |g_i - \mathrm{e}^{\mathrm{i}\theta}|^2 \leq 2 \sum_{(i,j) \in E} |g_i - g_j|^2.$$

Lemma 5. *Under the hypotheses of Theorem 4,*

$$2 \sum_{(i,j) \in E} |g_i - g_j|^2 \leq \frac{356}{\tau} \|\widetilde{X} - \widetilde{X}_0\|_F^2.$$

From these lemmas, the theorem follows immediately by observing $\sum_{i \in V} |g_i - \mathrm{e}^{\mathrm{i}\theta}|^2 = \|\widetilde{\mathbf{x}} - \mathrm{e}^{\mathrm{i}\theta} \widetilde{\mathbf{x}}_0\|_2^2$.

Proof of Lemma 4. We set $\alpha = \dfrac{\sum_{i \in V} \deg(i) g_i}{\mathrm{vol}(G)}$ and $w_i = g_i - \alpha$. Then

$$\mathbb{1}^* D \mathbf{w} = \sum_{i \in V} \deg(i)(g_i - \alpha) = 0,$$

so $D^{1/2}\mathbf{w}$ is orthogonal to $D^{1/2}\mathbb{1}$. Noting that the null space of L is spanned by $D^{1/2}\mathbb{1}$ when $\tau > 0$, and recalling that $L \succeq 0$, we have

$$\frac{(D^{1/2}\mathbf{w})^* L (D^{1/2}\mathbf{w})}{\mathbf{w}^* D \mathbf{w}} \geq \min_{\mathbf{y}^* D^{1/2} \mathbb{1} = 0} \frac{\mathbf{y}^* L \mathbf{y}}{\mathbf{y}^* \mathbf{y}} = \tau.$$

Therefore,

$$\begin{aligned} \tau \mathbf{w}^* D \mathbf{w} &\leq \mathbf{w}^*(D - W)\mathbf{w} &&= \mathbf{g}^*(D - W)\mathbf{g} \\ &= \sum_{i \in V} \deg(i) |g_i|^2 - \sum_{i \in V} g_i^* \sum_{(i,j) \in E} g_j &&= \sum_{(i,j) \in E} (1 - g_i^* g_j) \\ &= \frac{1}{2} \sum_{(i,j) \in E} |g_i - g_j|^2. \end{aligned}$$

We note that $\tau \mathbf{w}^* D \mathbf{w} = \tau \sum_{i \in V} \deg(i) |g_i - \alpha|^2$, while we seek a bound on $\sum_{i \in V} \deg(i) |g_i -$

$\mathrm{e}^{\mathrm{i}\theta}|^2$. To that end, we use the fact that $|g_i| = |\operatorname{sgn}(\alpha)| = 1$ to obtain

$$|g_i - \operatorname{sgn}(\alpha)| \leq |g_i - \alpha| + |\alpha - \operatorname{sgn}(\alpha)| \leq 2|g_i - \alpha|.$$

Setting $\theta := \arg \alpha$, we have the stated result. □

Proof of Lemma 5. Observe that for any two real numbers $a, b \in \mathbb{R}$, we have $\frac{1}{2}a^2 - b^2 \leq (a-b)^2$. Thus, by the reverse triangle inequality we have

$$\begin{aligned}
\sum_{(i,j)\in E} \left(\frac{1}{2}|g_i - g_j|^2 - |\Lambda_{ij} - 1|^2 \right) &\leq \sum_{(i,j)\in E} \left(|g_i - g_j| - |\Lambda_{ij} - 1| \right)^2 \\
&\leq \sum_{(i,j)\in E} |g_i - \Lambda_{ij} g_j|^2 \\
&= \sum_{(i,j)\in E} |\widetilde{\mathbf{x}}_i - \widetilde{X}_{ij} \widetilde{\mathbf{x}}_j|^2 \\
&= 2 \operatorname{vol}(G) \cdot \eta_{\widetilde{X}}(\widetilde{\mathbf{x}}),
\end{aligned} \tag{3.27}$$

as the denominator of (3.26) is $2\operatorname{vol}(G)$ whenever the entries of $\mathbf{y}$ all have unit modulus.

Lemma 3 now tells us that

$$\begin{aligned}
\sum_{(i,j)\in E} \left(\frac{1}{2}|g_i - g_j|^2 - |\Lambda_{ij} - 1|^2 \right) &\leq \frac{2 \cdot 44 \operatorname{vol}(G)}{\tau} \min_{\mathbf{y}\in\mathbb{C}^d} \eta_{\widetilde{X}}(\operatorname{sgn}(\mathbf{y})) \\
&\leq \frac{88 \operatorname{vol}(G)}{\tau} \eta_{\widetilde{X}}(\widetilde{\mathbf{x}}_0).
\end{aligned} \tag{3.28}$$

Moreover,

$$\begin{aligned}
\eta_{\widetilde{X}}(\widetilde{\mathbf{x}}_0) &= \frac{\displaystyle\sum_{(i,j)\in E} |(\widetilde{\mathbf{x}}_0)_i - \widetilde{X}_{ij} (\widetilde{\mathbf{x}}_0)_j|^2}{2 \displaystyle\sum_{i\in V} \deg(i) |(\widetilde{\mathbf{x}}_0)_i|^2} \\
&= \frac{\displaystyle\sum_{(i,j)\in E} |(\widetilde{\mathbf{x}}_0)_i (\widetilde{\mathbf{x}}_0)_j^* - \widetilde{X}_{ij}|^2}{2 \operatorname{vol}(G)} \\
&= \frac{\|\widetilde{X}_0 - \widetilde{X}\|_F^2}{2 \operatorname{vol}(G)},
\end{aligned}$$

so that $\sum_{(i,j)\in E} \frac{1}{2}|g_i - g_j|^2 \leq \frac{88}{\tau}\|X_0 - X\|_F^2 + \sum_{(i,j)\in E} |\Lambda_{ij} - 1|^2$. Considering also that

$$\sum_{(i,j)\in E} |\Lambda_{ij} - 1|^2 = \sum_{(i,j)\in E} \left|\widetilde{X}_{ij} - (\widetilde{X}_0)_{ij}\right|^2 = \left\|\widetilde{X} - \widetilde{X}_0\right\|_F^2 \tag{3.29}$$

and $\tau \leq 1$, this completes the proof. □

We may now use Theorem 4 to produce a perturbation bound for our banded matrix of phase differences $\widetilde{X}_0$.

Corollary 2. *Let $\widetilde{X}_0$ be the matrix in* (3.7)*, $\widetilde{\mathbf{x}}_0$ be the vector of true phases* (3.7)*, and $\widetilde{X}$ be as in line 3 of Algorithm 1 with $\widetilde{\mathbf{x}} = \operatorname{sgn}(\mathbf{u})$ where $\mathbf{u}$ is the top eigenvector of $\widetilde{X}$. Suppose that $\|\widetilde{X}_0 - \widetilde{X}\|_F \leq \eta\|\widetilde{X}_0\|_F$ for some $\eta > 0$. Then*

$$\min_{\theta\in[0,2\pi]} \|\widetilde{\mathbf{x}}_0 - \mathrm{e}^{\mathrm{i}\theta}\widetilde{\mathbf{x}}\|_2 \leq 12\frac{\eta d^{\frac{5}{2}}}{\delta^2}.$$

Proof. We apply Theorem 4 with the unweighted and undirected graph $G = (V, E)$, where $V = [d]$ and $E = \{(i, j) : |i - j| \bmod d < \delta\}$. Observe that G is also connected and $(2\delta - 1)$-regular so that $\min_{i\in V}(\deg(i)) = 2\delta - 1$. The spectral gap of G is $\tau > \frac{\pi^2}{6}\delta^2/d^2 > 0$ by Corollary 1. We know that $\|\widetilde{X}_0\|_F = \sqrt{d(2\delta - 1)}$, so that $\|\widetilde{X}_0 - \widetilde{X}\|_F \leq \eta\sqrt{d(2\delta - 1)}$. Finally, if $\mathbf{u}$ is the top eigenvector of $\widetilde{X}$ then it will also be an eigenvector of L_1 corresponding to its smallest eigenvalue since, here, $L_1 = I - \frac{1}{2\delta-1}\widetilde{X}$.

Combining these observations we have

$$\min_{\theta\in[0,2\pi]} \|\widetilde{\mathbf{x}}_0 - \mathrm{e}^{\mathrm{i}\theta}\widetilde{\mathbf{x}}\|_2 \leq 19\frac{\eta(d(2\delta - 1))^{1/2}}{\pi^2/6 \cdot \delta^2/d^2 \cdot (2\delta - 1)^{1/2}} \leq 12\frac{\eta d^{5/2}}{\delta^2}.$$

□

We are now properly equipped to analyze the robustness of Algorithm 1 to noise.

3.5 Recovery Guarantees for the Proposed Method

Herein we will assume Algorithm 1 is provided with measurements $\mathbf{y}$ of the form (3.8) such that the linear operator (3.5) is invertible on $T_\delta(\mathbb{C}^{d\times d})$ with condition number $\kappa > 0$. Unless otherwise stated, we follow the notation of Sections 3.1.1 and 3.1.2; therefore, our assumptions imply that $\|X - X_0\|_F \leq \kappa \frac{\|X_0\|_F}{\text{SNR}}$ or $\|X - X_0\|_F \leq \sigma_{\min}^{-1}\|\mathbf{n}\|_2$.

We now aim to bound the Frobenius norm of the perturbation error $(\tilde{X} - \tilde{X}_0)$ present in the matrix $\tilde{X}$ formed in line 2 of Algorithm 1. Toward this end we define the set of ρ-small indexes of $\mathbf{x}_0$ to be

$$S_\rho := \left\{ j \;\middle|\; |\,(x_0)_j\,| < \left(\frac{\sigma_{\min}^{-1}\|\mathbf{n}\|_2}{\rho}\right)^{\frac{1}{4}} \right\} \tag{3.30}$$

where $\rho \in \mathbb{R}^+$ is a free parameter. With the definition of S_ρ in hand we can bound the perturbation error $(\tilde{X} - \tilde{X}_0)$ using the next lemma.

Lemma 6. *Let $\tilde{X}$ be the matrix computed in line 2 of Algorithm 1. We have that*

$$\|\widetilde{X} - \widetilde{X}_0\|_{\text{F}} \leq 4\sqrt{\frac{\rho\frac{1}{\sigma_{\min}\delta}\|\mathbf{n}\|_2 + |S_\rho|}{d}} \cdot \|\widetilde{X}_0\|_{\text{F}} \tag{3.31}$$

holds for all $\rho \in \mathbb{R}^+$. In particular, setting $\rho = \frac{\sigma_{\min}^{-1}\|\mathbf{n}\|_2}{|(x_0)_{\min}|^4}$, where $|(x_0)_{\min}| = \min_j |(x_0)_j|$, we have that S_ρ is empty and

$$\|\widetilde{X} - \widetilde{X}_0\|_F \leq 2\frac{\sigma_{\min}^{-1}\|\mathbf{n}\|_2}{|(x_0)_{\min}|^2}\frac{\|\widetilde{X}_0\|_F}{\sqrt{d\delta}}. \tag{3.32}$$

Proof. Set $N_{jk} = X_{jk} - (X_0)_{jk}$ and consider that, for any $j, k \in S_\rho^c$,

$$\begin{aligned}
|(\widetilde{X}_0)_{jk} - \widetilde{X}_{jk}| &= \left|(\widetilde{X}_0)_{jk} - \text{sgn}\left(\frac{X_{jk}}{|(X_0)_{jk}|}\right)\right| \\
&\leq \left|(\widetilde{X}_0)_{jk} - \frac{X_{jk}}{|(X_0)_{jk}|}\right| + \left|\frac{X_{jk}}{|(X_0)_{jk}|} - \text{sgn}\left(\frac{X_{jk}}{|(X_0)_{jk}|}\right)\right| \\
&\leq 2\left|(\widetilde{X}_0)_{jk} - \frac{X_{jk}}{|(X_0)_{jk}|}\right| = 2\frac{|N_{jk}|}{|(X_0)_{jk}|} \leq 2\rho^{\frac{1}{2}}\frac{|N_{jk}|}{(\sigma_{\min}^{-1}\|\mathbf{n}\|_2)^{\frac{1}{2}}}.
\end{aligned} \tag{3.33}$$

Thus, we may write

$$\begin{aligned}\|\widetilde{X} - \widetilde{X}_0\|_{\mathrm{F}}^2 &\leq \sum_{j,k \in S_\rho^c} 4\rho \frac{|N_{jk}|^2}{\sigma_{\min}^{-1}\|\mathbf{n}\|_2} + \sum_{j \in S_\rho, \text{ or } k \in S_\rho} |(\widetilde{X}_0)_{jk} - \widetilde{X}_{jk}|^2 \\ &\leq 4\rho \frac{\|N\|_{\mathrm{F}}^2}{\sigma_{\min}^{-1}\|\mathbf{n}\|_2} + \sum_{j \in S_\rho} 4 \cdot (4\delta - 3) = 4\rho \frac{\|N\|_{\mathrm{F}}^2}{\sigma_{\min}^{-1}\|\mathbf{n}\|_2} + 4 \cdot (4\delta - 3)\,|S_\rho| \\ &\leq 16(\rho \sigma_{\min}^{-1}\|\mathbf{n}\|_2 + \delta\,|S_\rho|).\end{aligned}$$

The proof is completed by recalling that $\|\widetilde{X}_0\|_F = \sqrt{(2\delta - 1)d} \geq \sqrt{\delta d}$. □

Our robustness result relies additionally on the following lemma.

Lemma 7. *Suppose $X, \underline{X} \in \mathcal{H}^d$. Define $x, \underline{x} \in \mathbb{R}^d$ by $x_i = \sqrt{|X_{ii}|}$ and $\underline{x}_i = \sqrt{|\underline{X}_{ii}|}$. Then*

$$\|x - \underline{x}\|_2 \leq \sqrt{\|\mathrm{diag}(X - \underline{X})\|_1} \leq d^{1/4}\sqrt{\|X - \underline{X}\|_F}.$$

Proof of Lemma 7. We first claim that, for $f(x) = (1 - |1 - x|^{1/2})^2$, we have $f(x) \leq |x|$ for all $x \in \mathbb{R}$. To see this for $x \geq 1$, we set $t = |x - 1|^{1/2} = (x - 1)^{1/2}$ and observe

$$x = t^2 + 1 \geq t^2 - 2t + 1 = (t - 1)^2 = f(x).$$

For $0 \leq x \leq 1$, we set $t = |1 - x|^{1/2} = (1 - x)^{1/2}$ and see

$$x = 1 - t^2 = (1 - t)(1 + t) \geq (1 - t)(1 - t) = (1 - t)^2 = f(x).$$

For $x \leq 0$, we consider $g(t) = f(-t)$ for $t \geq 0$. Then

$$g(t) = (1 - (1 + t)^{1/2})^2 = 2 + t - 2\sqrt{1 + t} \leq t,$$

simply by bounding $\sqrt{1 + t} \geq 1$. From this, it follows that

$$\left(a - \left|a^2 - b\right|^{1/2}\right)^2 = a^2 \left(1 - \left|1 - \frac{b}{a^2}\right|^{1/2}\right)^2 \leq |b| \tag{3.34}$$

for any $a, b \in \mathbb{R}$.

Setting $N = X - \underline{X}$, we may write $X_{ij} = \underline{X}_{ij} + N_{ij}$. In particular, $X_{ii} = \underline{X}_{ii} + N_{ii} = \underline{x}_i^2 + N_{ii}$, so that $x_i = \sqrt{|\underline{x}_i^2 + N_{ii}|}$. Setting $n_i = N_{ii}$, (3.34) gives

$$||x_i| - |\underline{x}_i||^2 \leq |n_i|.$$

Trivially, we have $\|n\|_2 \leq \|X - \underline{X}\|_F$, so

$$\|x - \underline{x}\|_2 \leq \sqrt{\sum_{i=1}^{d} |n_i|} \leq \sqrt{\sqrt{d}\|n\|_2} \leq d^{1/4}\sqrt{\|X - \underline{X}\|_F},$$

as desired. □

We are finally ready to prove a robustness result for Algorithm 1.

Theorem 5. *Suppose $\delta > 2$ and $d \geq 4\delta$, and that $\tilde{X}$ and $\tilde{X}_0$ satisfy $\|\tilde{X} - \tilde{X}_0\|_{\mathrm{F}} \leq \eta\|\tilde{X}_0\|_F$ for some $\eta > 0$. Then, the estimate $\mathbf{x}$ produced by Algorithm 1 satisfies*

$$\min_{\theta \in [0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 \leq 12\|\mathbf{x}_0\|_\infty \left(\frac{d^{5/2}}{\delta^2}\right)\eta + d^{\frac{1}{4}}\sqrt{\sigma_{\min}^{-1}\|\mathbf{n}\|_2},$$

Alternatively, one can bound the error in terms of the size of the index set S_ρ from (3.30) *as*

$$\min_{\theta \in [0,2\pi]} \left\|\mathbf{x}_0 - \mathbb{e}^{\mathbb{i}\theta}\mathbf{x}\right\|_2 \leq 48\|\mathbf{x}_0\|_\infty \left(\frac{d}{\delta}\right)^2 \sqrt{\frac{\rho\|\mathbf{n}\|_2}{\sigma_{\min}\delta} + |S_\rho|} + d^{\frac{1}{4}}\sqrt{\sigma_{\min}^{-1}\|\mathbf{n}\|_2}, \tag{3.35}$$

for any desired $\rho \in \mathbb{R}^+$. Stated in terms of SNR*, these inequalities become*

$$\begin{aligned} \min_{\theta \in [0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 &\leq 12\|\mathbf{x}_0\|_\infty \left(\frac{d^{5/2}}{\delta^2}\right)\eta + d^{\frac{1}{4}}\sqrt{\kappa\frac{\|X_0\|_F}{\mathrm{SNR}}}, \\ \min_{\theta \in [0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 &\leq 48\|\mathbf{x}_0\|_\infty \left(\frac{d}{\delta}\right)^2 \sqrt{\rho\frac{\kappa\|X_0\|_F}{(\mathrm{SNR})\delta} + |S_\rho|} + d^{\frac{1}{4}}\sqrt{\kappa\frac{\|X_0\|_F}{\mathrm{SNR}}} \end{aligned} \tag{3.36}$$

Proof. Let $\phi \in [0, 2\pi)$ be arbitrary; then $\mathbb{e}^{\mathbb{i}\phi}\mathbf{x} = |\mathbf{x}| \circ \mathbb{e}^{\mathbb{i}\phi}\tilde{\mathbf{x}}$ and $\mathbf{x}_0 = |\mathbf{x}_0| \circ \tilde{\mathbf{x}}_0$, where $\circ$ denotes the entrywise (Hadamard) product.

We see that

$$\begin{aligned}\min_{\theta\in[0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 &= \min_{\theta\in[0,2\pi]} \left\| |\mathbf{x}_0| \circ \tilde{\mathbf{x}}_0 - |\mathbf{x}| \circ \mathrm{e}^{\mathrm{i}\theta}\tilde{\mathbf{x}} \right\|_2 \\ &\leq \min_{\theta\in[0,2\pi]} \left\| |\mathbf{x}_0| \circ \tilde{\mathbf{x}}_0 - |\mathbf{x}_0| \circ \mathrm{e}^{\mathrm{i}\theta}\tilde{\mathbf{x}} \right\|_2 + \left\| |\mathbf{x}_0| \circ \mathrm{e}^{\mathrm{i}\theta}\tilde{\mathbf{x}} - |\mathbf{x}| \circ \mathrm{e}^{\mathrm{i}\theta}\tilde{\mathbf{x}} \right\|_2\end{aligned}$$

where the second term is now independent of ϕ. As a result we have that

$$\min_{\theta\in[0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 \leq \|\mathbf{x}_0\|_\infty \left(\min_{\theta\in[0,2\pi]} \|\tilde{\mathbf{x}}_0 - \mathrm{e}^{\mathrm{i}\theta}\tilde{\mathbf{x}}\|_2 \right) + d^{1/4}\sqrt{\sigma_{\min}^{-1} \cdot \|\mathbf{n}\|_2}$$

Here, the bound on the second term follows from Lemma 7. The first inequality of the theorem now results from an application of Corollary 2 to the first term. The second inequality then follows from Lemma 6. □

Looking at the second inequality (3.35) in Theorem 5 we can see that the error bound there will be vacuous in most settings unless $S_\rho = \emptyset$. Recalling (3.30), one can see that S_ρ will be empty as soon as $\rho = \sigma_{\min}^{-1}\|\mathbf{n}\|_2 / |(x_0)_{\min}|^4$, where $(x_0)_{\min}$ is the smallest magnitude of any entry in $\mathbf{x}_0$. Utilizing this value of ρ in (3.35) leads to proving Theorem 1 as a corollary of Theorem 5 by quoting (3.32) from Lemma 6.

Corollary 3. *Suppose $\delta > 2$ and $d \geq 4\delta$. Let $(x_0)_{\min} := \min_j |(x_0)_j|$ be the smallest magnitude of any entry in $\mathbf{x}_0$. Then, the estimate $\mathbf{x}$ produced by Algorithm 1 satisfies*

$$\begin{aligned}\min_{\theta\in[0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 &\leq 24 \left(\frac{\|\mathbf{x}_0\|_\infty}{(x_0)_{\min}^2} \right) \frac{d^2}{\delta^{5/2}} \sigma_{\min}^{-1}\|\mathbf{n}\|_2 + d^{\frac{1}{4}} \sqrt{\sigma_{\min}^{-1}\|\mathbf{n}\|_2}, \\ \min_{\theta\in[0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 &\leq 24 \left(\frac{\|\mathbf{x}_0\|_\infty}{(x_0)_{\min}^2} \right) \frac{d^2}{\delta^{5/2}} \kappa \frac{\|X_0\|_F}{\mathrm{SNR}} + d^{\frac{1}{4}} \sqrt{\kappa \frac{\|X_0\|_F}{\mathrm{SNR}}}\end{aligned}$$

Corollary 3 yields a deterministic recovery result for any signal $\mathbf{x}_0$ which contains no zero entries. If desired, a randomized result can now be derived from Corollary 3 for arbitrary $\mathbf{x}_0$ by right multiplying the signal $\mathbf{x}_0$ with a random "flattening" matrix as done in [57].

3.6 Numerical Evaluation

We now present numerical simulations supporting the theoretical recovery guarantees in Section 3.5. In addition to illustrating the performance of our algorithm, we also compare it against other existing phase retrieval methods using *local* measurements. All the results presented here may be recreated using the open source *BlockPR* Matlab software package which is freely available at [58].

In Section 3.6.2 we utilize the sparse measurement masks of Example 2 (see Section 3.2 for details). For each choice of δ these measurements correspond to $2\delta - 1$ physical masks which are each shifted d times across the sample $\mathbf{x}_0$ by shifts of size 1. In Section 3.6.3, we then consider the Fourier measurement masks of Example 1 from Section 3.2. For each choice of δ these measurements correspond to $2\delta - 1$ Fourier measurements of *one* physical mask/illumination which is also shifted d times across the sample $\mathbf{x}_0$ by shifts of size 1. These Fourier measurements correspond particularly well to ptychographic measurements of a large sample in the small δ regime.

For completeness, we also present selected results comparing the proposed formulation against other well established phase retrieval algorithms (such as *Wirtinger Flow*) using *global* measurements such as coded diffraction patterns (CDPs) [21, §1.5]. These measurements correspond to those one might obtain from the diffraction patterns of the sample $\mathbf{x}_0$ after it has been masked by several different global random windows. Unless otherwise stated, we use i.i.d. zero-mean complex Gaussian random test signals with measurement errors modeled using an additive Gaussian noise model. Applied measurement noise and reconstruction error are both reported in decibels (dB) in terms of signal to noise ratios (SNRs),[10] with

$$\mathrm{SNR}_{\mathrm{db}} = 10\log_{10}\left(\frac{\sum_{j=1}^{D}|\langle \mathbf{a}_j, \mathbf{x}_0\rangle|^4}{D\sigma^2}\right),$$
$$\text{Error (dB)} = 10\log_{10}\left(\frac{\min_\theta \|\mathrm{e}^{\mathrm{i}\theta}\mathbf{x} - \mathbf{x}_0\|_2^2}{\|\mathbf{x}_0\|_2^2}\right),$$

[10] Note that we distinguish $\mathrm{SNR}_{\mathrm{db}}$ from $\mathrm{SNR} = \frac{\|\mathcal{A}(xx^*)\|_2}{\|n\|_2}$, though in the discussions of this section, we use "SNR" to refer to $\mathrm{SNR}_{\mathrm{db}}$.

where $\mathbf{a}_j, \mathbf{x}_0, \mathbf{x}, \sigma^2$ and D denote the measurement vectors, true signal, recovered signal, (Gaussian) noise variance and number of measurements respectively. All simulations were performed on a laptop computer running GNU/Linux (Ubuntu Linux 16.04 `x86_64`) with an Intel® Core™i3-3120M (2.5 GHz) processor, 4GB RAM and Matlab R2015b. Each data point in the timing and robustness plots was obtained as the average of 100 trials.

3.6.1 Numerical Improvements to Algorithm 1: Magnitude Estimation

Looking at the matrix X formed on line 1 of Algorithm 1 one can see that

$$X = X_0 + N',$$

where X_0 is the banded Hermitian matrix $T_\delta(\mathbf{x}_0\mathbf{x}_0^*)$ defined in (3.6), and N' contains arbitrary banded Hermitian noise. As stated and analyzed above, Algorithm 1 estimates the magnitude of each entry of $\mathbf{x}_0$ by observing that

$$X_{jj} = |(x_0)_j|^2 + N'_{jj}, \quad j \in [d].$$

Though this magnitude estimate suffices for our theoretical treatment above, it can be improved in practice by using slightly more general techniques.

Considering the component-wise magnitude of X, $|X| \in \mathbb{R}^{d\times d}$, one can see that its entries are

$$|X|_{jk} = \begin{cases} |(x_0)_j||(x_0)_k| + N''_{jk} & \text{if } |j-k| \bmod d < \delta \\ 0 & \text{otherwise} \end{cases},$$

where $N'' = |X| - |X_0|$ represents the changes in magnitude to the entries of $|X_0|$ due to noise. We may then let $D_j \in \mathbb{R}^{\delta\times\delta}$ denote the submatrix of $|X|$ given by

$$(D_j)_{kh} = |X|_{(j+k-1) \bmod d,\ (j+h-1) \bmod d},$$

for all $j \in [d]$; similarly we let N''_j denote the respective submatrices of N''. With this notation, it is clear that

$$D_j = |\mathbf{x}_0|^{(j)}(|\mathbf{x}_0|^{(j)})^* + N''_j,$$

where $|\mathbf{x}_0|^{(j)}_k = |\mathbf{x}_0|_{k+j-1}, k \in [\delta]$. This immediately suggests that we can estimate the magnitudes of the entries of $\mathbf{x}_0$ by calculating the top eigenvectors of these approximately rank one D_j matrices.

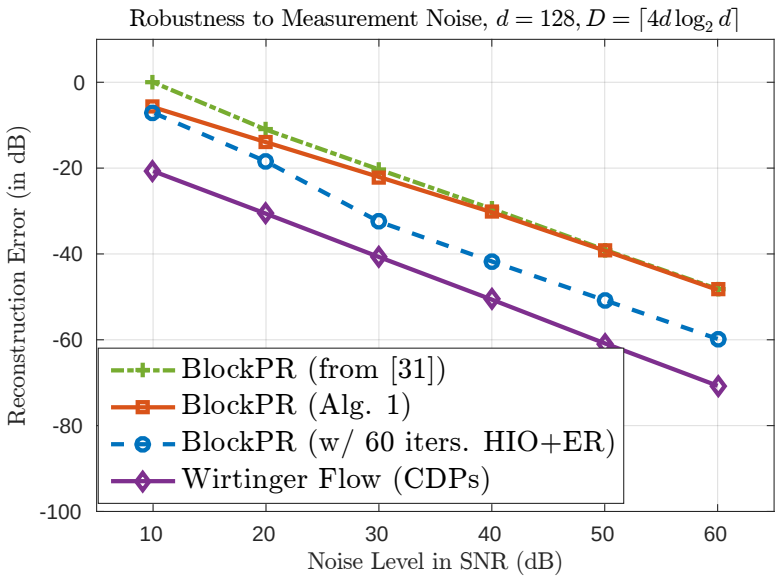
Indeed, if we do so for all of $D_1, \ldots, D_d \in \mathbb{R}^{\delta \times \delta}$, we will produce δ estimates of each $(x_0)_j$ entry's magnitude. A final estimate of each $|(x_0)_j|$ can then be computed by taking the average, median, etc. of the δ different estimates of $|(x_0)_j|$ provided by each of the leading eigenvectors of $D_{j-\delta+1}, \ldots, D_j$; in our experiments, we used the arithmetic mean. Of course, one need neither use all d possible D_j matrices, nor make them have size $\delta \times \delta$. More generally, to reduce computational complexity, one may instead use d/s matrices, $\tilde{D}_{j'} \in \mathbb{R}^{\gamma \times \gamma}$, of size $1 \leq \gamma \leq \delta$ and with shifts $s \leq \gamma$ (dividing d), having entries

$$(\tilde{D}_{j'})_{k,h} = |X|_{(sj'+k-s) \bmod d,\ (sj'+h-s) \bmod d}.$$

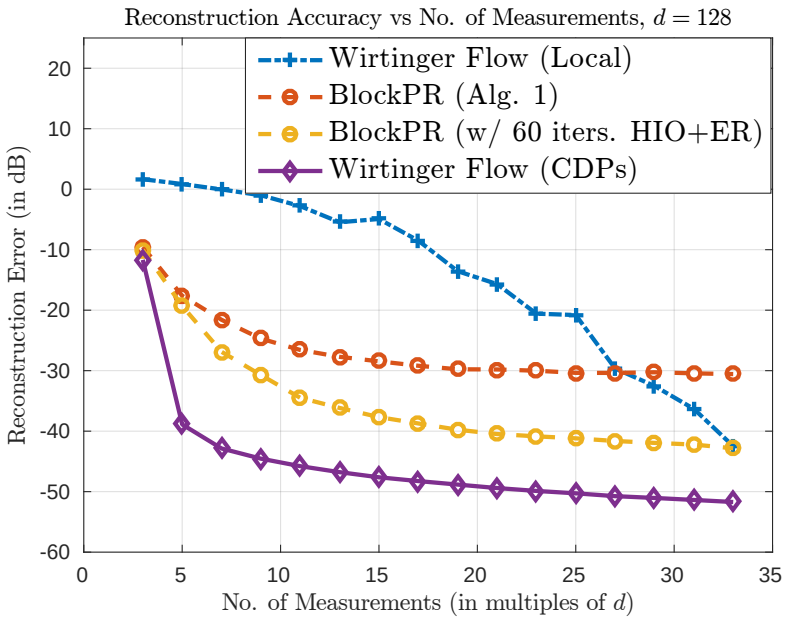
Computing the leading eigenvectors of $\tilde{D}_{j'}$ for all $j' \in [d/s]$ will then produce (multiple) estimates of each magnitude $|(x_0)_j|$ which can then be combined as desired to produce our final magnitude estimates. As we shall see below, one can achieve better numerical robustness to noise using this technique than what can be achieved using the simpler magnitude estimation technique presented in line 4 of Algorithm 1.

3.6.2 Experiments with Measurements from Example 2 of Section 3.2

We begin by presenting results in Fig. 3.2a demonstrating the improved noise robustness of the proposed method over the formulation in [57]. Recall that [57] uses a greedy angular synchronization method instead of the eigenvector-based procedure analyzed in this paper. Fig. 3.2a plots the reconstruction error when recovering a $d = 128$ length complex



(a) Improved Robustness to Measurement Noise – Comparing Variants of the *BlockPR* algorithm



(b) Reconstruction Error vs. No. of Measurements; (Reconstruction at 40dB SNR)

Figure 3.2: Robust Phase Retrieval – Local vs. Global Measurements

Gaussian test signal using $D = \lceil 4d \log_2 d \rceil$ measurements at different added noise levels. As mentioned above, the local correlation measurements described in Example 2 of Section 3.2 are utilized in this plot and in all the ensuing experiments in this subsection unless otherwise indicated. Three variants of the proposed algorithm are plotted in Fig. 3.2a:

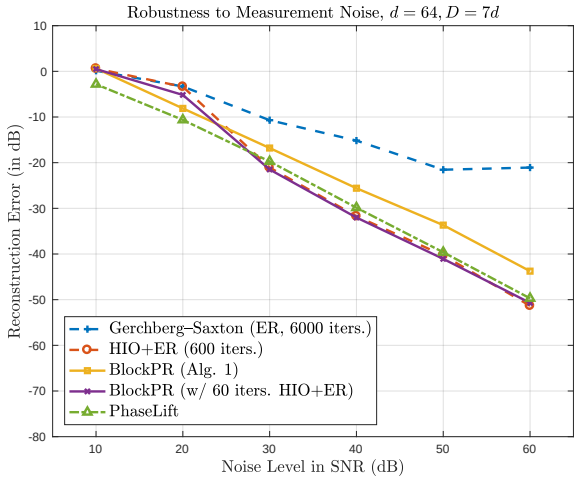
1. an implementation of Algorithm 1 (denoted by $\square$'s),

2. an implementation of Algorithm 1 post-processed using 60 iterations of the Hybrid Input–Output (HIO) and Error Reduction (ER) algorithms (implemented in two successive sets, with each set consisting of 20 iterations of HIO followed by 10 ER iterations; denoted by $\circ$'s), and

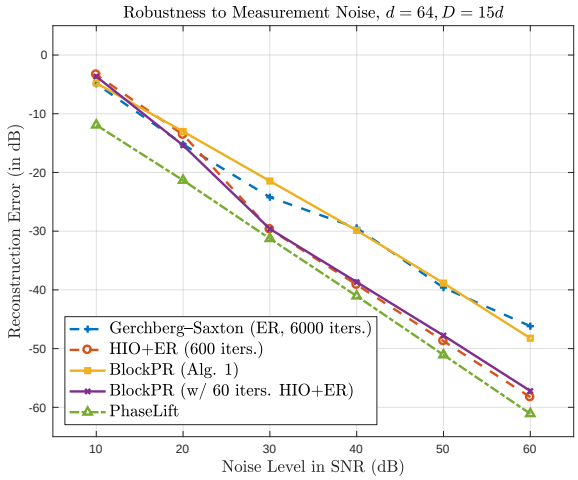
3. the algorithmic implementation from [57] (denoted by +'s).

We see that the eigenvector-based angular synchronization method proposed in this paper provides more accurate reconstructions – especially at low SNRs – over the greedy angular synchronization of [57]. Moreover, post-processing using the HIO+ER algorithms as detailed above yields a significant improvement in reconstruction errors over the two other variants. For reference, we also include reconstruction errors with the *Wirtinger Flow* algorithm (denoted by $\lozenge$'s) when using (*global*) coded diffraction pattern (CDP) measurements. Specifically, we use $2\delta - 1$ (with $\delta = 2 \log_2 d = 14$) octanary modulations/codes as described in [21, §1.5, (1.9)] to construct the CDP measurements. Clearly, using global measurements such as coded diffraction patterns provides superior noise tolerance; however, they are not applicable to the local measurement model considered here. Indeed, when the *Wirtinger Flow* algorithm is used with local measurements such as those described in this paper, the noise tolerance significantly deteriorates. Fig. 3.2b illustrates this phenomenon by plotting the reconstruction error in recovering a $d = 128$ length complex Gaussian test signal at 40 dB SNR when using different numbers of measurements, D. *Wirtinger flow*, for example, requires a large number of local measurements before returning accurate reconstructions. The wide disparity in reconstruction accuracy between local and global measurements for *Wirtinger Flow* illustrates the significant challenge in phase retrieval from local measurements. Furthermore, we see that the *BlockPR* method proposed in this paper is more noise

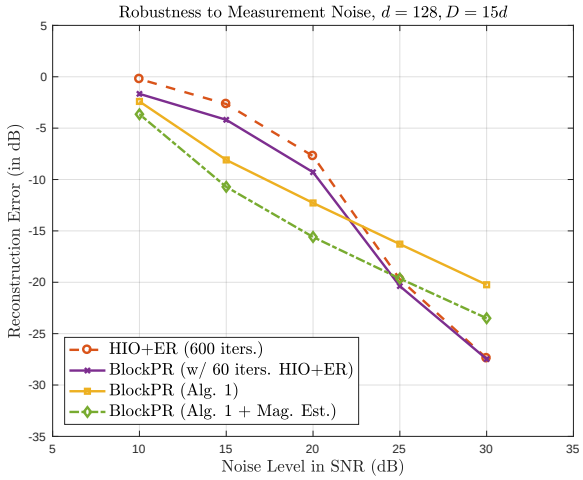
tolerant than *Wirtinger Flow* for local measurements.



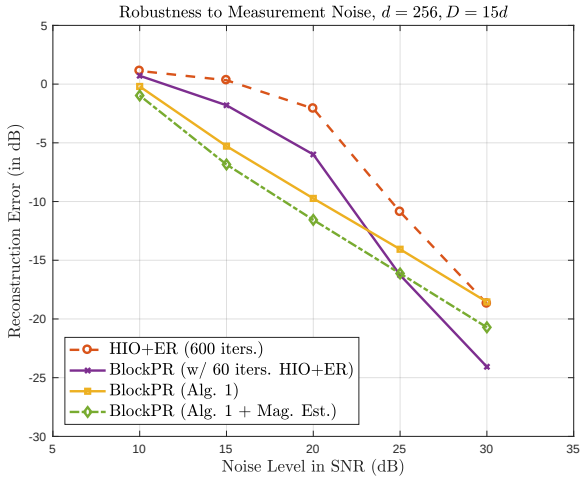
(a) Using $D = 7d$ measurements.



(b) Using $D = 15d$ measurements.



(c) Low SNR Simulations: Using $D = 15d$ measurements, Problem Size $d = 128$



(d) Low SNR Simulations: Using $D = 15d$ measurements, Problem Size $d = 256$

Figure 3.3: Robustness to measurement noise – Phase Retrieval from deterministic local correlation measurements.

Given the weaker performance of *Wirtinger Flow* with local measurements, we now restrict our attention to the empirical evaluation of the proposed method (Alg. 1, as well as the post-processed variant (with HIO+ER iterations)) against *PhaseLift* and alternating projection algorithms. Although numerical simulations suggest that these methods work with local measurements, we note that (to the best of our knowledge) there are no theoretical recovery or robustness guarantees for these methods and measurements. The *PhaseLift*

algorithm was implemented as a trace regularized least-squares problem using CVX [48, 49] – a package for specifying and solving convex programs in Matlab. We consider two variants from the family of alternating projection methods – *Gerchberg–Saxton* (sometimes referred to as the Error Reduction (ER) algorithm) and Hybrid Input-Output (HIO). For both algorithms, the following two projections were utilized: (i) projection onto the measured magnitudes, and (ii) projection onto the span of the measurement vectors $\{\mathbf{a}_j\}_{j=1}^{D}$. This formulation as well as other details and connections to convex optimization theory can be found in [12]. For the ER and HIO implementations, the initial guess was set to be the all-zero vector.[11] For the HIO implementation, as is popular practice (see, for example, [40]) every few (20) HIO iterations were followed by a small number of (10) ER iterations, with the maximum number of HIO+ER iterations limited to 600 – this choice of iteration count ensures convergence of the algorithm (see Fig. 3.4) while comparing favorably with the computational cost (see Fig. 3.5b) of the proposed *BlockPR* method. For the ER implementation, $6,000$ iterations were necessary to ensure convergence.

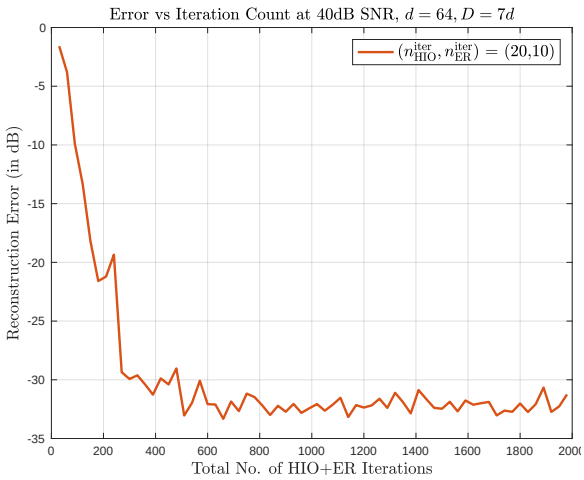


Figure 3.4: Reconstruction Error vs. Iteration Count for HIO+ER Implementation

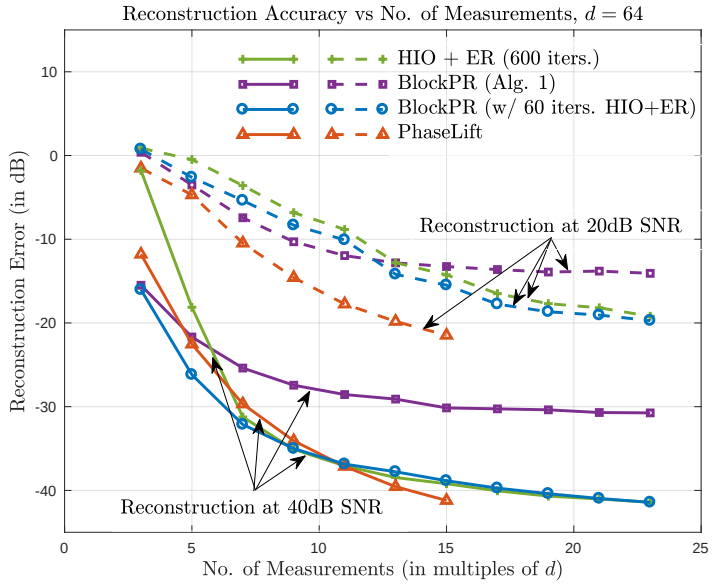
We begin by presenting numerical results evaluating the robustness to measurement noise. Figs. 3.3a and 3.3b plot the error in reconstructing a $d = 64$ length complex vector $\mathbf{x}_0$ using $D = 7d$ and $D = 15d$ local correlation-based phaseless measurements respectively.

[11] We note that using a random starting guess does not change the qualitative nature of the empirical results.

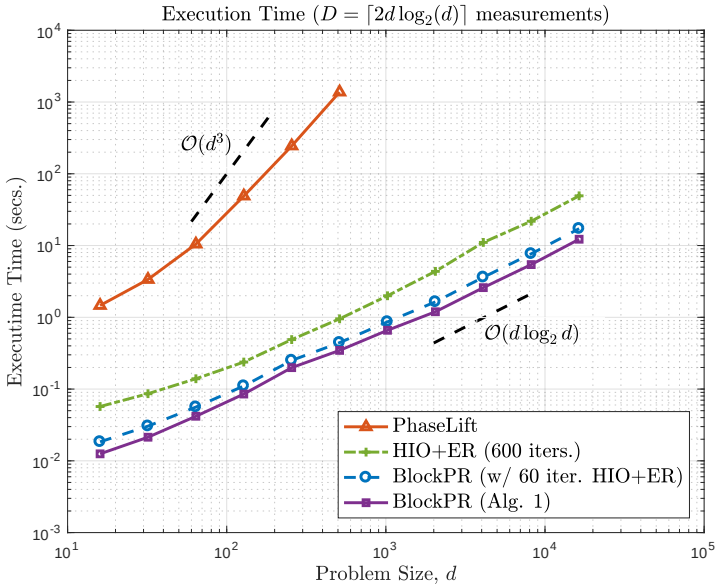
Note that this corresponds to using $\delta = 4$ (and the associated $2\delta - 1 = 7$ masks) and $\delta = 8$ (and the corresponding $2\delta - 1 = 15$ masks) respectively. Moreover, the well-conditioned *deterministic* (and sparse) measurement construction defined in Example 2 of Section 3.2 was utilized along with additive Gaussian measurement noise. We see from Fig. 3.3 that the method proposed in this paper (denoted *BlockPR* in the figure) performs reliably across a wide range of SNRs and compares favorably against existing popular phase retrieval algorithms. When using a small number of measurements (as in Fig. 3.3a, and modeling real-world situations), both variants of the proposed method – Alg. 1, as well as Alg. 1 post-processed using 60 HIO+ER iterations – outperform the ER algorithm by significant margins and compare well with the popular HIO+ER algorithm. When more measurements are available (as in Fig. 3.3b), the performance of the ER and HIO+ER algorithms approaches the performance of the *BlockPR* variants proposed in this paper. In addition, the proposed methods also compare well with the significantly more expensive *PhaseLift* reconstructions. We emphasize that the superior performance of the proposed methods demonstrated here comes with rigorous theoretical recovery guarantees for local measurements – something that cannot be said of any of the other methods in Fig. 3.3.

Additionally, Figs. 3.3c and 3.3d compare the performance of the HIO+ER algorithm at low SNRs for problems sizes $d = 128$ and $d = 256$ respectively, with the various *BlockPR* implementations – including one with the improved magnitude estimation procedure detailed in Section 3.6.1 (with $s = 1$ and using the average of the obtained $\tilde{D}_{j\prime}$ block magnitude estimates). These figures demonstrate the value of the magnitude estimation procedure from Section 3.6.1 at low SNRs over the HIO+ER post-processing method utilized in the other figures (and over the HIO+ER algorithm); we defer a more detailed study of this to future work.

Next, Fig. 3.5a plots the reconstruction error in recovering a $d = 64$-length complex vector as a function of the number of measurements used. This corresponds to using values of δ ranging from 2 to 12 (and the associated $2\delta - 1$ masks). As with Fig. 3.3, the deterministic correlation-based measurement constructions of Section 3.2 (Example construction 2) were utilized along with an additive Gaussian noise model. Plots are provided for simulations at



(a) Reconstruction Error vs. No. of Measurements



(b) Execution Time vs. Problem Size

Figure 3.5: Performance Evaluation and Comparison of the Proposed Phase Retrieval Method (with Deterministic Local Correlation Measurements of Example 2, Section 3.2 and Additive Gaussian Noise)

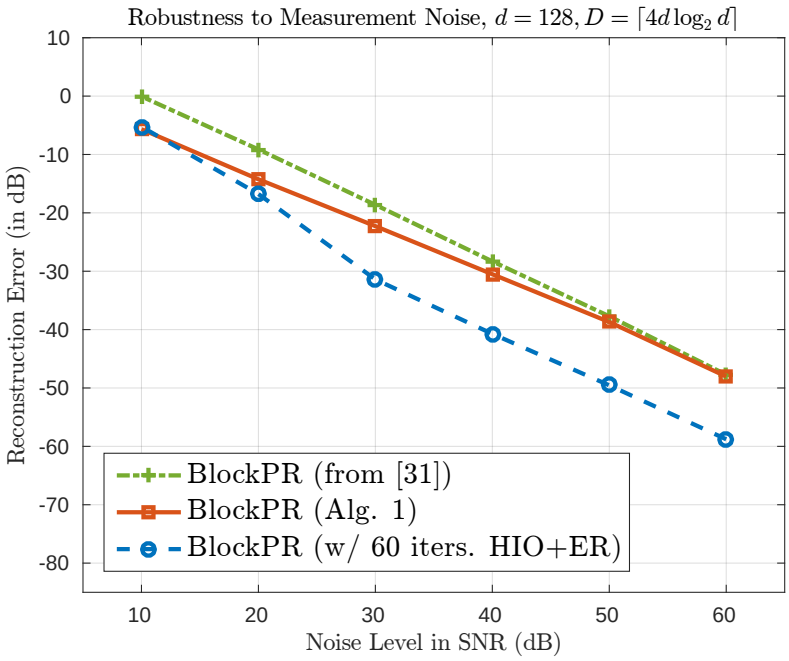
two noise levels – 20 dB and 40 dB. Comparing the performance of the two *BlockPR* variants, we observe that the HIO+ER post-processing procedure provides improved reconstruction errors – with the margin of improvement increasing when more measurements are available. We also notice that both variants of *BlockPR* compare particularly well with the other algorithms (HIO+ER and *PhaseLift*) when small numbers of measurements are available.

Finally, Fig. 3.5b plots the average execution time (in seconds) required to solve the phase retrieval problem using $D = \lceil 2d \log_2 d \rceil$ measurements. For comparison, execution times for the *PhaseLift* and HIO+ER alternating projection algorithms are provided. We observe that both variants of the proposed method are several orders of magnitude faster than the *PhaseLift* algorithm,[12] and between 2–5 times faster than the HIO+ER implementation. Moreover, the plot illustrates the essentially FFT-time computational complexity (see Section 3.1.3) of the proposed method. Between the two *BlockPR* variants, we see that there is a small trade-off between reduced execution time and improved accuracy; one of the two variants may be more appropriate depending on the application requirements.

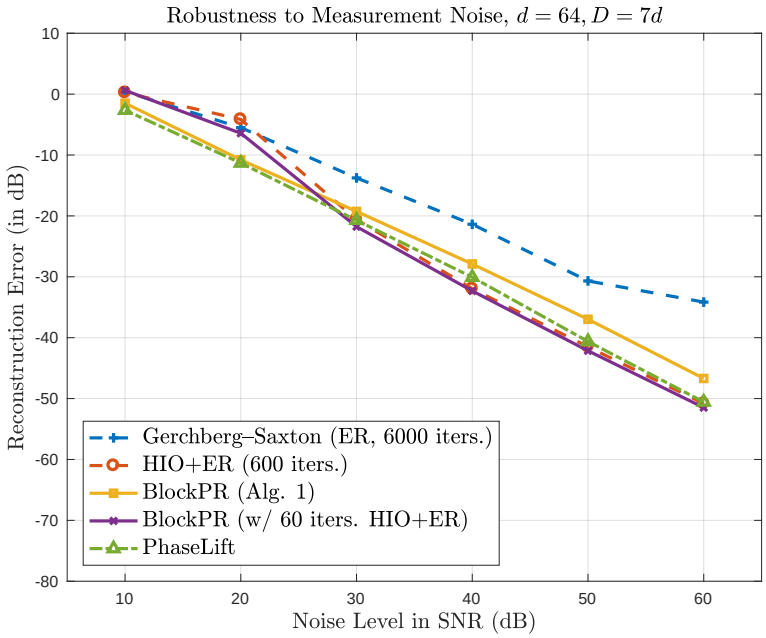
3.6.3 Experiments with Ptychographic Measurements from Example 1 of Section 3.2

We now present a selection of empirical results demonstrating the accuracy, robustness, and efficiency of the proposed method when using the ptychographic measurements from Example 1, Section 3.2 and observe that, in this case, the comparison between the proposed algorithm and existing methods is similar to that in section Section 3.6.2. Recall that (see Example 1 from Section 3.2 and Section 3.1.4 for details) this measurement construction corresponds to a discretization of the ptychographic measurements $\left|\mathcal{F}[\widetilde{h} \cdot S_l f](\omega)\right|^2$ as per (3.11), where f is the unknown specimen and $\widetilde{h}(t) = \frac{e^{-t/a}}{\sqrt[4]{2\delta - 1}}$ denotes the *single, deterministic* illumination function (or mask). As before, δ defines the local support of the illumination function and we consider a discretization involving $2\delta - 1$ Fourier modes and

[12] For computational efficiency and due to memory constraints, the *PhaseLift* plot in Fig. 3.5b was generated using the TFOCS software package (http://cvxr.com/tfocs/) instead of the more computationally expensive CVX software package.



(a) Improved Robustness to Measurement Noise – Comparing Variants of the *BlockPR* algorithm



(b) Phase Retrieval from $D = 7d$ Measurements – Comparison with Other Phase Retrieval Algorithms

Figure 3.6: Numerical Evaluation of the Proposed Algorithm with the Ptychographic Measurements from Example 1 of Section 3.2

sample/specimen shifts of 1 unit or pixel (yielding a total of d shifts in the discrete problem formulation).

We begin with Fig. 3.6a, which demonstrates the relative performance of the *BlockPR* variants described in [57] and this paper. More specifically, we plot the error in reconstructing a $d = 128$ length complex Gaussian test signal using $D = \lceil 4d \log_2 d \rceil$ measurements at different added noise levels. As with Fig. 3.2a, we plot results for three different variants of the *BlockPR* algorithm: (i) Algorithm 1, (ii) Algorithm 1 with the HIO+ER post-processing procedure as described in Section 3.6.2, and (iii) the implementation from [57]. We again observe (as in Fig. 3.2a) that the methods described in this paper (which use eigenvector-based angular synchronization) are more accurate than that detailed in [57] (which uses a greedy angular synchronization method), with the improvement in performance being especially significant at low SNRs. Next, Fig. 3.6b studies the performance of the proposed method(s) against three other popular phase retrieval algorithms – *PhaseLift*, the *Gerchberg–Saxton* Error Reduction (ER) algorithm, and the Hybrid Input–Output (HIO) algorithm (with implementation parameters identical to those in Section 3.6.2). As with Fig. 3.3a, we consider the reconstruction of a $d = 64$ length complex vector x_0 using $D = 7d$ measurements in the presence of additive Gaussian noise. We observe that the methods proposed in this paper outperform the ER algorithm and compare very well with the HIO+ER and (the significantly more expensive) *PhaseLift* algorithms across a wide range of SNRs. Finally, we note that the execution time plot for this ptychographic measurement construction is qualitatively and quantitatively similar to Fig. 3.5b – the proposed methods are faster (by a factor of 2–5) than an equivalent HIO+ER implementation and orders of magnitude faster than convex optimization approaches such as *PhaseLift*.

3.7 Concluding Remarks

In this paper new and improved deterministic robust recovery guarantees are proven for the phase retrieval problem using local correlation measurements. In addition, a new practical phase retrieval algorithm is presented which is both faster and more noise robust

than previously existing approaches (e.g., alternating projections) for such local measurements.

Future work might include the exploration of more general classes of measurements which are guaranteed to lead to well conditioned linear systems of the type used to reconstruct $X \approx X_0$ in line 1 of Algorithm 1. Currently two deterministic measurement constructions are known (recall, e.g., Section 3.2) – it should certainly be possible to construct more general families of such measurements.

Other interesting avenues of inquiry include the theoretical analysis of the magnitude estimate approach proposed in Section 3.6.1 in combination with the rest of Algorithm 1. Alternate phase retrieval approaches might also be developed by using such local block eigenvector-based methods for estimating phases too, instead of just using the single global top eigenvector as currently done in line 3 of Algorithm 1.

Finally, more specific analysis of the performance of the proposed methods using masked/windowed Fourier measurements (recall Section 3.1.5) would also be interesting. In particular, an analysis of the performance of such approaches as a function of the bandwidth of the measurement mask/window could be particularly enlightening. One might also consider extending the discrete results of this paper to the analytic setting by, e.g., expanding on [72].

3.8 Alternate Perturbation Bounds

In this section we present a simpler (and easier to derive), albeit weaker, perturbation result in the spirit of Section 3.4, which is associated with the analysis of line 3 of Algorithm 1. Specifically, we will derive an upper bound on $\min_{\theta \in [0,2\pi]} \|\widetilde{\mathbf{x}}_0 - \mathrm{e}^{\mathrm{i}\theta}\widetilde{\mathbf{x}}\|_2$ (provided by Theorem 6), which scales like d^3. While this dependence is strictly worse than the one derived in Section 3.4, it is easier to obtain and the technique may be of independent interest.

We will begin with a result concerning the top eigenvector of any Hermitian matrix.

Lemma 8. *Let $X_0 = \sum_{j=1}^{d} \nu_j \mathbf{x}_j \mathbf{x}_j^*$ be Hermitian with eigenvalues $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d$ and orthonormal eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_d \in \mathbb{C}^d$. Suppose that $X = \sum_{j=1}^{d} \lambda_j \mathbf{v}_j \mathbf{v}_j^*$ is Hermitian with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_d \in \mathbb{C}^d$, and $\|X - X_0\|_{\mathrm{F}} \leq \eta \|X_0\|_F$ for some $\eta \geq 0$. Then,*

$$\left(1 - |\langle \mathbf{x}_1, \mathbf{v}_1 \rangle|^2\right) \leq \frac{4\eta^2 \|X_0\|_{\mathrm{F}}^2}{(\nu_1 - \nu_2)^2}.$$

Proof. An application of the $\sin\theta$ theorem [27, 88] (see, e.g., the proof of Corollary 1 in [106]) tells us that

$$\sin\left(\arccos\left(|\langle \mathbf{x}_1, \mathbf{v}_1 \rangle|\right)\right) \leq \frac{2\eta \|X_0\|_F}{|\nu_1 - \nu_2|}.$$

Squaring both sides we then learn that

$$\left(1 - |\langle \mathbf{x}_1, \mathbf{v}_1 \rangle|^2\right) = \sin^2\left(\arccos\left(|\langle \mathbf{x}_1, \mathbf{v}_1 \rangle|\right)\right) \leq \frac{4\eta^2 \|X_0\|_F^2}{(\nu_1 - \nu_2)^2}, \tag{3.37}$$

giving us the desired inequality. □

The following variant of Lemma 8 concerning rank 1 matrices X_0 is of use in the analysis of many other phase retrieval methods, and can be used, e.g., to correct and simplify the proof of equation (1.8) in Theorem 1.3 of [19].

Lemma 9. *Let $\mathbf{x}_0 \in \mathbb{C}^d$, set $X_0 = \mathbf{x}_0 \mathbf{x}_0^*$, and let $X \in \mathbb{C}^{d \times d}$ be Hermitian with $\|X - X_0\|_F \leq \eta \|X_0\|_F = \eta \|\mathbf{x}_0\|_2^2$ for some $\eta \geq 0$. Furthermore, let λ_i be the i-th largest magnitude eigenvalue of X and $\mathbf{v}_i \in \mathbb{C}^d$ an associated eigenvector, such that the $\mathbf{v}_i$ form an orthonormal eigenbasis. Then*

$$\min_{\theta \in [0, 2\pi]} \|e^{\mathbb{i}\theta} \mathbf{x}_0 - \sqrt{|\lambda_1|} \mathbf{v}_1\|_2 \leq (1 + 2\sqrt{2}) \eta \|\mathbf{x}_0\|_2.$$[13]

Proof. In this special case of Lemma 8 we have $\nu_1 = \|X_0\|_F = \|\mathbf{x}_0\|_2^2$ and $\mathbf{x}_1 := \mathbf{x}_0 / \|\mathbf{x}_0\|$.

[13] *It is interesting to note that similar bounds can also be obtained using simpler techniques (see, e.g., [56]).*

Choose $\phi \in [0, 2\pi]$ such that $\langle e^{\mathbb{i}\phi}\mathbf{x}_0, \mathbf{v}_1\rangle = |\langle \mathbf{x}_0, \mathbf{v}_1\rangle|$. Then,

$$\begin{aligned}\|e^{\mathbb{i}\phi}\mathbf{x}_0 - \sqrt{\nu_1}\mathbf{v}_1\|_2^2 &= 2\nu_1 - 2\nu_1 \cdot |\langle \mathbf{x}_0/\|\mathbf{x}_0\|, \mathbf{v}_1\rangle| = 2\nu_1 - 2\nu_1 \cdot |\langle \mathbf{x}_1, \mathbf{v}_1\rangle| \\ &\leq 2\nu_1 \left(1 - |\langle \mathbf{x}_1, \mathbf{v}_1\rangle|\right)\left(1 + |\langle \mathbf{x}_1, \mathbf{v}_1\rangle|\right) \\ &= 2\nu_1 \left(1 - |\langle \mathbf{x}_1, \mathbf{v}_1\rangle|^2\right) \leq 8\eta^2 \|X_0\|_{\mathrm{F}}\end{aligned} \tag{3.38}$$

where the last inequality follows from Lemma 8 with $\nu_1 = \|X_0\|_F = \|\mathbf{x}_0\|_2^2$. Finally, by the triangle inequality, Weyl's inequality (see, e.g., [55]), and (3.38), we have

$$\begin{aligned}\|e^{\mathbb{i}\phi}\mathbf{x}_0 - \sqrt{|\lambda_1|}\mathbf{v}_1\|_2 &\leq \|e^{\mathbb{i}\phi}\mathbf{x}_0 - \sqrt{\nu}_1\mathbf{v}_1\|_2 + \|\sqrt{\nu}_1\mathbf{v}_1 - \sqrt{|\lambda_1|}\mathbf{v}_1\|_2 \\ &\leq 2\sqrt{2}\cdot\eta\sqrt{\nu}_1 + \left|\sqrt{\nu}_1 - \sqrt{|\lambda_1|}\right| \\ &\leq 2\sqrt{2}\cdot\eta\sqrt{\nu}_1 + \frac{|\nu_1 - \lambda_1|}{\sqrt{\nu}_1 + \sqrt{|\lambda_1|}} \\ &\leq 2\sqrt{2}\cdot\eta\sqrt{\nu}_1 + \frac{\eta\nu_1}{\sqrt{\nu}_1 + \sqrt{|\lambda_1|}} \\ &\leq (1 + 2\sqrt{2})\eta\sqrt{\nu}_1.\end{aligned}$$

The desired result now follows. □

We may now use Lemma 8 to produce a perturbation bound for our banded matrix of phase differences $\tilde{X}_0$ from (3.7).

Theorem 6. *Let $\widetilde{X}_0 = T_\delta(\widetilde{\mathbf{x}}_0\widetilde{\mathbf{x}}_0^*)$ where $|(\widetilde{\mathbf{x}}_0)_i| = 1$ for each i. Further suppose $\widetilde{X} \in T_\delta(\mathcal{H}^d)$ has $\widetilde{\mathbf{x}}$ as its top eigenvector, where $||\widetilde{\mathbf{x}}||_2 = \sqrt{d}$. Suppose that $\|\widetilde{X}_0 - \widetilde{X}\|_F \leq \eta\|\widetilde{X}_0\|_F$ for some $\eta > 0$. Then, there exists an absolute constant $C \in \mathbb{R}^+$ such that*

$$\min_{\theta\in[0,2\pi]} \|\widetilde{\mathbf{x}}_0 - e^{\mathbb{i}\theta}\widetilde{\mathbf{x}}\|_2 \leq C\frac{\eta d^3}{\delta^{\frac{5}{2}}}.$$

Proof. Recall that the phase vectors $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}_0$ are normalized so that $\|\tilde{\mathbf{x}}\|_2 = \|\tilde{\mathbf{x}}_0\|_2 = \sqrt{d}$. Combining Lemmas 2 and 8 after noting that $\|\tilde{X}_0\|_{\mathrm{F}}^2 = d(2\delta - 1)$ we learn that

$$\left(1 - \frac{1}{d^2}|\langle \tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}\rangle|^2\right) \leq C'\eta^2\left(\frac{d}{\delta}\right)^5 \tag{3.39}$$

for an absolute constant $C' \in \mathbb{R}^+$. Let $\phi \in [0, 2\pi)$ be such that $\operatorname{Re}\left(\langle \tilde{\mathbf{x}}_0, \mathbb{e}^{\mathbb{i}\phi}\tilde{\mathbf{x}}\rangle\right) = |\langle \tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}\rangle|$. Then,

$$\begin{aligned}\|\tilde{\mathbf{x}}_0 - \mathbb{e}^{\mathbb{i}\phi}\tilde{\mathbf{x}}\|_2^2 &= 2d - 2\operatorname{Re}\left(\langle \tilde{\mathbf{x}}_0, \mathbb{e}^{\mathbb{i}\phi}\tilde{\mathbf{x}}\rangle\right) \\ &= 2d\left(1 - \frac{1}{d}\left|\langle \tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}\rangle\right|\right) \leq 2d\left(1 - \frac{1}{d^2}\left|\langle \tilde{\mathbf{x}}_0, \tilde{\mathbf{x}}\rangle\right|^2\right).\end{aligned}$$

Combining this last inequality with (3.39) concludes the proof. □

Acknowledgements

The authors would like to thank Felix Krahmer for helpful discussions regarding Lemma 9. MI and RS would like to thank the Hausdorff Institute of Mathematics, Bonn for its hospitality during its Mathematics of Signal Processing Trimester Program. A portion of this work was completed during that time. In addition, the authors would like to thank and acknowledge the Institute for Mathematics and its Applications (IMA) for the workshop "Phaseless Imaging in Theory and Practice: Realistic Models, Fast Algorithms, and Recovery Guarantees" hosted there in a August of 2017. The revision of this paper benefitted greatly from many discussions with the participants there. In particular, conversations with James Fienup, Andreas Menzel, and Irene Waldspurger were exceedingly helpful. MI was supported in part by NSF DMS-1416752. RS was supported in part by a Hellman Fellowship and NSF DMS-1517204.

Acknowledgement of joint authorship

Sections 3.1–3.8, in full, are a reprint of the material accepted for publication by Applied Computational and Harmonic Analysis. Iwen, Mark A.; Preskitt, Brian; Saab, Rayan; Viswanathan, Aditya. *Phase retrieval from local measurements: Improved robustness via eigenvector-based angular synchronization.* Available online as of 18 June 2018.

Chapter 4

Invertible Local Measurement Systems

4.1 Introduction

In Chapter 3, we proposed the algorithm for robustly solving phase retrieval whose analysis and extension forms the basis of this dissertation. This algorithm, stated in Algorithm 1, operates in two steps: the first step is to solve a linear system, obtained by recasting the magnitude-only linear measurements $y_{(\ell,j)} = |\langle x_0, S^\ell m_j\rangle|^2 + \eta_{j\ell}$ as linear measurements on the space of Hermitian matrices, by rewriting

$$|\langle x_0, S^\ell m_j\rangle|^2 = \mathrm{Tr}\left(S^\ell m_j m_j^* S^{-\ell} x_0 x_0^*\right) = \langle S^\ell m_j m_j^* S^{-\ell}, x_0 x_0^*\rangle.$$

By design, it is clear that the "vectors" $\{S^\ell m_j m_j^* S^{-\ell}\}$ are all contained in the subspace $T_\delta(\mathcal{H}^d)$ of $\mathcal{H}^d$ (defined in (3.3)), where d is the ambient dimension (meaning $x_0, m_j \in \mathbb{C}^d$) and δ is the support size of the masks (so $\mathrm{supp}(m_j) \subseteq [\delta]$). Hence, we used our measurements to solve for (an estimate of) the projection of $x_0 x_0^*$ onto $T_\delta(\mathcal{H}^d)$, namely $T_\delta(x_0 x_0^* + N) =: X$, where N is some perturbation.

After the linear step, the second step is recovering x_0 from X. We accomplish this by inferring the magnitudes of x_0's entries from the main diagonal of X (as the main diagonal is preserved by the projection T_δ) and finding their relative phases through an angular synchronization problem (see Section 3.4 and Chapter 5).

The subject of this chapter is to study in more detail the linear system solved in the first step. Most crucially, we consider that, in the work of Chapter 3, we only produced two examples of collections of vectors $\{m_j\}_{j\in[2\delta-1]} \subseteq \mathbb{C}^d$ such that this linear system was invertible. Considering that one of the major contributions of our algorithm is that it admits measurement models that intend to replicate laboratory conditions, our knowledge of which vectors are compatible with our algorithm and the theory built for it is critical in promoting the applicability of this work. Therefore, in Section 4.2, we make a study of the conditioning of the linear system in (3.5) as a function of the set of masks $\{m_j\}_{j=1}^D$, which we state and prove in Theorem 7. As a consequence, this result also gives us a complete characterization of all sets of masks $\{m_j\}_{j=1}^D$ that are capable of spanning $T_\delta(\mathcal{H}^d)$, in the sense that we have a checkable condition that indicates whether the linear system of (3.5) is invertible.

We consider the act of inverting $\mathcal{A}$ from a practical perspective in Section 4.4. We write its inverse explicitly and analyze the runtime of calculating $\mathcal{A}^{-1}(y)$ in Section 4.4.1. In Section 4.4.3, we analyze the variance of the individual entries of $\mathcal{A}^{-1}(y)$ when the measurements are exposed to uniform, Gaussian noise. In Section 4.3, we provide a few examples of explicit $\gamma \in \mathbb{R}^d$ that are proven to satisfy the conditions to span $T_\delta(\mathcal{H}^d)$, and we illustrate the results of this chapter with numerical studies.

4.1.1 Definitions and Notation

Before we begin these analyses, we introduce and unify some definitions and notational elements that will make discussion of the mathematical details more convenient. We say that $\{m_j\}_{j=1}^D \subseteq \mathbb{C}^d$ is a *local measurement system* or *family of masks* of support δ if $1 \in \text{supp}(m_j)$ and $\text{supp}(m_j) \subseteq [\delta]$ for each j.

If we further have that each m_j satisfies $m_j = \mathcal{R}_d(\sqrt{K} f_j^K) \circ \gamma$ for some $K \geq \max(\delta, D), \gamma \in \mathbb{C}^d$ satisfying $\mathrm{supp}(\gamma) = [\delta]$, then we call $\{m_j\}_{j=1}^D$ a *local Fourier measurement system* of support δ with mask γ and modulation index K. If $K = D = 2\delta - 1$, then we simply refer to $\{m_j\}_{j=1}^D$ as a *local Fourier measurement system* of support δ with mask γ. We add that, if we say that $\{m_j\}_{j=1}^D$ is a local Fourier measurement system with support δ and mask γ, this implies an assertion that $\mathrm{supp}(\gamma) = [\delta]$.

Given a local measurement system $\{m_j\}_{j=1}^D$ in $\mathbb{C}^d$, the associated *lifted measurement system* is the set $\mathcal{L}_{\{m_j\}} = \{S^\ell m_j m_j^* S^{-\ell}\}_{(\ell,j)\in[d]_0\times[D]} \subseteq \mathbb{C}^{d\times d}$. We then say that a family of masks $\{m_j\}_{j=1}^D \subseteq \mathbb{C}^d$ of support δ is a *spanning family* if $\mathrm{span}_{\mathbb{R}} \mathcal{L}_{\{m_j\}} = T_\delta(\mathcal{H}^d)$.

The *measurement operator* associated with a local measurement system is the operator

$$\begin{aligned} \mathcal{A} &: T_\delta(\mathbb{C}^{d\times d}) \to \mathbb{C}^{[d]_0\times[D]} \\ \mathcal{A}(X)_{(\ell,j)} &= \langle S^\ell m_j m_j^* S^{-\ell}, X\rangle. \end{aligned} \tag{4.1}$$

The *canonical matrix representation* of $\mathcal{A}$ is the matrix $A \in \mathbb{C}^{dD\times d(2\delta-1)}$, defined by

$$\left(A \begin{bmatrix} \mathrm{diag}(X, 1-\delta) \\ \vdots \\ \mathrm{diag}(X, \delta-1) \end{bmatrix} \right)_{(j-1)d+\ell} = \mathcal{A}(X)_{(\ell-1,j)}. \tag{4.2}$$

For convenience, we define the *diagonal vectorization operator* $\mathcal{D}_I : \mathbb{C}^{d\times d} \to \mathbb{C}^{|I|d}$ for any

subset $\{\ell_i\}_{i=1}^{|I|} = I \subseteq [d]$ and $\mathcal{D}_k : \mathbb{C}^{d\times d} \to \mathbb{C}^{(2k-1)d}$ for any integer $k \leq \frac{d+1}{2}$ by

$$\mathcal{D}_I(X) = \begin{bmatrix} \operatorname{diag}(X, \ell_1) \\ \vdots \\ \operatorname{diag}(X, \ell_{|I|}) \end{bmatrix} \tag{4.3}$$

$$\mathcal{D}_k(X) = \mathcal{D}_{[2k-1]_{1-k}}(X) = \begin{bmatrix} \operatorname{diag}(X, 1-k) \\ \vdots \\ \operatorname{diag}(X, k-1) \end{bmatrix}, \tag{4.4}$$

so that (4.2) becomes $A\mathcal{D}_\delta(X)_{(j-1)d+\ell} = \mathcal{A}(X)_{(\ell-1,j)}$. We remark that, when $2k-1 \leq d$, $\mathcal{D}_k$ is invertible on $T_k(\mathbb{C}^{d\times d})$, and for $v \in \mathbb{C}^{d(2k-1)}$, we use $\mathcal{D}_k^{-1}(v)$ or $\mathcal{D}_k^*(v)$ to represent the matrix in $T_k(\mathbb{C}^{d\times d})$ whose diagonals are given by the $2k-1$ distinct d-length blocks of v.

4.2 Condition Number

In this section, we calculate the singular values, and therefore the condition number, of the measurement operator $\mathcal{A}$ for an arbitrary local measurement system $\{m_j\}_{j=1}^d$ in Theorem 7, the main result of this section. We remark that this is an important element in using the error bounds proven in Section 3.5, most notably Theorem 5 and Corollary 3, since they all unavoidably rely on the condition number κ and singular values of the linear system solved in line 1 of Algorithm 1, and leaving this quantity unknown and unestimated renders these bounds far less useful. We also remark that, in this section, as in Sections 4.3 and 4.5, we focus on calculations of the condition numbers κ of the linear systems at hand, since κ is scale invariant and reflects more completely on the strength of a local measurement system than does $\sigma_{\min}^{-1}$. Indeed, if $\{m_j\}_{j\in[D]}$ is a local measurement system with $\sigma_{\min}^{-1} = s$, then $\{tm_j\}_{j\in[D]}$ has $\sigma_{\min}^{-1} = \frac{s}{t^2}$, so it would appear that simply making t large could arbitrarily improve the recovery guarantees of Section 3.5 which refer to $\sigma_{\min}^{-1}$; of course, this action would correspond to simply multiplying the observed measurements by the scalar t^2, and slips in its advantage by ignoring that $\|n\|_2$ would also scale as t^2 in such a case. This

process clearly buys us nothing, so the $\sigma_{\min}^{-1}$ inequalities fail, in this way, to consider a sense of scale between $\|n\|_2$ and $\|\mathcal{A}(xx^*)\|_2$. Therefore, by referencing κ and $\text{SNR} = \frac{\|\mathcal{A}(xx^*)\|}{\|n\|_2}$ in the alternative statements of Theorem 5 and Corollary 3, we have inequalities that more accurately – if less tightly, by always aggravating the estimate with a factor of $\frac{\sigma_{\max}\|xx^*\|_F}{\|\mathcal{A}(xx^*)\|_2}$ – describe the relationship between the design of the linear system and the accuracy of the estimate produced by Algorithm 1.

We emphasize further that, perhaps equally or even more significantly, this result gives a complete characterization of all local measurement systems that are usable for phase retrieval in Algorithm 1, since we may simply check whether $\{m_j\}_{j\in[D]}$ leads to a system with any singular values of 0. This is an important contribution to our understanding of the work presented in Chapter 3, since previously we only possessed two examples of families of masks that produced invertible linear systems.

Theorem 7. *Given a family of masks $\{m_j\}_{j\in[D]}$ of support $\delta \leq \frac{d+1}{2}$, we define $g_m^j = \text{diag}(m_j m_j^*, m)$,*

$$H = P^{(d,D)} \begin{bmatrix} R\overline{g}_{1-\delta}^1 & \cdots & R\overline{g}_{\delta-1}^1 \\ \vdots & \ddots & \vdots \\ R\overline{g}_{1-\delta}^D & \cdots & R\overline{g}_{\delta-1}^D \end{bmatrix},$$

and $M_j = \sqrt{d}\left(f_j^d \otimes I_D\right)^* H$. Then the singular values of $\mathcal{A}$ as defined in (4.1) are $\{\sigma_i(M_j)\}_{(i,j)\in[D]\times[d]}$ and its condition number is

$$\kappa(\mathcal{A}) = \frac{\max\limits_{i\in[d]} \sigma_{\max}(M_i)}{\min\limits_{i\in[d]} \sigma_{\min}(M_i)}.$$

Although Theorem 7 is satisfyingly general, perhaps the most important and useful result proven in this section is the strictly narrower Proposition 2. This result generalizes Example 1 of Section 3.2, stated in (3.17), by abstracting out the term $\frac{\mathrm{e}^{-n/a}}{\sqrt[4]{2\delta-1}}$ and considering what "base vectors" $\gamma \in \mathbb{R}^d$ will produce well-conditioned spanning families of masks when m_j is taken to be the local Fourier measurement system with mask γ, and we determine a choice for the modulation index $K \in \mathbb{N}$ that guarantees the cleanest expression for the condition number.

Proposition 2. *Let $\{m_j\}_{j=1}^D \subseteq \mathbb{C}^d$ be a local Fourier measurement system with support*

δ, mask γ, and modulation index K, where $D = \min(d, 2\delta - 1)$. Let $\mathcal{A}$ be the associated measurement operator as in (4.1)*, with canonical matrix representation A as in* (4.2)*.*

If we additionally assume that $K = D$, then the condition number of $\mathcal{A}$ is

$$\kappa(\mathcal{A}) = \frac{d^{-1/2}\|\gamma\|_2^2}{\min\limits_{m\in[\delta]_0, j\in[d]} |F_d^*(\gamma \circ S^{-m}\gamma)_j|}. \tag{4.5}$$

Otherwise, if $K > D$, we may bound the condition number by

$$\kappa(\mathcal{A}) \leq \frac{d^{-1/2}\|\gamma\|_2^2}{\min\limits_{m\in[\delta]_0, j\in[d]} |F_d^*(\gamma \circ S^{-m}\gamma)_j|} \kappa(\widetilde{F}_K), \tag{4.6}$$

where $\widetilde{F}_K \in \mathbb{C}^{D\times D}$ is the $D \times D$ principal submatrix of F_K.

We recall, as discussed in Section 3.2, that the design of local Fourier measurement systems is motivated by the application of ptychography, described in Section 3.1.4 – in this type of laboratory setup, γ represents the "illumination function," describing the intensity of radiation applied to each segment of the sample – so it is appropriate to the end user that our simplest and most conveniently applied result pertains to a realistic, broad class of local measurement systems. In particular, we are grateful to have determined precisely which illumination functions are admissible for our phase retrieval algorithm: following almost immediately from Proposition 2, we have sufficient conditions for a local Fourier measurement system to be a spanning family, which we state in Corollary 4.

Corollary 4. *Let $\{m_j\}_{j\in D}$ be a local Fourier measurement system of support δ with mask $\gamma \in \mathbb{R}^d$ and modulation index K, where $D = \min(d, 2\delta - 1)$. Then $\{m_j\}_{j\in[D]}$ is a spanning family if $F_d^*(\gamma \circ S^{-m}\gamma)_j \neq 0$ for all $m \in [\delta]_0, j \in [d]$ and $K \geq D$.*

The proofs of Theorem 7 and Proposition 2 require some preliminary work in defining and studying a few new operators pertaining to the structure of (4.1). These definitions and a number of results concerning them are contained in Section 4.2.1. Section 4.2.3 contains some remarks about the "commonness" of masks γ that produce spanning Fourier families, as well as a technical result that supplements the simple, sufficient condition of Corollary 4

with a necessary and sufficient condition that covers the case $2\delta - 1 > d$.

4.2.1 Interleaving Operators and Circulant Structure

To set the stage for the proof, we introduce a certain collection of permutation operators and study their interactions with circulant and block-circulant matrices. The structure we identify here will be of much use to us in unraveling the linear systems we encounter in our model for phase retrieval with local correlation measurements.

To this end, we introduce the *interleaving operators* $P^{(d,N)} : \mathbb{C}^{dN} \to \mathbb{C}^{dN}$ for any $d, N \in \mathbb{N}$, each of which is a permutation defined by

$$(P^{(d,N)}v)_{(i-1)N+j} = v_{(j-1)d+i}. \tag{4.7}$$

We can view this is beginning with $v \in C^{dN}$ written as N blocks of d entries, and interleaving them into d blocks each of N entries. Additionally, for $\ell, N_1, N_2 \in \mathbb{N}, v \in \mathbb{C}^{\ell N_1}, k \in [\ell]$, and $H \in \mathbb{C}^{\ell N_1 \times N_2}$, we define the block circulant operator circ^{N_1} by

$$\begin{aligned}
\mathrm{circ}_k^{N_1}(v) &= \begin{bmatrix} v & S^{N_1}v & \cdots & S^{(k-1)N_1}v \end{bmatrix} \\
\mathrm{circ}_k^{N_1}(H) &= \begin{bmatrix} H & S^{N_1}H & \cdots & S^{(k-1)N_1}H \end{bmatrix},
\end{aligned}$$

where, as with $\mathrm{circ}(\cdot)$, when we omit the subscript we define $\mathrm{circ}^{N_1}(H) = \mathrm{circ}_\ell^{N_1}(H)$ and $\mathrm{circ}^{N_1}(v) = \mathrm{circ}_\ell^{N_1}(v)$. We now proceed with the following lemmas; the first establishes the inverse of $P^{(d,N)}$.

Lemma 10. *For $d, N \in \mathbb{N}$, we have*

$$(P^{(d,N)})^{-1} = P^{(d,N)*} = P^{(N,d)}.$$

Proof of Lemma 10. To prove the statement, we simply take $v \in \mathbb{C}^{dN}$ and calculate, for

$i \in [d], j \in [N]$,

$$\begin{aligned}(P^{(d,N)}P^{(N,d)}v)_{(i-1)N+j} &= (P^{(d,N)}(P^{(N,d)}v))_{(i-1)N+j} \\ &= (P^{(N,d)}v)_{(j-1)d+i} \\ &= v_{(i-1)N+j},\end{aligned}$$

with these equalities coming from the definition in (4.7). □

We now observe some useful ways in which the interleaving operators commute with the construction of circulant matrices.

Lemma 11. *Suppose* $V_i \in \mathbb{C}^{k\times n}, v_{ij} \in \mathbb{C}^k, w_j \in \mathbb{C}^{kN_1}$ *for* $i \in [N_1], j \in [N_2]$ *and*

$$M_1 = \begin{bmatrix} \operatorname{circ}(V_1) \\ \vdots \\ \operatorname{circ}(V_{N_1}) \end{bmatrix}, \quad M_2 = \begin{bmatrix} \operatorname{circ}^{N_1}(w_1) & \cdots & \operatorname{circ}^{N_1}(w_{N_2}) \end{bmatrix}, \text{ and}$$

$$M_3 = \begin{bmatrix} \operatorname{circ}(v_{11}) & \cdots & \operatorname{circ}(v_{1N_2}) \\ \vdots & \ddots & \vdots \\ \operatorname{circ}(v_{N_1 1}) & \cdots & \operatorname{circ}(v_{N_1 N_2}) \end{bmatrix}.$$

Then

$$P^{(k,N_1)}M_1 = \operatorname{circ}^{N_1}\left(P^{(k,N_1)}\begin{bmatrix} V_1 \\ \vdots \\ V_{N_1} \end{bmatrix}\right) \tag{4.8}$$

$$M_2 P^{(k,N_2)*} = \operatorname{circ}^{N_1}\left(\begin{bmatrix} w_1 & \cdots & w_{N_2} \end{bmatrix}\right) \tag{4.9}$$

$$P^{(k,N_1)}M_3P^{(k,N_2)*} = \operatorname{circ}^{N_1}\left(P^{(k,N_1)}\begin{bmatrix} v_{11} & \cdots & v_{1N_2} \\ \vdots & \ddots & \vdots \\ v_{N_1 1} & \cdots & v_{N_1 N_2} \end{bmatrix}\right). \tag{4.10}$$

Proof of lemma 11. We index the matrices to check the equalities. For (4.8), we take

$(a, b, \ell, j) \in [d] \times [N_1] \times [k] \times [n]$ and have

$$\begin{aligned}(P^{(k,N_1)} M_1)_{(a-1)N_1+b,(\ell-1)n+j} &= (M_1)_{(b-1)k+a,(\ell-1)n+j} \\ &= \begin{bmatrix} S^{\ell-1}V_1 \\ \vdots \\ S^{\ell-1}V_{N_1} \end{bmatrix}_{(b-1)k+a,j} \\ &= (S^{\ell-1}V_b)_{a,j} = (V_b)_{a+\ell-1,j}\end{aligned}$$

and

$$\begin{aligned}\operatorname{circ}^{N_1}\left(P^{(k,N_1)} \begin{bmatrix} V_1 \\ \vdots \\ V_{N_1} \end{bmatrix}\right)_{(a-1)N_1+b,(\ell-1)n+j} &= \left(P^{(k,N_1)} \begin{bmatrix} V_1 \\ \vdots \\ V_{N_1} \end{bmatrix}\right)_{(a-1)N_1+b+(\ell-1)N_1,j} \\ &= (V_b)_{a+\ell-1,j}\end{aligned}$$

For (4.9), we take $(a, b, j) \in [k] \times [N_2] \times [kN_1]$ and have

$$(P^{(k,N_2)} M_2^*)_{(a-1)N_2+b,j} = (M_2)_{j,(b-1)k+a} = (w_b)_{j+(a-1)N_1}$$

and

$$\left(\operatorname{circ}^{N_1}\left(\begin{bmatrix} w_1 & \cdots & w_{N_2} \end{bmatrix}\right)\right)_{j,(a-1)N_2+b} = (S^{N_1(a-1)} w_b)_j = (w_b)_{j+N_1(a-1)}$$

(4.10) follows immediately by combining (4.8) and (4.9). □

Lemma 12 introduces a few useful identities relating the interleaving operators to kronecker products.

Lemma 12. *For* $v \in \mathbb{C}^N, V = \begin{bmatrix} V_1 & \cdots & V_\ell \end{bmatrix} \in \mathbb{C}^{N \times \ell}, A = \begin{bmatrix} A_1 & \cdots & A_m \end{bmatrix} \in \mathbb{C}^{d \times m}$, *and*

$B_i \in \mathbb{C}^{m\times k}, i \in [\ell]$, *we have*

$$P^{(d,N)}(v \otimes A) = A \otimes v \tag{4.11}$$

$$P^{(d,N)}(V \otimes A) = \begin{bmatrix} A \otimes V_1 & \cdots & A \otimes V_\ell \end{bmatrix} \tag{4.12}$$

$$P^{(d,N)}(V \otimes A)P^{(\ell,m)} = A \otimes V \tag{4.13}$$

$$(V \otimes A)\begin{bmatrix} B_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_\ell \end{bmatrix} = \begin{bmatrix} V_1 \otimes AB_1 & \cdots & V_\ell \otimes AB_\ell \end{bmatrix} \tag{4.14}$$

Proof of Lemma 12. For (4.11), we see that, for $i, j, k \in [d] \times [N] \times [m]$, we have

$$\begin{aligned}(P^{(d,N)}v \otimes A)_{(i-1)N+j,k} &= (v \otimes A)_{(j-1)d+i,k} \\ &= v_j A_{ik}, \text{ while} \\ (A \otimes v)_{(i-1)N+j,k} &= A_{ik}v_j,\end{aligned}$$

and (4.12) follows by considering that $V \otimes A = \begin{bmatrix} V_1 \otimes A & \cdots & V_\ell \otimes A \end{bmatrix}$. To get (4.13), we trace the positions of columns, considering that $(V \otimes A)e_{(i-1)m+j} = V_j \otimes A_i$. From (4.12), we observe that $P^{(d,N)}(V \otimes A)e_{(i-1)m+j} = A_j \otimes V_i$, so

$$\begin{aligned}P^{(d,N)}(V \otimes A)P^{(m,\ell)}e_{(j-1)\ell+i} &= P^{(d,N)}(V \otimes A)e_{(i-1)m+j} \\ &= A_j \otimes V_i = (A \otimes V)e_{(j-1)\ell+i}.\end{aligned}$$

As for (4.14), we remark that

$$(V \otimes A)\begin{bmatrix} B_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_\ell \end{bmatrix} = (V \otimes A)\begin{bmatrix} e_1^\ell \otimes B_1 & \cdots & e_\ell^\ell \otimes B_\ell \end{bmatrix}$$

$$= \begin{bmatrix} (V \otimes A)(e_1^\ell \otimes B_1) & \cdots & (V \otimes A)(e_\ell^\ell \otimes B_\ell) \end{bmatrix} = \begin{bmatrix} V_1 \otimes AB_1 & \cdots & V_\ell \otimes AB_\ell \end{bmatrix},$$

as desired. □

The following lemma is a standard result (e.g., Theorem 13.26 in [65]) regarding the kronecker product.

Lemma 13. *We have* $\operatorname{vec}(ABC) = (C^T \otimes A)\operatorname{vec}(B)$ *for any* $A \in \mathbb{C}^{m\times n}, B \in \mathbb{C}^{n\times p}, C \in \mathbb{C}^{p\times k}$. *In particular, for* $a, b \in \mathbb{C}^d, \operatorname{vec} ab^* = \overline{b} \otimes a$, *and*

$$\operatorname{vec} E_{jk}(\operatorname{vec} E_{j'k'})^* = E_{kk'} \otimes E_{jj'}. \tag{4.15}$$

The next lemma covers the standard result concerning the diagonalization of circulant matrices, as well as a generalization to block-circulant matrices.

Lemma 14. *For any* $v \in \mathbb{C}^d$, *we have*

$$\operatorname{circ}(v) = F_d \operatorname{diag}(\sqrt{d}F_d^* v)F_d^* = \sqrt{d}\sum_{j=1}^{d}(f_j^{d*}v)f_j^d f_j^{d*} \tag{4.16}$$

Suppose $V \in C^{kN\times m}$, *then* $\operatorname{circ}^N(V)$ *is block diagonalizable by*

$$\operatorname{circ}^N(V) = (F_k \otimes I_N)\left(\operatorname{diag}(M_1, \ldots, M_k)\right)(F_k \otimes I_m)^*, \tag{4.17}$$

where

$$\sqrt{k}\,(F_k \otimes I_N)^* V = \begin{bmatrix} M_1 \\ \vdots \\ M_k \end{bmatrix}, \quad \text{or} \quad M_j = \sqrt{k}(f_j^k \otimes I_N)^* V \tag{4.18}$$

Proof of lemma 14. The diagonalization in (4.16) is a standard result: see, e.g., Theorem 7 of [50].

To prove (4.17), we set V_i to be the $k \times m$ blocks of V such that $V^* = \begin{bmatrix} V_1^* & \cdots & V_k^* \end{bmatrix}$ and begin by observing that, for $u \in \mathbb{C}^k$ and $W \in \mathbb{C}^{m\times p}$, the ℓ^{th} $k\times p$ block of $\operatorname{circ}^N(V)(u\otimes W)$ is given by

$$\left(\operatorname{circ}^N(V)(u \otimes W)\right)_{[\ell]} = \sum_{i=1}^{k} u_i (S^{N(i-1)}V)_\ell W = \sum_{i=1}^{k} u_i V_{\ell-i+1} W.$$

Taking $u = f_j^k$ and $W = I_m$, this gives

$$\begin{aligned}\left(\mathrm{circ}^N(V)(f_j^k \otimes I_m)\right)_{[\ell]} &= \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \omega_k^{(j-1)(i-1)} V_{\ell-i+1} I_m \\ &= \frac{1}{\sqrt{k}} \omega_k^{(j-1)(\ell-1)} \sum_{i=1}^{k} \omega_k^{-(j-1)(i-1)} V_i \\ &= (f_j^k)_\ell \left(\sqrt{k}(f_j^k \otimes I_N)^* V\right) = (f_j^k)_\ell M_j.\end{aligned}$$

This relation is equivalent to having

$$\mathrm{circ}^N(V)(f_j^k \otimes I_m) = (f_j^k \otimes M_j) = (f_j^k \otimes I_N) M_j,$$

which is the statement of the lemma. □

Lemma 14 immediately gives the following corollary regarding the conditioning of $\mathrm{circ}^N(V)$, with which we return to considering spanning families of masks.

Corollary 5. *With notation as in lemma 14, the condition number of* $\mathrm{circ}^N(V)$ *is*

$$\frac{\max\limits_{i\in[k]} \sigma_{\max}(M_i)}{\min\limits_{i\in[k]} \sigma_{\min}(M_i)}.$$

4.2.2 Proofs of Main Results

We begin with the proof of Theorem 7, of which Proposition 2 is a special case.

Proof of Theorem 7. We consider the rows of the matrix A representing the measurement operator $\mathcal{A}$, defined in (4.2) and (4.1). We vectorize $X \in T_\delta(\mathbb{C}^{d\times d})$ by its diagonals with $\mathcal{D}_\delta$,

as in (4.4) and set $\chi_m = \mathrm{diag}(X, m), m = 1-\delta, \ldots, \delta-1$. Each measurement then looks like

$$\begin{aligned}\mathcal{A}(X)_{(\ell,j)} &= \langle S^\ell m_j m_j^* S^{-\ell}, X\rangle \\ &= \sum_{m=1-\delta}^{\delta-1} \langle S^\ell g_m^j, \chi_m\rangle,\end{aligned}$$

so that the definition of A in (4.2) immediately yields its $(j-1)d+\ell^{\text{th}}$ row as $\begin{bmatrix} g_{1-\delta}^{j*} S^{1-\ell} & \cdots & g_{\delta-1}^{j*} S^{1-\ell}\end{bmatrix}$. Transposing this expression, we see that the $(j-1)d+1^{\text{st}}$ through $(j-1)d+d^{\text{th}}$ rows of A compose $\begin{bmatrix}\mathrm{circ}(g_{1-\delta}^j)^* & \cdots & \mathrm{circ}(g_{\delta-1}^j)^*\end{bmatrix}$. Together with $\mathrm{circ}(v)^* = \mathrm{circ}(R\overline{v})$, this determines A to be the block matrix given by

$$A = \begin{bmatrix}\mathrm{circ}(g_{1-\delta}^1)^* & \cdots & \mathrm{circ}(g_{\delta-1}^1)^* \\ \vdots & \ddots & \vdots \\ \mathrm{circ}(g_{1-\delta}^D)^* & \cdots & \mathrm{circ}(g_{\delta-1}^D)^*\end{bmatrix} = \begin{bmatrix}\mathrm{circ}(R\overline{g}_{1-\delta}^1) & \cdots & \mathrm{circ}(R\overline{g}_{\delta-1}^1) \\ \vdots & \ddots & \vdots \\ \mathrm{circ}(R\overline{g}_{1-\delta}^D) & \cdots & \mathrm{circ}(R\overline{g}_{\delta-1}^D)\end{bmatrix},$$

which may be transformed, by (4.10) of Lemma 11, to

$$P^{(d,D)} A P^{(d,2\delta-1)*} = \mathrm{circ}^D\left(P^{(d,D)}\begin{bmatrix}R\overline{g}_{1-\delta}^1 & \cdots & R\overline{g}_{\delta-1}^1 \\ \vdots & \ddots & \vdots \\ R\overline{g}_{1-\delta}^D & \cdots & R\overline{g}_{\delta-1}^D\end{bmatrix}\right) = \mathrm{circ}^D(H). \tag{4.19}$$

Quoting Corollary 5 establishes the theorem. □

We are now able to prove Proposition 2.

Proof of Proposition 2. We begin by remarking that, for any $j, m \in [d]$, the Cauchy-Schwarz inequality gives us

$$\begin{aligned} f_j^{d*}(\gamma \circ S^{-m}\gamma) &= \left(D_{f_j^{d*}\gamma}\right)\cdot\left(S^{-m}\gamma\right) \\ &\leq \frac{1}{\sqrt{d}}\|\gamma\|_2^2.\end{aligned}$$

Observing that $f_1^{d*}(\gamma \circ \gamma) = \frac{1}{\sqrt{d}}\|\gamma\|_2^2$, we have that

$$\max_{(j,m)\in[d]\times[\delta]_0} F_d^*(\gamma \circ S^{-m}\gamma)_j = \frac{1}{\sqrt{d}}\|\gamma\|_2^2. \tag{4.20}$$

For the moment, we assert that $D = 2\delta - 1 \leq d$ and set $\widetilde{F}_K \in \mathbb{C}^{2\delta-1\times 2\delta-1}, (\widetilde{F}_K)_{ij} = \frac{1}{\sqrt{K}}\omega_K^{(i-1)(j-\delta)}$ to be the principal submatrix of $\operatorname{diag}(\sqrt{K}f_{2-\delta}^K)F_K$. For a local Fourier measurement system, we have

$$\begin{aligned} g_m^j &= \operatorname{diag}(m_j m_j^*, m) = \operatorname{diag}((\gamma \circ v_j)(\gamma \circ v_j)^*, m) \\ &= \operatorname{diag}(D_{v_j}\gamma\gamma^* D_{\overline{v_j}}, m) = \omega_K^{-m(j-1)} \operatorname{diag}(\gamma\gamma^*, m), \end{aligned} \tag{4.21}$$

so, setting $g_m = \operatorname{diag}(\gamma\gamma^*, m)$, we have $g_m^j = \operatorname{diag}(m_j m_j^*, m) = \omega_K^{-m(j-1)} g_m$. Therefore, we label the $2\delta - 1 \times 2\delta - 1$ blocks of H by $H^* = \begin{bmatrix} H_1^* & \cdots & H_d^* \end{bmatrix}$, so that

$$(H_\ell)_{ij} = (R\overline{g}_{j-\delta}^i)_\ell = \omega_K^{(i-1)(j-\delta)}(Rg_{j-\delta})_\ell$$

and $M_\ell = \sqrt{d}(f_j^d \otimes I_D)^* H = \sum_{k=1}^d \omega_d^{-(\ell-1)(k-1)} H_k$, giving

$$\begin{aligned} (M_\ell)_{ij} &= \sum_{k=1}^d \omega_d^{-(\ell-1)(k-1)}(H_k)_{ij} = \omega_K^{(i-1)(j-\delta)} \sum_{k=1}^d \omega_d^{-(\ell-1)(k-1)}(Rg_{j-\delta})_k \\ &= \sqrt{d}\omega_K^{(i-1)(j-\delta)}(\overline{F}_d^* g_{j-\delta})_\ell. \end{aligned}$$

In other words,

$$M_\ell = \sqrt{dK}\,\widetilde{F}_K \operatorname{diag}(f_{2-\ell}^{d*}g_{1-\delta}, \ldots, f_{2-\ell}^{d*}g_{\delta-1}). \tag{4.22}$$

If $K = 2\delta - 1$, then $\widetilde{F}_K$ is unitary, and the singular values of M_ℓ are $\{\sqrt{dK}|f_\ell^{d*}g_j|\}_{j=1-\delta}^{\delta-1}$ (since $|f_{2-\ell}^{d*}g_j| = |\overline{f}_{2-\ell}^{d*}g_j| = |f_\ell^{d*}g_j|$). Recognizing that $S^j g_j = g_{-j}$, then Theorem 7 and (4.20) take us to (4.5).

If $D = 2\delta - 1 < K$, then the argument remains unchanged, except that the singular values of M_ℓ, instead of being known explicitly, are bounded above and below by $\max_{|j|<\delta}|f_\ell^{d*}g_j|\sigma_{\max}(\widetilde{F}_K)$ and $\min_{|j|<\delta}|f_\ell^{d*}g_j|\sigma_{\min}(\widetilde{F}_K)$ respectively, which gives the more general re-

sult of (4.6).

If $2\delta - 1 > d = D$, then, by an argument similar to that used in Theorem 7, we may obtain

$$A = \begin{bmatrix} \mathrm{circ}(R\overline{g}_0^1) & \cdots & \mathrm{circ}(R\overline{g}_{d-1}^1) \\ \vdots & \ddots & \vdots \\ \mathrm{circ}(R\overline{g}_0^d) & \cdots & \mathrm{circ}(R\overline{g}_{d-1}^d) \end{bmatrix},$$

and a similar application of (4.10) gives that

$$P^{(d,d)} A P^{(d,d)*} = \mathrm{circ}^d \left(P^{(d,d)} \begin{bmatrix} R\overline{g}_0^1 & \cdots & R\overline{g}_{d-1}^1 \\ \vdots & \ddots & \vdots \\ R\overline{g}_0^d & \cdots & R\overline{g}_{d-1}^d \end{bmatrix} \right).$$

Setting

$$H = P^{(d,d)} \begin{bmatrix} R\overline{g}_0^1 & \cdots & R\overline{g}_{d-1}^1 \\ \vdots & \ddots & \vdots \\ R\overline{g}_0^d & \cdots & R\overline{g}_{d-1}^d \end{bmatrix},$$

and defining $H_\ell \in \mathbb{C}^{d\times d}$ by $H^* = \begin{bmatrix} H_1^* & \cdots & H_d^* \end{bmatrix}$, we have

$$(H_\ell)_{ij} = \omega_K^{(i-1)(j-1)} (R g_{j-1})_\ell \quad \text{and} \quad (M_\ell)_{ij} = \sqrt{d}\,\omega_K^{(i-1)(j-1)} (\overline{F}_d^* g_{j-1})_\ell,$$

giving $M_\ell = \sqrt{dK}\,\mathcal{R}_{d\times d}(F_K)\,\mathrm{diag}(f_{2-\ell}^{d*} g_0, \ldots, f_{2-\ell}^{d*} g_{d-1})$, which immediately gives us (4.6). We remark that indexing only over the diagonals $m \in [\delta]_0$ in (4.6) suffices, again because $S^j g_j = g_{-j}$, so having $2\delta - 1 > d$ makes $1-\delta, \ldots, -1$ redundant. □

4.2.3 Remarks on Spanning Family Results

Having proven Proposition 2 and Corollary 4, we remark that the condition for a local Fourier measurement system to be a spanning family is generic, in the sense that it fails to hold only on a subset of $\mathbb{R}^d$ with Lebesgue measure zero, except possibly when $\delta > d/2$. We consider that, whenever $m \neq d/2$, the set of $\gamma \in \mathbb{R}^d$ giving at least one zero in $F_d^*(\gamma \circ S^{-m}\gamma)$

is a finite union of zero sets of non-trivial quadratic polynomials and hence a set of zero measure. Indeed, when $m \neq d/2$, we have that

$$F_d^*((e_1 + e_{m+1}) \circ S^m(e_1 + e_{m+1}))_k = f_k^{d*} e_{m+1} = \omega^{-m(k-1)},$$

so $\gamma \mapsto F_d^*(\gamma \circ S^m \gamma)_k$ is a non-zero, homogeneous quadratic polynomial and therefore has a zero locus of measure zero.

However, when $d = 2m$, then $\gamma \circ S^m \gamma$ is periodic with period m and $F_d^*(\gamma \circ S^m \gamma)_{2i} = 0$ for $i \in [m]_0$. In particular, if $\delta \geq d/2$, then $D = d$ and, with reference to the notation of Theorem 7, M_{2i} will be singular for all i, making A itself singular!

Nonetheless, there is a way to have $\mathcal{L}_\gamma = \mathcal{H}^d$ when $d/2$; the reason this does not contradict Theorem 7 and Proposition 2 is that, as was observed in the proof of Corollary 4, these consider A and $\mathcal{A}$ as operators on the complex vector space $\mathbb{C}^{d \times d}$, and restricting to $\mathcal{H}^d$ as a vector space over $\mathbb{R}$ changes the behavior of the system in a way that must be accounted for.

In this case, the behavior is changed in our favor, and while Proposition 2 prohibits us from constructing a local Fourier measurement system that spans $\mathbb{C}^{d \times d}$ when $2 \mid d$, Proposition 3 shows that we may have $\text{span}_{\mathbb{R}}\{S^\ell m_j m_j^* S^{-\ell}\} = \mathcal{H}^d$. Necessary and sufficient conditions to procure a spanning family of $T_\delta(\mathcal{H}^d)$ in general, and $\mathcal{H}^d$ in particular when $\delta \geq \frac{d+1}{2}$ are stated in Proposition 3. However, since the proof of this result is tedious, and since it represents an edge case that is largely uninteresting to the intended application of this dissertation, we relegate its proof to Appendix A.

Proposition 3. *Suppose $\{m_j\}_{j \in [D]} \subseteq \mathbb{C}^d$ is a local Fourier measurement system of support δ with mask $\gamma \in \mathbb{R}^d$ and modulation index $K \geq D = \min(2\delta - 1, d)$. Then $\{m_j\}_{j \in [D]}$ is a spanning family, that is,*

$$\text{span}_{\mathbb{R}}\{S^\ell m_j m_j^* S^{-\ell}\}_{(\ell,j) \in [d]_0 \times [D]} = T_\delta(\mathcal{H}^d)$$

if and only if each of the sets $J_k := \{m \in [\delta]_0 : (F_d^*(\gamma \circ S^{-m}\gamma))_k \neq 0\}$*, for all* $k \in [d]$*, satisfy*

$$2|J_k| - \mathbb{1}_{0 \in J_k} \geq D. \tag{4.23}$$

Equivalently, we may require

$$\#\text{nnz}\left(f_k^{d*} D_\gamma \begin{bmatrix} S^{1-\delta}\gamma & \cdots & S^{\delta-1}\gamma \end{bmatrix}\right) \geq D, \text{ for all } k \in [d]. \tag{4.24}$$

We remark that the maximum value of $2|J_k| - \mathbb{1}_{0 \in J_k}$ is $2\delta - 1$, so when $D = 2\delta - 1 \leq d$, the condition of (4.23) is equivalent to Corollary 4: $F_d^*(\gamma \circ S^{-m}\gamma)$ must contain no zeros for all $m \in [\delta]_0$. However, when $D = d \leq 2\delta - 1$, we notice that (4.23) can be satisfied, even when some terms $F_d^*(\gamma \circ S^{-m}))_k$ are zero, as long as no particular Fourier component f_k^d produces too many zeros as m sweeps through $[\delta]_0$; this reforumlation of the condition is made somewhat more explicit in (4.24).

4.3 Explicit Examples of Spanning Families

In this section, we analyze three explicit examples of masks $\gamma \in \mathbb{R}^d$ and their corresponding local Fourier measurement systems, and prove under what conditions these constitute spanning families. The goal is to constructively provide examples of spanning families that are well-conditioned, and which are scalable in the sense that they may be used for any choice of d and δ. Specifically, in Section 4.3.1, we analyze Example 1 of Section 3.2 (also known as "exponential masks," as we take $\gamma_i = Ca^i$ for some $C, a \in \mathbb{R}$) with the new theory of this chapter, and find improvements on the bounds of its condition number, which scales roughly like $\kappa \approx \delta^2$. In Section 4.3.2, a new set of masks is studied. These masks, referred to as "near-flat masks," are constructed by taking $\gamma = ae_1 + \mathbb{1}_{[\delta]} \in \mathbb{R}^d$, and we provide a choice of a that achieves a condition number that is asymptotically linear in δ – a notable improvement over the conditioning of the exponential masks. Finally, in Section 4.3.3, we note the somewhat curious case of a constant mask, $\gamma = \mathbb{1}_{[\delta]}$. Here, γ produces a spanning

local Fourier measurement system – with poor conditioning – when d's prime divisors are each greater than δ.

4.3.1 Exponential Masks

For our analysis of Example 1 of Section 3.2, we rephrase the definition of (3.17) somewhat slightly. Here, we will take $d \in \mathbb{N}$ to be the ambient dimension and $\delta \leq \frac{d+1}{2}$. Then, we let $\{m_j(a)\}_{j=1}^{2\delta-1}$ be the local Fourier measurement system with mask $\gamma(a) \in \mathbb{R}^d$ defined by $\gamma(a)_i = a^{i-1}$, for some $0 < a \in \mathbb{R}, a \neq 1$. Denoting the condition number of the measurement operator associated to $\{m_j(a)\}_{j\in[2\delta-1]}$ by $kappa(a)$, we remark that it suffices to compute (or estimate) $\kappa(a)$ for all $a > 1$, as $\kappa(a) = \kappa(1/a)$, which is proven in Proposition 4.

Proposition 4. *Given $d \in \mathbb{N}, \delta \leq \frac{d+1}{2}$, and $a \in (0,1) \cup (1,\infty)$, we define $\{m_j(a)\}_{j\in[2\delta-1]}$ to be the local Fourier measurement system of support δ with mask $\gamma(a) \in \mathbb{R}^d$ defined by*

$$\gamma(a)_i = \begin{cases} a^{i-1}, & i \in [\delta] \\ 0, & otherwise \end{cases}.$$

Let $\kappa(a) := \kappa(\mathcal{A})$ be the condition number of the measurement operator associated with $\{m_j(a)\}_{j\in[2\delta-1]}$. Then $\kappa(a) = \kappa(1/a)$.

Proof of Proposition 4. Fix $a > 1$. We begin by noting that $\gamma(a) = a^{\delta-1}S^{\delta-1}R\gamma(a^{-1})$, so that $\|\gamma(a)\|_2^2 = a^{2\delta-2}\|\gamma(a^{-1})\|_2^2$ and

$$\begin{aligned} \gamma(a) \circ S^{-m}\gamma(a) &= a^{2\delta-2}\left(S^{\delta-1}R\gamma(a^{-1}) \circ S^{-m}S^{\delta-1}R\gamma(a^{-1})\right) \\ &= a^{2\delta-2}S^{\delta-1}RS^m(\gamma(a^{-1}) \circ S^{-m}\gamma(a^{-1})), \end{aligned}$$

so that

$$f_j^{d*}(\gamma(a)\circ S^{-m}\gamma(a)) = a^{2\delta-2}\omega_d^{-(j-1)(\delta-m-1)} f_{2-j}^{d*}(\gamma(a^{-1})\circ S^{-m}\gamma(a^{-1}))$$
$$= a^{2\delta-2}\omega_d^{-(j-1)(\delta-m-1)}\overline{f_j^{d*}(\gamma(a^{-1})\circ S^{-m}\gamma(a^{-1}))},$$

which gives, by Proposition 2, that

$$\kappa(a) = \frac{d^{-1/2}\|\gamma(a)\|_2^2}{\min\limits_{m\in[\delta]_0, j\in[d]}\left|f_j^{d*}(\gamma(a)\circ S^{-m}\gamma(a))\right|}$$
$$= \frac{d^{-1/2}\|\gamma(a^{-1})\|_2^2}{\min\limits_{m\in[\delta]_0, j\in[d]}\left|f_j^{d*}(\gamma(a^{-1})\circ S^{-m}\gamma(a^{-1}))\right|} = \kappa(1/a)$$

□

We now produce an estimate of $\kappa(a)$ in Proposition 5.

Proposition 5. *With $d, \delta, \{m_j(a)\}_{j\in[2\delta-1]}, \gamma(a), \kappa(a)$ defined as in Proposition 4, when $a > 1$ we have*

$$\kappa(a) \leq \frac{a^{2\delta}-1}{a^{\delta+1}-a^{\delta-1}}\cdot\frac{a^2+1}{a^2-1}, \tag{4.25}$$

and when $a \in (0,1), \kappa(a) = \kappa(1/a)$. Taking

$$a_\delta = \begin{cases} 1+\frac{4}{\delta-2}, & \delta \geq 5 \\ \frac{1+\sqrt{5}}{2}, & otherwise \end{cases},$$

we may obtain

$$\kappa(a_\delta) \leq \left(\frac{\mathrm{e}(\delta+2)}{4}\right)^2. \tag{4.26}$$

Proof of Proposition 5. We fix $a > 1$, and we remark that

$$\|\gamma(a)\|_2^2 = \sum_{i=1}^{\delta} a^{2(i-1)} = \frac{1-a^{2\delta}}{1-a^2} = \frac{a^{2\delta}-1}{a^2-1}, \tag{4.27}$$

so all that remains is to bound $\min_{j,m}|f_j^{d*}(\gamma(a)\circ S^{-m}\gamma(a))|$ from below. For convenience,

we define $h(j,m) := \sqrt{d} f_j^{d*}(\gamma(a) \circ S^{-m}\gamma(a))$. We begin by computing

$$\begin{aligned} |h(j,m)| &= \left| \sum_{i=1}^{\delta-m} a^{m+2(i-1)} \omega_d^{-(j-1)(i-1)} \right| \\ &= \left| \frac{a^m - a^{2\delta-m}\omega_d^{-(j-1)(\delta-m)}}{1 - a^2\omega_d^{-(j-1)}} \right| \\ &\geq \left| \frac{a^{2\delta-m} - a^m}{a^2+1} \right|. \end{aligned} \tag{4.28}$$

which is clearly decreasing with m, so it is minimized when $m = \delta - 1$. This gives that $\min_{j,m} |f_j^{d*}(\gamma(a) \circ S^{-m}\gamma(a))| \geq \frac{1}{\sqrt{d}} \frac{a^{\delta+1}-a^{\delta-1}}{a^2+1}$, which, with (4.27) and Propositions 2 and 4, yields (4.25).

To obtain (4.26), we consider that (4.28) is necessarily suboptimal. Consider that, trivially, $h(j,\delta-1) = |\gamma(a)_1 \gamma(a)_\delta| = a^{\delta-1}$, whereas we have bounded it below by $\frac{a^{\delta+1}-a^{\delta-1}}{a^2+1} = a^{\delta-1}\frac{a^2-1}{a^2+1}$. Hence, a tighter lower bound may be to use $m = \delta - 2$ in the right hand side of (4.26). Using

$$\frac{a^{\delta+2} - a^{\delta-2}}{a^2+1} = a^{\delta-2}\frac{a^4-1}{a^2+1} = a^{\delta-2}(a^2-1),$$

this yields

$$h(j,m) \geq \min\left(a^{\delta-1}, a^{\delta-2}(a^2-1)\right).$$

We further observe that $a^{\delta-2}(a^2-1) = a^{\delta-1}$ exactly when $a^2 - a - 1 = 0$, yielding

$$h(j,m) \geq \begin{cases} a^{\delta-1}, & a \geq \frac{1+\sqrt{5}}{2} \\ a^{\delta-2}(a^2-1), & 1 < a \leq \frac{1+\sqrt{5}}{2} \end{cases}.$$

Therefore, for $\delta \geq 5$, we have $(1 + \frac{4}{\delta-2})^{1/2} \leq \frac{1+\sqrt{5}}{2}$ so that, choosing $a_\delta = 1 + \frac{4}{\delta-2}$, we may guarantee

$$\begin{aligned} \kappa(a_\delta) &\leq \frac{a_\delta^{2\delta} - 1}{a_\delta^{\delta-2}(a_\delta^2-1)^2} \leq \frac{a_\delta^{\delta+2}}{(a_\delta^2-1)^2} \\ &= \frac{(1+\frac{4}{\delta-2})^{(\delta+2)/2}}{(4/(\delta-2))^2} \leq \frac{\mathrm{e}^2(1+\frac{4}{\delta-2})^2}{(4/(\delta-2))^2} = \left(\frac{\mathrm{e}(\delta+2)}{4}\right)^2, \end{aligned} \tag{4.29}$$

where the third inequality comes from $(1 + \frac{2}{x})^x \leq \mathrm{e}^2$ for $x > 0$. For $\delta \in [2,5)$, setting

$\phi = \frac{1+\sqrt{5}}{2}$, we may upper bound

$$\kappa(\phi) \leq \frac{\phi^{2\delta} - 1}{\phi^{\delta-1}},$$

and numerically verify that this implies $\kappa(\phi) < (\mathrm{e}(\delta+2)/4)^2$ for $\delta \in [2, 5)$.

□

We remark that the bound in (4.26) is slightly stronger than a similar bound proven in Theorem 4 of [57], the work by Iwen, Viswanathan, and Wang where this family of masks was first studied. They bounded the condition number for the optimal choice of a by

$$\kappa < \max\left\{144\mathrm{e}^2, \left(\frac{3\mathrm{e}(\delta-1)}{2}\right)^2\right\}.$$

For small delta, the term $144\mathrm{e}^2 > 1,000$ is much too loose, and even if this term was omitted, $\frac{\delta+2}{4} < \frac{3(\delta-1)}{2}$ for all $\delta > 1$. For asymptotically large δ, the bound proven above in Proposition 5 is stronger by a constant factor of 36.

4.3.2 Near-Flat Masks

We now analyze masks of the form $\gamma = ae_1 + \mathbb{1}_{[\delta]}$ in Proposition 6.

Proposition 6. *Let $\{m_j\}_{j\in[D]} \subseteq \mathbb{C}^d$ be the local Fourier measurement system of support $\delta \leq \frac{d+1}{2}$ with mask $\gamma \in \mathbb{R}^d$ given by $\gamma = ae_1 + \mathbb{1}_{[\delta]}$ where $a > \delta - 1$. Then this is a spanning family with condition number bounded above by*

$$\kappa \leq \frac{a^2 + 2a + \delta}{a - \delta + 1}. \tag{4.30}$$

If we choose $a = 2\delta - 1$, we have $\kappa \leq 4\delta + 1$.

Proof of Proposition 6. We calculate the condition number directly. We immediately have $\|\gamma\|_2^2 = (a+1)^2 + (\delta - 1) = a^2 + 2a + \delta$, which is the numerator of (4.30), so it remains only

to provide a lower bound on $\sqrt{d}f_j^{d*}(\gamma \circ S^{-m}\gamma)$. To achieve this, we remark that, for $m \geq 1$,

$$\sqrt{d}f_j^{d*}(\gamma \circ S^{-m}\gamma) = a + \sum_{i=1}^{\delta-m} \omega_d^{(j-1)(i-1)} = \begin{cases} a+\delta-m, & j=1 \\ a + \dfrac{1-\omega_d^{(j-1)(\delta-m)}}{1-\omega_d^{j-1}}, & \text{otherwise} \end{cases}.$$

Clearly, this expression has its maximum absolute value when $j = 1$, as $|a+\sum_{i=1}^{\delta-m} \omega_d^{(j-1)(i-1)}| \leq a + \sum_{i=1}^{\delta-m} |\omega_d^{(j-1)(i-1)}| = a + \delta - m$, so we consider that, for $j \neq 1$, we have

$$\left|\sqrt{d}f_j^{d*}(\gamma \circ S^{-m}\gamma)\right| \geq \left|\mathrm{Re}\left(a + \frac{1-\omega_d^{(j-1)(\delta-m)}}{1-\omega_d^{j-1}}\right)\right|. \tag{4.31}$$

We then reduce the term $\mathrm{Re}\left(\frac{1-\omega_d^{(j-1)(\delta-m)}}{1-\omega_d^{j-1}}\right)$ by setting $\mathrm{e}^{\mathrm{i}\theta} := \omega_d^{j-1}$ and $k := \delta - m$ and finding

$$\begin{aligned} \mathrm{Re}\left(\frac{1-\mathrm{e}^{\mathrm{i}k\theta}}{1-\mathrm{e}^{\mathrm{i}\theta}}\right) &= \mathrm{Re}\left(\frac{(1-\mathrm{e}^{\mathrm{i}k\theta})(1-\mathrm{e}^{-\theta\mathrm{i}})}{2-2\cos\theta}\right) \\ &= \frac{(1-\cos\theta)+\cos(k-1)\theta-\cos k\theta}{2-2\cos\theta} \\ &= \frac{1}{2} + \frac{\sin(k-\frac{1}{2})\theta \sin\frac{\theta}{2}}{2\sin^2\frac{\theta}{2}}, \end{aligned} \tag{4.32}$$

where the third line comes from $\sin^2 x = (1-\cos(2x))/2$ and $\cos a - \cos b = \frac{1}{2}\sin\left(\frac{a+b}{2}\right)\sin\left(\frac{b-a}{2}\right)$. Using $|\sin n\theta| \leq n|\sin\theta|$, this gives us that $\mathrm{Re}\left(\frac{1-\mathrm{e}^{\mathrm{i}k\theta}}{1-\mathrm{e}^{\mathrm{i}\theta}}\right) \geq \frac{1}{2} - \frac{2k-1}{2} = -k$, and hence

$$\left|\sqrt{d}f_j^{d*}(\gamma \circ S^{-m}\gamma)\right| \geq a - \delta + m \geq a - \delta + 1 \tag{4.33}$$

for all $1 \leq m < \delta$. For $m = 0$, a similar calculation gives that

$$\left|\sqrt{d}f_j^{d*}(\gamma \circ \gamma)\right| \geq a^2 + 2a - \delta, \tag{4.34}$$

and we notice that this bound is *greater* than the bound for the case $1 \leq m$, stated in (4.33) whenever $a \geq \frac{1+\sqrt{5}}{2}$. This means (4.33) is tighter than (4.34) whenever $a > \delta - 1$, which is necessary for these bounds to be positive and meaningful. Therefore, we may restrict to choices of $a > \delta - 1$ and use the bound $\left|\sqrt{d}f_j^{d*}(\gamma \circ S^{-m}\gamma)\right| \geq a - \delta + 1$. This completes the

proof. □

We remark that this result is a very welcome contribution to our library of well-conditioned local measurement systems. In Example 2 of Section 3.2, we provided an example of a local measurement system that achieves a condition number that scales as $\mathcal{O}(\delta)$, but it was a somewhat cumbersome construction. Each of its members m_j was incredibly sparse, having either 1 or 2 nonzero entries, which is somewhat unrealizable since it is not even a local Fourier measurement system (which corresponds to the diffraction process that we expect to govern our measurement apparatus). Furthermore, its sparsity at the level of discretization means a practitioner would have to have perfect control over the illumination of their sample at the scale of resolution they want for their image, which puts an unrealistic demand on the accuracy of the equipment being used (imagine taking a photograph by separately illuminating each pixel in a scene and recording the reflected light). By contrast, the example proposed in Proposition 6, while not necessarily simple to achieve in the lab, at least has the merit of accomodating the Fresnel diffraction model which motivates the study of local Fourier measurement systems, and its conditioning asymptotically equals that of the sparse construction that was previously our most well-conditioned measurement system.

We further remark that a small improvement can be made to this bound. By following the work of Mercer in [71], we can asymptotically (as $d, \delta \to \infty$) bound the real part of $\sum_{i=1}^{m} \omega_d^{(j-1)(i-1)}$ below by $C\delta$ where $C > -1/4$. At the very least, a choice of $a = 2\delta$ would in this case yield a condition number that asymptotically converges to no more than $4\delta^2/(7/4\delta) = \frac{16}{7}\delta$, but this yields an improvement of no more than a factor of 2. In fact, optimizing the choice of a more carefully, with this improved constant, can get us down to $\kappa < \frac{7}{8}\delta$ as $\delta \to \infty$, but this proof is technical, so we relegate it to Appendix B and illustrate it numerically in Section 4.5.1.

4.3.3 Constant Masks

After these two examples, we also remark that the simplest type of mask – a constant mask, where $\gamma = \mathbb{1}_{[\delta]}$ – can actually produce a spanning family. The conditions required of δ and d to admit this are upsettingly number theoretical, so we present this result in Proposition 7 primarily as an incidental curiosity.

Proposition 7. *Take $d \in \mathbb{N}$ and $\delta \leq d$. Then, with $D = \min(d, 2\delta - 1)$, the local Fourier measurement system $\{m_j\}_{j=1}^{D}$ of support δ and mask $\gamma = \mathbb{1}_{[\delta]}$ is a spanning family if and only if d is strictly δ-rough, in the sense that $k \mid d \implies k > \delta$. In this event, and if we additionally take $d > 4$, the condition number of $\mathcal{A}$ is bounded by $\kappa \leq \delta d^2/8$.*

Proof of Proposition 7. We begin by remarking that $\gamma \circ S^{-(\delta-k)}\gamma = \mathbb{1}_{[k]}$, such that

$$\sqrt{d} f_j^{d*}(\gamma \circ S^{-(\delta-k)}\gamma) = \sum_{i=1}^{k} \omega_d^{(j-1)(i-1)} = \begin{cases} k, & j = 1 \\ \dfrac{1 - \omega_d^{(j-1)k}}{1 - \omega_d^{(j-1)}}, & \text{otherwise} \end{cases}, \tag{4.35}$$

and hence $f_j^{d*}(\gamma \circ S^{-(\delta-k)}\gamma) = 0$ iff $(j-1)k = nd$ for some positive integer n. By Corollary 4, this means that γ produces a spanning family iff there does not exist a pair $(j, k) \in [d] \times [\delta]_0$ such that $jk = nd$ for some positive integer n. This condition occurs iff there is no pair $(j, k) \in [d] \times [\delta]$ satisfying $jk = d$, which is to say that γ produces a spanning family iff d is strictly δ-rough.

To get the condition number, consider that $\|\gamma\|_2^2 = \delta$. It suffices, then, to get a lower bound on $\sqrt{d}\left|f_j^{d*}\mathbb{1}_{[k]}\right|$ for all $k \in [\delta]$. Trivially following from (4.35), we may write

$$\sqrt{d}\left|f_j^{d*}\mathbb{1}_{[k]}\right| \geq \frac{|1 - \omega_d|}{2} \geq \frac{|\mathrm{Re}(1 - \omega_d)|}{2} = \frac{1 - \cos\left(\frac{2\pi}{d}\right)}{2}.$$

For $d > 4$, we use $1 - \cos(x) \geq (2x/\pi)^2$ to get that $(1 - \cos(\frac{2\pi}{d}))/2 \geq 8/d^2$, which completes the proof. (We also note that the only possible case when $d \leq 4$ is $d = 3, \delta = 2$, and we can find by exhaustive calculation that $\kappa = \frac{2}{2-\sqrt{3}}$). □

This result shows that, while constant masks can produce spanning families in some circumstances, the condition number of the resulting linear system is remarkably unstable as a function of the parameters of the discretization, d and δ. At the very least, we have that if (δ, d) admits a constant spanning local Fourier measurement system, then $(\delta, d+1)$ will not. Since d is intended to represent the number of pixels in the sensor array, this is a prohibitively specific requirement to be made of the discretization of the phase retrieval problem, so we emphasize that the result of Proposition 7 is of strictly mathematical interest.

4.4 Inverting $\mathcal{A}$

In this section, we use the results of Section 4.2 to explicitly state the inverse of the measurement operator $\mathcal{A}$, as well as the computational complexity of calculating its inverse. Additionally, we calculate the variance in *each entry* of $\mathcal{A}^{-1}(y)$ when $y = \mathcal{A}(T_\delta(xx^*)) + \eta$ is produced under an i.i.d. Gaussian noise model, which will be useful to us in later analysis.

4.4.1 Explicit inverse of $\mathcal{A}$

We begin by fixing a local measurement system $\{m_j\}_{j=1}^D$ with support δ, taking $\mathcal{A}$ to be the associated measurement operator and A its canonical matrix representation as in Eqs. (4.1) and (4.2). We then remark from (4.19) and Lemma 14 that

$$A = P^{(D,d)}(F_d \otimes I_D)\,\mathrm{diag}(M_\ell)_{\ell=1}^d (F_d \otimes I_D)^* P^{(d,D)},$$

where we recall M_ℓ from (4.18). In the case where $\{m_j\}_{j=1}^D$ is a local Fourier measurement system with mask γ and modulation index K, we define $Z \in \mathbb{C}^{D\times d}$ by

$$Z_{m\ell} = \sqrt{dK}\, f_{2-\ell}^{d*} g_{m-\delta}. \tag{4.36}$$

Setting $z_\ell = Ze_\ell$, we have

$$M_\ell = \widetilde{F}_K D_{z_\ell} \text{ and } \operatorname{diag}(M_\ell)_{\ell=1}^d = (I_d \otimes \widetilde{F}_K)\operatorname{diag}(\operatorname{vec}(Z)), \tag{4.37}$$

by which we further reduce A to

$$A = P^{(D,d)}(F_d \otimes I_D)(I_d \otimes \widetilde{F}_K)\operatorname{diag}(\operatorname{vec}(Z))(F_d \otimes I_D)^* P^{(d,D)}.$$

This reasoning immediately produces the inverse of A, which we state in Proposition 8. While Proposition 8 covers the case of general local measurement systems, the rest of this section will be restricted to local Fourier measurement systems.

Proposition 8. *Let $A \in \mathbb{C}^{dD\times dD}$ be the canonical representation of the measurement operator $\mathcal{A}$ associated with a local Fourier measurement system $\{m_j\}_{j=1}^d$ of support $\delta \leq \frac{d+1}{2}$ with mask $\gamma \in \mathbb{R}^d$. Defining Z as in* (4.36)*, we have*

$$A^{-1} = P^{(D,d)}(F_d \otimes I_D)(\operatorname{diag}(\operatorname{vec}(Z)))^{-1}(I_d \otimes \widetilde{F}_D^*)(F_d \otimes I_D)^* P^{(d,D)}. \tag{4.38}$$

If $\{m_j\}_{j=1}^D$ is a general local measurement system, and $\mathcal{A}$ is invertible, then its inverse is given by

$$A^{-1} = P^{(D,d)}(F_d \otimes I_D)\operatorname{diag}(M_\ell^{-1})_{\ell=1}^d (F_d \otimes I_D)^* P^{(d,D)}.$$

This formulation makes it straightforward to deduce the computational complexity of inverting $\mathcal{A}$, as previously stated in Section 3.1.3. Namely, each permutation $P^{(D,d)}$ may be run on a vector with $\mathcal{O}(dD)$ operations. Since, by (4.13) of Lemma 12, we have that $F_d \otimes I_D = P^{(d,D)}(I_D \otimes F_d)P^{(D,d)}$, and considering also that $I_D \otimes F_d$ comprises D Fourier transforms of dimension d, multiplication by $F_d \otimes I_D$ costs $\mathcal{O}(dD + Dd\log d) = \mathcal{O}(dD\log d)$ operations. Since $\widetilde{F}_D = F_D S^{1-\delta}$, multiplying by $I_d \otimes \widetilde{F}_D^*$ takes $\mathcal{O}(dD\log D)$ operations. Finally, multiplying by $\operatorname{diag}(\operatorname{vec}(Z))^{-1}$ trivially has a cost of $\mathcal{O}(dD)$. Putting all these considerations together, and recalling that $d \geq D = 2\delta - 1$, the cost of inverting A comes

out to

$$\mathcal{O}(dD) + \mathcal{O}(dD \log d) + \mathcal{O}(dD \log D) + \mathcal{O}(dD) + \mathcal{O}(dD \log d) + \mathcal{O}(dD) = \mathcal{O}(dD \log d),$$

or $\mathcal{O}(\delta\, d \log d)$, as concluded in Section 3.1.3 and [57].

4.4.2 Preliminaries in Probability

Prior to discussing the variance of the images of vectors under A^{-1} or $\mathcal{A}^{-1}$, we review some preliminaries regarding the real and complex multivariate Gaussian distributions. These results and the notation with which we express them may be found in many standard texts, for example [93] for the real case and section 7.8.1 of [43] for the complex. For $\mu \in \mathbb{R}^n$ and $0 \prec \Sigma \in \mathcal{S}^n, \mathcal{N}(\mu, \Sigma)$ refers to the multivariate normal distribution on $\mathbb{R}^n$ with mean μ and covariance matrix $\Sigma = \mathbb{E}_x[(x-\mu)(x-\mu)^T]$, and is determined by its probability density function

$$f_{\mathcal{N}}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right).$$

For $\nu \in \mathbb{C}^n$ and $0 \prec \Xi \in \mathcal{H}^n, \mathcal{CN}(\nu, \Xi)$ is the (circularly symmetric about ν) multivariate complex normal distribution on $\mathbb{C}^n$ with mean ν, covariance $\Xi = \mathbb{E}_z[(z-\nu)(z-\nu)^*]$, and density function

$$f_{\mathcal{CN}}(z; \nu, \Xi) = \frac{1}{\pi^n \det(\Xi)} \exp(-(z-\nu)^* \Xi^{-1} (z-\nu))$$

We remark that circular symmetry is defined by having that the real and imaginary parts of $z \sim \mathcal{CN}(0, \Xi)$ be i.i.d., and is ensured by tacitly requiring, as we shall throughout this dissertation, that $\mathbb{E}[(z-\nu)(z-\nu)^T] = 0$ (see Theorem 7.8.1 of [43]).

We now relate a standard result from the literature concerning linear transformations of Gaussian random vectors.

Proposition 9 (Theorem 3.3.3 in [93] and Section 7.8.1 of [43])**.** *Suppose* $x \sim \mathcal{N}(\mu, \Sigma)$.

Then for $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$,

$$Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T). \tag{4.39}$$

Suppose $z \sim \mathcal{CN}(\nu, \Xi)$. *Then for* $B \in \mathbb{C}^{m\times n}, c \in \mathbb{C}^m$,

$$Bz + c \sim \mathcal{CN}(B\nu + c, B\Xi B^*). \tag{4.40}$$

We remark that the result regarding complex Gaussian vectors implies that linear transformations preserve the property of circular symmetry.

4.4.3 Distribution of variance

In the interest of studying the propagation of error through our recovery algorithm, in this section, we describe the probability distribution of the noise on each entry of $\mathcal{A}^{-1}(\mathcal{A}(xx^*)+\eta)$ as a function of the noise vector's distribution. To keep things tractable, we assume that the entries of η are identically and independently distributed; specifically, we will assume that $\eta_{(\ell,j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ for some $\sigma \geq 0$.

Before we begin, we remark that the distribution of $\mathcal{A}^{-1}(\eta)_{i,i+m}$ will depend only on m. This follows intuitively by noting that $\mathcal{A}$ commutes nicely with "diagonal shifts" in the sense that

$$\mathcal{A}(S^k X S^{-k})_{(\ell,j)} = \langle S^\ell m_j m_j^* S^{-\ell}, S^k X S^{-k}\rangle = \langle S^{\ell-k} m_j m_j^* S^{-(\ell-k)}, X\rangle = \mathcal{A}(X)_{(\ell-k,j)},$$

so that if $y_1, y_2 \in \mathbb{R}^{[d]\times[D]}$ satisfy $(y_1)_{\ell,j} = (y_2)_{\ell-k,j}$ for some $k \in \mathbb{N}$, we will have $\mathcal{A}^{-1}(y_1) = S^k\mathcal{A}^{-1}(y_2)S^{-k}$. In particular, if the entries of y_1 are identically, independently distributed random variables and y_2 is *defined* by $(y_2)_{\ell-k,j} = (y_1)_{\ell,j}$, then y_1 and y_2 are drawn from the same distribution on $\mathbb{R}^{[d]\times[D]}$. This means that $\mathcal{A}^{-1}(y_1)$ and $\mathcal{A}^{-1}(y_2)$ are identically distributed, but $\mathcal{A}^{-1}(y_1) = S^k\mathcal{A}^{-1}(y_2)S^{-k}$, so the distributions of $\mathcal{A}^{-1}(y_1)$ and $S^k\mathcal{A}^{-1}(y_1)S^{-k}$

are identical. In particular, the distribution of the image of i.i.d. noise under $\mathcal{A}^{-1}$ is invariant under such diagonal shifts, so $\mathcal{A}^{-1}(\eta)_{i,i+m}$ is distributed identically to (though not necessarily independently from!) $\mathcal{A}^{-1}(\eta)_{1,1+m}$, and this conclusion holds for all i and m.

To make this precise, and to discover the distribution of $\mathcal{A}^{-1}(\eta)_{1,1+m}$ exactly, we will present another means by which $\mathcal{A}^{-1}(y)$ may be calculated from y. With this in mind, we remark that (4.12) of Lemma 12 gives that

$$P^{(D,d)}(F_d \otimes I_D) = \begin{bmatrix} I_D \otimes f_1^d & \cdots & I_D \otimes f_d^d \end{bmatrix},$$

so A may be transformed by (4.14) to give

$$A = \begin{bmatrix} M_1 \otimes f_1^d & \cdots & M_d \otimes f_d^d \end{bmatrix} \begin{bmatrix} I_D \otimes f_1^d \\ \vdots \\ I_D \otimes f_d^d \end{bmatrix} = \sum_{j=1}^{d} M_j \otimes f_j^d f_j^{d*}.$$

Setting $X = \begin{bmatrix} \chi_{1-\delta} & \cdots & \chi_{\delta-1} \end{bmatrix} \in \mathbb{C}^{d \times 2\delta-1}$, Lemma 13 gives us

$$A \begin{bmatrix} \chi_{1-\delta} \\ \vdots \\ \chi_{\delta-1} \end{bmatrix} = \operatorname{vec}\left(\sum_{j=1}^{d} f_j^d f_j^{d*} X M_j^T \right).$$

Recalling (4.37) along with $\widetilde{F}_D^{-T} = (\widetilde{F}_D^T)^* = \overline{\widetilde{F}}_D$, we have

$$f_j^{d*} \operatorname{mat}_{(d,D)} \left(A \begin{bmatrix} \chi_{1-\delta} \\ \vdots \\ \chi_{\delta-1} \end{bmatrix} \right) \overline{\widetilde{F}}_D = f_j^{d*} X D_{z_j},$$

so that, for $\ell \in [2\delta - 1]$, and recalling $\overline{\widetilde{F}}_D e_\ell = \overline{f}^D_{\ell+1-\delta} = f^D_{\delta+1-\ell}$, we have

$$f_j^{d*} \operatorname{mat}_{(d,D)}(A \operatorname{vec}(X)) f^D_{\delta+1-\ell} = f_j^{d*} \operatorname{mat}_{(d,D)}(A \operatorname{vec}(X)) \overline{\widetilde{F}}_D e_\ell = f_j^{d*} \chi_{\ell-\delta} Z_{\ell j}.$$

In this way, from $A\,\mathrm{vec}(X)$ we may recover

$$b_\ell := F_d^* \,\mathrm{mat}_{(d,D)}(A\,\mathrm{vec}(X)) f_{\delta+1-\ell}^D = \mathrm{vec}(f_j^{d*} Z_{\ell j} \chi_{\ell-\delta})_{j=1}^d = D_{Z^T e_\ell} F_d^* \chi_{\ell-\delta}$$

for each ℓ, from which $\chi_{\ell-\delta}$ is determined by taking $\chi_{\ell-\delta} = F_d D^{-1}_{Z^T e_\ell} b_\ell$. In other words, for $y \in \mathbb{R}^{dD}$, the m^{th} diagonal of $\mathcal{D}_\delta^{-1}(A^{-1}y)$, for $m = 1-\delta, \ldots, \delta-1$, is given by

$$\chi_m = F_d D^{-1}_{Z^T e_{m+\delta}} F_d^* \,\mathrm{mat}_{(d,D)}(y) f_{1-m}^D, \tag{4.41}$$

and we use this expression to deduce the distribution of noise on the m^{th} diagonal when y is a random variable. For instance, if we consider that

$$\mathcal{D}_\delta^{-1} A^{-1}(A\mathcal{D}_\delta(T_\delta(xx^*)) + \eta) = T_\delta(xx^*) + \mathcal{D}_\delta^{-1}(A^{-1}\eta),$$

where $\eta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma^2)$ for $j \in [d(2\delta-1)]$, knowing the distribution of $A^{-1}\eta$ will tell us the distribution of noise on our recovered estimate of $T_\delta(xx^*)$. We deduce this result directly from (4.41), and summarize it in Proposition 10. We remark that, in the statement and proof of this proposition, exponentiation of vectors is entry-wise.

Proposition 10. *Suppose that $\{m_j\}_{j=1}^{2\delta-1}$ is a local Fourier measurement system with support $\delta \le \frac{d+1}{2}$, mask γ, and modulation index $K = D = 2\delta-1$, with associated measurement operator and representation matrix $\mathcal{A}$ and A. Suppose further that $\eta \in \mathbb{R}^{d(2\delta-1)}$ has entries that are i.i.d. Gaussian random variables, namely $\eta_j \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$ for $j \in [d(2\delta-1)]$ and some $\sigma \ge 0$. Then, setting $N = \mathcal{D}_\delta^{-1}A^{-1}\eta \in T_\delta(\mathbb{C}^{d\times d})$, then N is Hermitian, its $0,\ldots,\delta-1^{st}$ diagonals are distributed independently from one another, and the m^{th} diagonal of N is distributed by*

$$\begin{aligned}
\mathrm{diag}(N,m) &\sim \mathcal{CN}(0,\sigma^2\,\mathrm{circ}(s_m)), m \in [\delta-1], \text{ and} \\
\mathrm{diag}(N,0) &\sim \mathcal{N}(0,\sigma^2\,\mathrm{circ}(s_0)), \text{ where} \\
s_m &= \frac{1}{Dd^{3/2}} F_d^* |F_d^* g_m|^{-2}.
\end{aligned} \tag{4.42}$$

Proof of Proposition 10. To establish that N is Hermitian, we remark that $\mathbb{C}^{d\times d} = \mathcal{H}^d \oplus \mathrm{Skew}(d)$, where

$$\mathrm{Skew}(d) = \{B \in \mathbb{C}^{d\times d} : B^* = -B\}$$

is the set of skew-Hermitian matrices and where $\oplus$ represents the direct product. In other words, given any $M \in \mathbb{C}^{d\times d}$, we have a unique $H \in \mathcal{H}^d, B \in \mathrm{Skew}(d)$ such that $M = H + B$. We now consider that, if $H \in \mathcal{H}^d$ and $B \in \mathrm{Skew}(d)$, we have that

$$\begin{aligned}\mathrm{Tr}(H^*B) &= \mathrm{Tr}(HB) = \mathrm{Tr}(BH)\\ &= -\mathrm{Tr}(B^*H) = -\overline{\mathrm{Tr}(H^*B)},\end{aligned}$$

meaning that $\mathrm{Tr}(H^*B) \in \mathrm{i}\mathbb{R}$. Additionally, we remark that the Hermitian matrices are a real Hilbert space, so for another $H' \in \mathcal{H}^d$, we have $\mathrm{Re}\langle H', M\rangle = \langle H', H\rangle$ and $\mathrm{Im}\langle H', M\rangle = \langle H', B\rangle/\mathrm{i}$. Therefore, since all the measurement matrices $S^\ell m_j m_j^* S^{-\ell}$ appearing in $\mathcal{A}$ are Hermitian, given $M \in \mathbb{C}^{d\times d}$ decomposed into its Hermitian and skew-Hermitian parts $M = H + B$, we have that $\mathrm{Re}\,\mathcal{A}(M) = \mathcal{A}(H)$ and $\mathrm{Im}\,\mathcal{A}(M) = \mathcal{A}(B)/\mathrm{i}$. In particular, $\mathcal{A}^{-1}(\mathbb{R}^{[d]_0\times[D]}) \subseteq \mathcal{H}^d$, and N, being the inverse image of a real vector η, will be Hermitian.

For convenience, throughout the remainder of the proof we will set $\chi_m = \mathrm{diag}(N, m)$ for $m = 1-\delta, \ldots, \delta - 1$. To prove the independence of $\{\chi_m\}_{m\in[\delta]_0}$, we consider that (4.41) tells us that

$$\chi_m = F_d D^{-1}_{Z^T e_{m+\delta}} F_d^* \left(\mathrm{mat}_{(d,D)}(\eta) f^D_{1-m}\right),$$

and we focus on the term $\mathrm{mat}_{(d,D)}(\eta) f^D_{1-m}$. Considering that $\eta \sim \mathcal{N}(0, \sigma^2 I_{dD})$, we have that the rows $\begin{bmatrix} r_1 & \cdots & r_d\end{bmatrix}^T$ of $\mathrm{mat}_{(d,D)}(\eta)$ are distributed according to $r_i = \mathrm{mat}_{(d,D)}(\eta)^T e_i \sim \mathcal{N}(0, \sigma^2 I_D)$. At this point, it would be convenient if we could merely cite Proposition 9 to establish the distribution of $r_i^T f^D_{1-m}$ or indeed $\mathrm{mat}_{(d,D)}(\eta) f^D_{1-m}$, but we remark that we don't have a result for the image of a real Gaussian vector under a complex linear transformation. Therefore, we consider the real and imaginary parts of $\mathrm{mat}_{(d,D)}(\eta) f^D_{1-m}$ separately, setting $v_m = \mathrm{mat}_{(d,D)}(\eta) f^D_{1-m}$ and seeing that

$$(v_m)_i = r_i^T \,\mathrm{Re}(f^D_{1-m}) + \mathrm{i}\, r_i^T \,\mathrm{Im}(f^D_{1-m}) = \mathrm{Re}(f^D_{1-m})^T r_i + \mathrm{i}\,\mathrm{Im}(f^D_{1-m})^T r_i.$$

Since $\{f_1^D\} \cup \{\sqrt{2}\,\mathrm{Re}(f_{1-m}^D)\}_{m\in[\delta-1]} \cup \{\sqrt{2}\,\mathrm{Im}(f_{1-m}^D)\}_{m\in[\delta-1]}$ is an orthonormal basis for $\mathbb{R}^D$, then compiling these vectors into a matrix Q, we have from Proposition 9 that $Q^T r_i \sim \mathcal{N}(0, \sigma^2 Q^T Q) = \mathcal{N}(0, \sigma^2 I_D)$, meaning the real and imaginary components of $v_1, \ldots, v_{\delta-1}$ and v_0 are all independent, with $v_0 \sim \mathcal{N}(0, \sigma^2 I_d), \mathrm{Re}(v_m) \sim \mathcal{N}(0, \frac{\sigma^2}{2} I_d)$, and $\mathrm{Im}(v_m) \sim \mathcal{N}(0, \frac{\sigma^2}{2} I_d)$ for $m \in [\delta - 1]$. Therefore, $v_m \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, \sigma^2 I_d)$ and $v_0 \sim \mathcal{N}(0, \sigma^2 I_d)$, independently from the other v_m. Since the v_m are independent, clearly their images under non-singular, fixed (independently of the random process) linear transformations will also be independent. In particular, the diagonals $\chi_m = F_d D^{-1}_{Z^T e_{m+\delta}} F_d^* v_m$ will be independent of one another for $m \in [\delta]_0$.

To get the actual distribution of the χ_m for $m \in [\delta - 1]$, we simply quote Proposition 9 again to get that

$$\chi_m \sim \mathcal{CN}(0, \sigma^2 F_d D^{-1}_{Z^T e_{m+\delta}} F_d^* I_d F_d D^{-1*}_{Z^T e_{m+\delta}} F_d^*) = \mathcal{CN}(0, \sigma^2 F_d D^{-2}_{|Z^T e_{m+\delta}|} F_d^*).$$

The covariance matrix becomes, by recalling (4.16) and the definition of Z in (4.36),

$$\sigma^2 F_d D^{-2}_{|Z^T e_{m+\delta}|} F_d^* = \mathrm{circ}\left(d^{-1/2} F_d^* |Z^T e_{m+\delta}|^{-2}\right) = \mathrm{circ}(s_m).$$

The only distinction for χ_0 is that the distributions are all $\mathcal{N}$ instead of $\mathcal{CN}$; the same calculations as above give $\chi_0 \sim \mathcal{N}(0, \sigma^2\,\mathrm{circ}(s_0))$. To verify that $s_0 \in \mathbb{R}^d$, we consider that $g_0 = \gamma \circ \gamma$, so $\tilde{g} := F_d^* g_0 = \frac{1}{\sqrt{d}}(F_d^*\gamma) * (F_d^*\gamma)$ satisfies $\tilde{g}_i = \tilde{g}_{2-i}$ (equivalently, $R\tilde{g} = \tilde{g}$). Similarly, $R(\tilde{g}^{-2}) = \tilde{g}^{-2}$, which guarantees that $s_0 = \frac{1}{D\sqrt{d}} F_d^* \tilde{g}^{-2} \in \mathbb{R}^d$. □

4.4.4 Slight Improvement to Magnitude Estimation Bounds

One useful corollary of the calculations in Section 4.4.3 allows us to sharpen our bounds on the error of the magnitude estimation step of Algorithm 1 studied in Section 3.5. In particular, because (4.41) shows us how to calculate a specific diagonal χ_m from the measurement data y, we can get a bound on $\|\mathrm{diag}(X - X_0)\|_1$ that is sharper than the $\|X - X_0\|_F$ that we used in the proof of Lemma 7. We briefly state and prove this improvement

here, and cite it later in the finalized, sharpest bounds presented in Sections 5.3 and 7.4.

To describe precisely what we mean, we define the diagonal projection operators $P_{\mathrm{diag}(\mathbb{C}^{m\times n},\ell)} : \mathbb{C}^{m\times n} \to \mathbb{C}^m$ by $P_{\mathrm{diag}(\mathbb{C}^{m\times n},\ell)}(A) = \mathrm{diag}(A,\ell)$. With this in hand, we consider that (4.41) may be rewritten as

$$(P_{\mathrm{diag}(\mathbb{C}^{d\times d},m)} \circ A^{-1})y = F_d D^{-1}_{Z^T e_{m+\delta}} F_d^* \,\mathrm{mat}_{(d,D)}(y) f^D_{1-m}, \tag{4.43}$$

where all terms are defined as in Section 4.4.3. We may rewrite this, using Lemma 13, as

$$(P_{\mathrm{diag}(\mathbb{C}^{d\times d},m)} \circ A^{-1})y = (f^{DT}_{1-m} \otimes F_d D^{-1}_{Z^T e_{m+\delta}} F_d^*)y.$$

The singular values of $f^{DT}_{1-m} \otimes F_d D^{-1}_{Z^T e_{m+\delta}} F_d^*$ are the singular values of $F_d D^{-1}_{Z^T e_{m+\delta}} F_d^*$, which in turn are simply the entries of $Z^T e_{m+\delta}$. When we consider the main diagonal χ_0, we notice from (4.36) that this gives $\sigma = \sqrt{dK}|F_d^*(\gamma \circ \gamma)|$. We state this result, along with specific bounds on $\sigma_{\min}$ and κ for $\gamma_{\exp}$ and γ_{flat} studied in Section 4.3, in Proposition 11.

Proposition 11. *Consider a local Fourier measurement system with support δ with $D := 2\delta - 1 \le d$, mask γ, and matrix representation $A \in \mathbb{C}^{dD\times dD}$. Set $Y = (P_{\mathrm{diag}(\mathbb{C}^{d\times d},0)} \circ A^{-1})$. Then*

$$\begin{aligned} \sigma_{\max}(Y) &= \sqrt{D}\|\gamma\|_2^2 \\ \sigma_{\min}(Y) &= \sqrt{dD}\min_{j\in[d]}|f_j^{d*}(\gamma\circ\gamma)|. \end{aligned} \tag{4.44}$$

When $\gamma = \gamma_{\exp}$ or γ_{flat} with the recommended parameter a_δ from Propositions 5 and 22, then we have, for $\delta \ge 5$,

$$\begin{aligned} \sigma_{\min}(Y_{\exp}) &\ge 20\sqrt{D},\ \ \kappa(Y_{\exp}) \le \delta/2 \\ \sigma_{\min}(Y_{\mathrm{flat}}) &\ge \frac{\sqrt{D}}{6}(\delta-1)^2,\ \ \kappa(Y_{\mathrm{flat}}) \le 1 + \frac{18}{\delta-1} \end{aligned} \tag{4.45}$$

Proof of Proposition 11. The main statement, (4.44), is proven in the remarks preceding the proposition. The inequalities of (4.45) are proven by straightforward calculation.

To get $\sigma_{\min}(Y_{\exp})$, we write that, for any $\omega \in \mathbb{S}^1$,

$$\begin{aligned}\frac{1}{\sqrt{D}}\sigma_{\min}(Y_{\exp}) &\geq \left|\sum_{i=1}^{\delta} a_\delta^{2(i-1)}\omega^{i-1}\right| = \left|\frac{1-a_\delta^{2\delta}\omega^\delta}{1-a_\delta^2\omega}\right| \\ &\geq \frac{a_\delta^{2\delta}-1}{a_\delta^2+1} = \frac{\left(1+\frac{4}{\delta-2}\right)^\delta - 1}{2+\frac{4}{\delta-2}},\end{aligned}$$

which may be shown numerically to exceed 20. For $\kappa(Y_{\exp})$, we use $\frac{1}{\sqrt{D}}\sigma_{\min}(Y_{\exp}) \geq \frac{a_\delta^{2\delta}-1}{a_\delta^2+1}$ and $\frac{1}{\sqrt{D}}\sigma_{\max}(Y_{\exp}) = \frac{a_\delta^{2\delta}-1}{a_\delta^2-1}$ to get

$$\kappa(Y_{\exp}) \leq \frac{a_\delta^2+1}{a_\delta^2-1} = \frac{2+\frac{4}{\delta-2}}{\frac{4}{\delta-2}} = \delta/2.$$

To get $\sigma_{\min}(Y_{\text{flat}})$, we remark that

$$\begin{aligned}\frac{1}{\sqrt{D}}\sigma_{\min}(Y_{\text{flat}}) &\geq (a_\delta+1)^2 - (\delta-1) = 4|C_0|^2\delta^2 + (4|C_0|-1)\delta + 2 \\ &\geq \frac{1}{6}\delta^2 - \frac{1}{6}\delta + 2 \geq \frac{1}{6}(\delta-1)^2,\end{aligned}$$

where we have merely used numerical bounds on the constant $|C_0| \approx 0.434467$ discussed in the proof of Proposition 22. For $\kappa(Y_{\text{flat}})$, we use that $\frac{1}{\sqrt{D}}\sigma_{\max}(Y_{\text{flat}}) = 4C_0^2\delta^2 + (4C_0+1)\delta + 2 = \frac{1}{\sqrt{D}}\sigma_{\min}(Y_{\text{flat}}) + 2(\delta-1)$, such that

$$\kappa(Y_{\text{flat}}) = 1 + \frac{2\delta-1}{\frac{1}{\sqrt{D}}\sigma_{\min}(Y_{\text{flat}})} \leq 1 + \frac{18}{\delta-1},$$

where we have used that $2\delta - 1 \leq 3\delta - 3$. □

4.5 Numerical Study of Conditioning

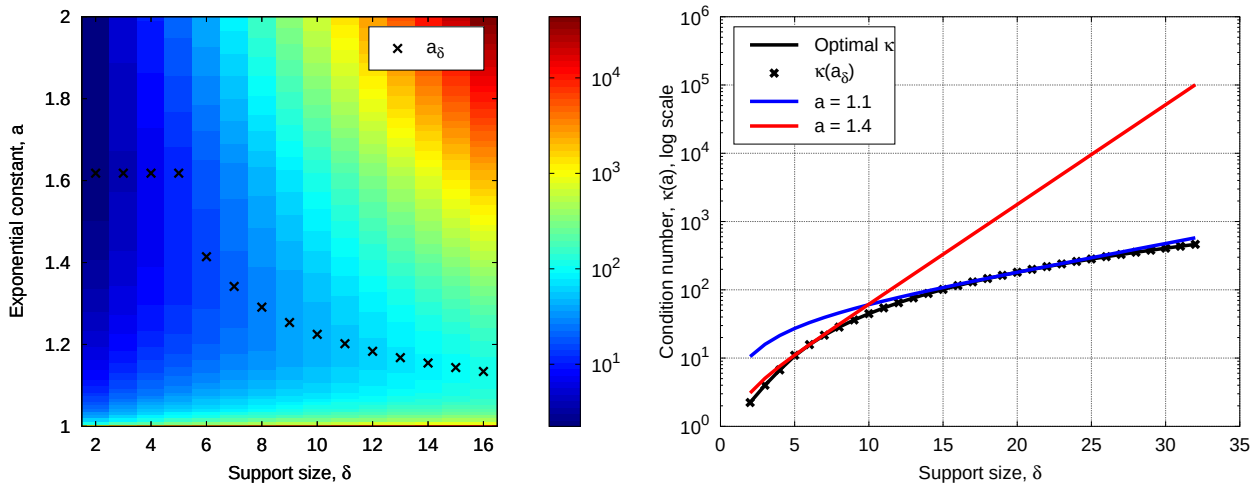
In this section, we demonstrate the results of Theorem 7 and Proposition 2 by numerically computing the conditioning of various spanning families. In Section 4.5.1, we verify the bounds for the condition number of the exponential and near-flat mask local Fourier mea-

surement systems studied in Sections 4.3.1 and 4.3.2. Additionally, we consider local Fourier measurement systems using a physically "natural" mask, where the illumination function is concentrated in the center and decays symetrically to the sides according to a Gaussian bell curve; although we have no theoretical guarantees for such a construction, conversations with crystallographers (in particular, the principal author of [85]) have suggested that this is a realistic model, so we take a practical interest in its suitability for the algorithm proposed in Chapter 3. Section 4.5.2 visualizes the propagation of variance studied in Section 4.4.3 for particular instantiations of these same masks, suggesting that the condition number belies the typical level of noise magnification. In Section 4.5.3, we perform a Monte Carlo simulation to estimate the distribution of condition numbers among randomly generated local measurement systems, both general and Fourier.

4.5.1 Conditioning of Exponential and Near-Flat Masks

Here, we make a numerical study of the conditioning of the local Fourier measurement systems analyzed in Section 4.3 to verify the tightness of the bounds we discovered in Propositions 5 and 6. Figure 4.1 shows two visualizations of κ for the exponential mask $\gamma_i = a^i, i \in [\delta]$, which depends on a, δ, and d. The first, in Fig. 4.1a, shows a heatmap representing how κ varies over a and δ for a fixed value of $d = 32$. Noticing the logarithmic scale on the colorbar, it's clear that choosing a correctly is extremely important for having a decently conditioned system: once δ exceeds 14 or so, there's a factor of greater than 10^2 between the optimal condition number and the condition number for $a = 2$. The black $\times$'s in the diagram represent the choice of a, a_δ, recommended for each value of δ in Proposition 5, and we observe that the values of a_δ appear to track the optimal basin quite closely.

We confirm this intuition further in Fig. 4.1b, where we plot κ vs. δ for four distinct choices of a as a function of δ. As described in the legend, the blue and red curves represent constant choices of a, the black curve represents an estimate of the optimal κ (computed by performing an exhaustive search to a resolution of $1/128$ over $a \in (1, 2]$ for each δ), and the black $\times$'s represent $\kappa(a_\delta)$, as in Fig. 4.1a. This figure shows precisely the behavior we would

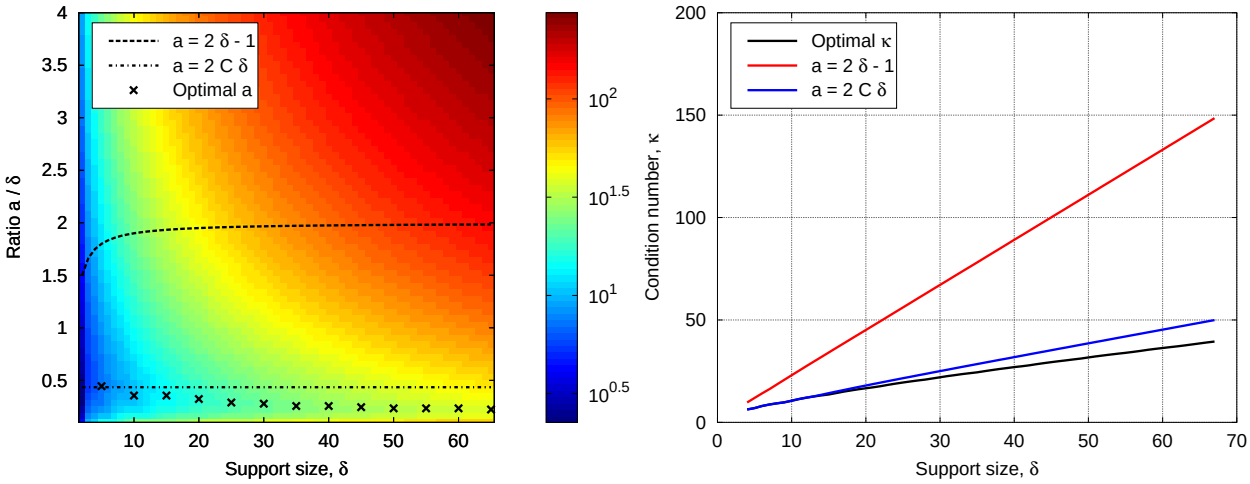


(a) $\log \kappa$ over a and δ, $d = 32$ (b) κ over δ; $d = 80$

Figure 4.1: Condition number for exponential mask local measurement systems

expect if the bound of (4.25) were tight: for a fixed a, the condition number is asymptotically bounded by $Ca^{\delta+1}$, which on a log scale becomes $\delta \log a + b$. We see that, after about $\delta = 10$, both of the constant a curves appear linear. Of course, they intersect the "optimal a" curve when the optimal a coincides with these constant choices, which may be judged by comparing with Fig. 4.1a. The other interesting result is that this chart corroborates our expectation that the recommended choice a_δ is close to optimal: indeed, the $\kappa(a_\delta)$ marks are practically indistinguishable from the optimal κ curve. We should expect this observation at least asymptotically in δ, since our bound on $\kappa(a_\delta)$ is asymptotically tight: the value $a_\delta = (1 + \frac{4}{\delta-2})^{1/2}$ optimizes the expression $\frac{a^{\delta+2}}{(a^2-1)^2}$, which is an intermediate term in (4.29), and every inequality in (4.29) is individually tight as d and δ grow large.

In Fig. 4.2, we make a similar visualization of the conditioning of the near-flat mask measurement systems. Figure 4.2a shows how κ varies against δ and a/δ, and we see that the recommended choice $a = 2\delta - 1$ is quite suboptimal; one simple improvement suggested by the work in Appendix B is shown, where we take $a = 2C\delta$ for $C \approx 0.217$, and the improvement is clear, although this choice still is not optimal (the optimal value of a/δ is plotted as $\times$'s). Figure 4.2b compares the condition number achieved by each of these three selections of a, where again the optimal value is obtained, to a resolution of $1/128$, by a linear search. As expected, the two choices of a which are affine in δ produce condition



(a) κ over a/δ and δ, $d = 6\delta - 1$

(b) κ over δ, $d = 6\delta - 1$

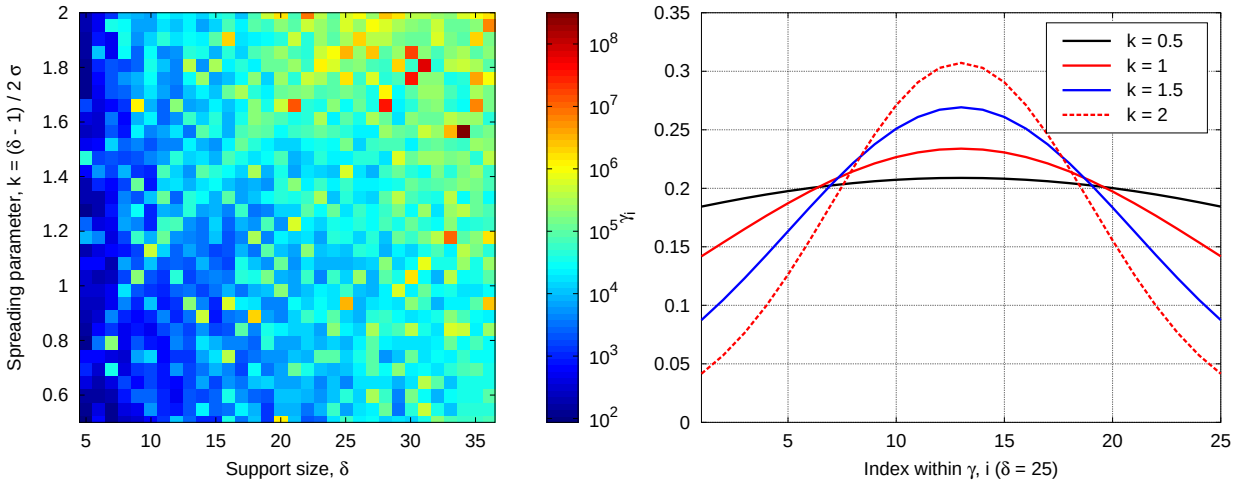
Figure 4.2: Condition number for near-flat mask local measurement systems

numbers roughly linear in δ, and the optimal value doesn't appear to be of a lower order. At any rate, the guarantees provided in Propositions 6 and 22 are validated, at least in upper bounding the growth of κ with the size of the problem.

Figure 4.3a gives a heatmap of the condition number for the linear system corresponding to a "natural" illumination function, where $\gamma \in \mathbb{R}^d$ is defined by

$$\gamma_i = \begin{cases} C\mathrm{e}^{-\frac{(i-\mu)^2}{2\sigma^2}}, & i \in [\delta] \\ 0, & \text{otherwise} \end{cases}$$

with $\mu = \frac{\delta+1}{2}$ (so that the illumination is concentrated at the center of the mask), C chosen such that $\|\gamma\|_2 = 1$, and $\sigma = \frac{\delta-1}{2k}$, where k is the width of $[\mu, \delta]$ in standard deviations. We consider this choice of γ to be natural in the sense that it roughly models the brightness of a laser beam which is pointed at the center of our window $[1, \delta]$. To visualize how concentrated or spread out are the masks under consideration in Fig. 4.3a, γ is plotted for a few sample values of k in Fig. 4.3b. Unfortunately, though, as we can see in Fig. 4.3a, the conditioning of masks of this variety is overwhelmingly unstable, with little to no recognizable pattern in the dependence of κ on k and δ. In many spots, varying δ by 1 (liable to occur as an artifact of discretization) or changing k by ≈ 0.05 (not a large change; see Fig. 4.3b) can worsen κ



(a) κ over k and δ, $d = 6\delta - 1$ (b) γ for different values of $k, \delta = 25$

Figure 4.3: Gaussian illumination masks

by multiple orders of magnitude. As κ is smooth neither in δ nor k, these types of masks appear to be unsuitable for Algorithm 1.

4.5.2 Numerical Distribution of Variance

Figures 4.4 and 4.5 illustrate the results of Proposition 10 in section Section 4.4.3. The experiment run for Fig. 4.5 was to fix a family of masks $\{m_j\}_{j\in[D]}$ and solve the linear system $\mathcal{A}(X) = y$ for several pseudo-randomly generated vectors of measurements $y \sim \mathcal{N}(0, I_{dD})$, and then calculate and plot the sample variance for each entry of $X = \mathcal{A}^{-1}(y)$. We repeated this for one particular realization (fixing d, δ, and parameters) of the exponential, near-flat, and Gaussian masks discussed in Section 4.5.1; we will refer to these three mask vectors as $\gamma_{\text{exp}}, \gamma_{\text{flat}}$, and γ_{gauss} for the remainder of this section. The figure captions in Fig. 4.5 show the parameters d, δ, and a (or k) used for each experiment. These heatmaps display $d \times d$ grids, where the $(i, j)^{\text{th}}$ pixel represents the sample variance of X_{ij}, and the perfectly blue bands observed are simply the parts of X that are "zeroed out" by the $T_\delta(X)$ restriction. First of all, we remark the scales used on each of these graphs: the exponential and near-flat masks exhibit variance at similar orders of magnitude, while the Gaussian measurement system (using masks from Fig. 4.3b with $a = 2$) produces variance on the entries that is

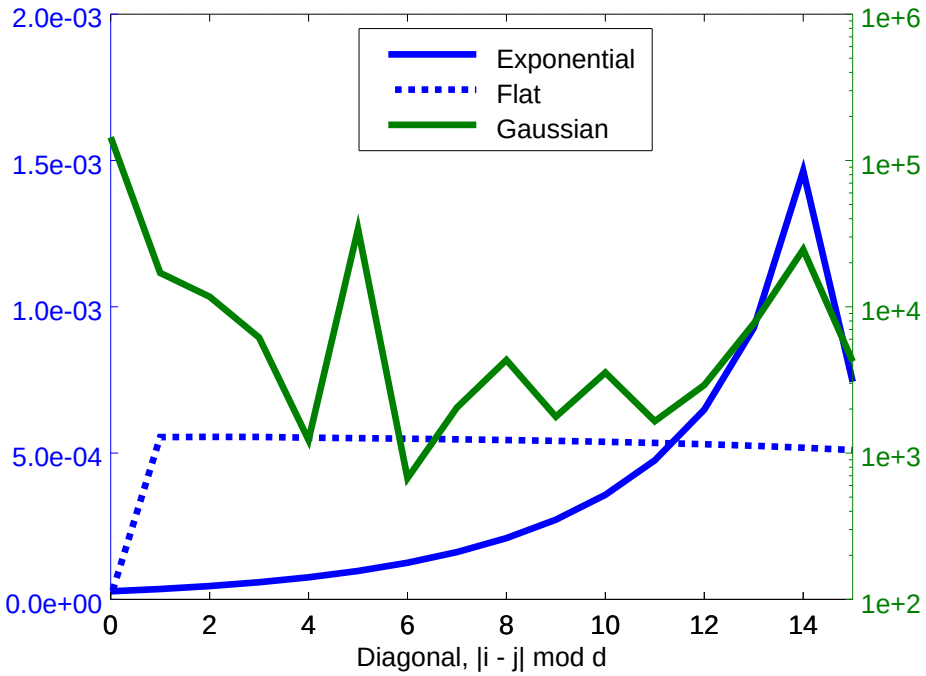


Figure 4.4: Actual variance $|s_m|_1$ along each diagonal for exponential, flat, and Gaussian masks. $d = 64, \delta = 16$.

several orders of magnitude higher. This is absolutely an issue with the conditioning of the Gaussian system, since our design of the masks has given us

$$\|\gamma_{\text{exp}}\|_2^2 = 191.65, \|\gamma_{\text{flat}}\|_2^2 = 78.226, \|\gamma_{\text{gauss}}\|_2^2 = 1,$$

so unit-variance noise on the γ_{gauss} measurements ought to be no more than $200x$ as significant as similar noise on the exponential or flat measurements, by a rough Cauchy-Schwarz estimate:

$$\mathcal{A}(xx^*)_{(\ell,j)} = |\langle x, S^\ell m_j \rangle|^2 \leq \|\gamma\|_2^2 \|x\|_2^2.$$

And this result – where the Gaussian masks are far less competitive – is unsurprising, considering the horrendous stability observed in Fig. 4.3a.

More interesting, then, is to view how the variance is distributed over the diagonals when using the exponential and flat masks. Referring to Figs. 4.5a and 4.5b, the main diagonal appears to have the most accurate entries in both cases. For the flat masks, the

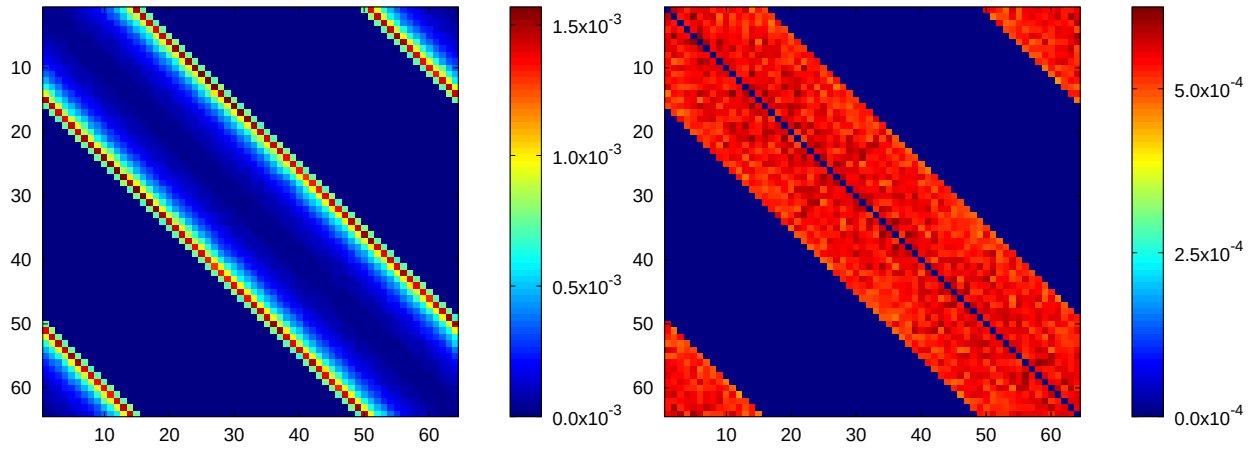
off-diagonal entries of $\mathcal{A}^{-1}(y)$ appear to have roughly equal uncertainty, while the entries almost monotonically increase in variance as you get farther away from the main diagonal with the exponential masks. Interestingly, the penultimate off-diagonal has the greatest variance with the exponential masks. These observations are confirmed in Fig. 4.4, where we have used Proposition 10 to calculate the *actual* variance (specifically, we calculate $|s_m|_1$ as a function of $m \in [\delta]_0$) of the entries of $\mathcal{A}^{-1}(y)$) for all three of these measurement systems. We emphasize that the Gaussian variance curve uses the logarithmic scale on the right side of the graph, while the other two curves use the linear scale on the left.

One useful feature of this analysis is that it suggests that randomly-distributed noise (in the sense that $y_{(\ell,j)} = |\langle x, S^\ell m_j \rangle|^2 + \eta_{(\ell,j)}$ with $\eta_{(\ell,j)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$) will typically cooperate with the magnitude estimation step in line 4 of Algorithm 1 when we use exponential or near-flat masks. In particular, if $x_0 \in \mathbb{C}^d$ is the ground truth and $X = \mathcal{A}^{-1}(y)$ is our estimate of $X_0 = T_\delta(x_0 x_0^*)$, then we might expect $\|\text{diag}(X - X_0)\|_1$ to be considerably smaller than $d^{1/4}\|X - X_0\|_F$; considering Lemma 7 along with these numerical results, this suggests that typical noise will not be adversarial to our simple magnitude estimation technique.

4.5.3 Conditioning of Randomized Measurements

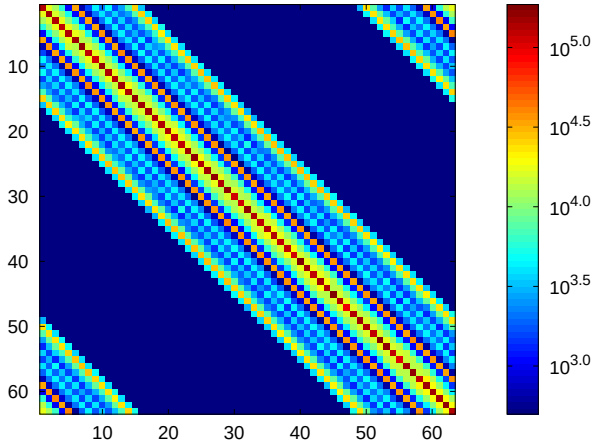
In Section 4.2.3, we proved that local Fourier measurement systems that are spanning families are "generic," meaning the set $\{\gamma \in \mathbb{R}^d : \text{span}_{\mathbb{R}} \mathcal{L}_\gamma \neq T_\delta(H^d)\}$ has Lebesgue measure zero. In turn, this means that drawing $\gamma \sim \mathcal{D}$ from any probability distribution $\mathcal{D}$ which is absolutely continuous with respect to $\mathcal{N}(0, I_\delta)$ will produce a spanning local Fourier measurement system with probability 1.[1] In this section, we verify this result numerically, with three main questions in mind: how are the condition numbers of such randomly generated local Fourier measurement systems distributed? Does a similar genericness result hold for general local measurement systems, and how are these systems typically conditioned? Finally, in the style of Fig. 4.4, what is the "typical" distribution of entry-wise variances?

[1]Here, we write that we are drawing γ from a probability distribution on $\mathbb{R}^\delta$, since $\gamma = \mathbb{1}_{[\delta]} \circ \gamma$ only has δ non-zero entries. This is a slight overloading of notation, equivalent to identifying $\gamma \in \mathbb{R}^d$ with $\mathcal{R}_\delta \gamma \in \mathbb{R}^\delta$.



(a) Exponential masks.
$d = 64, \delta = 16, a = a_\delta \approx 1.134$

(b) Near-flat masks.
$d = 64, \delta = 16, a = 2C\delta \approx 6.952$



(c) Natural/gaussian masks.
$d = 63, \delta = 16, k = 2$

Figure 4.5: Variance of entries of $\mathcal{A}^{-1}(y)$ for $y \sim \mathcal{N}(0, I_{dD})$, $N = 512$ samples

We consider the first of these questions in Fig. 4.7. In each case, we fixed d and δ, then generated a pseudorandom γ from a folded normal distribution (so $\gamma = |g|, g \sim \mathcal{N}(0, I_\delta)$) and calculated the condition number $\kappa = \frac{\|\gamma\|_2^2/\sqrt{d}}{\min_{j,m}|f_j^{d*}(\gamma \circ S^{-m}\gamma)|}$ of $\mathcal{A}$ for each. The folded normal distribution was chosen because (a) masks are supposed to represent illumination strength, or alternatively, some attenuation imposed upon the radiation by a physical screen, each of which is best physically modelled by non-negative values of γ (b) the rotational symmetry of the Gaussian ensures that any mask $\gamma \geq 0$ is "equally likely" to be used by this distribution (we notice that $\kappa(\gamma)$ is invariant over multiplying γ by a non-zero scalar, so rotational symmetry ensures that all rays have equal probability density). This experiment was run 2048 times for each (d, δ) pair, and the resulting values of κ are visualized in the histograms of Fig. 4.7. For reference, we calculated the condition numbers of γ_{exp} and γ_{flat}, using the recommended parameters $a_{\text{exp}} = \left(1 + \frac{4}{\delta - 2}\right)^{1/2}, a_{\text{flat}} = 0.43447\delta$ in each case, and marked the two data points $\kappa_{\text{exp}}, \kappa_{\text{flat}}$ to see how our theoretically-vetted examples compare to "generic" masks.

The biggest impression we can gather from these results is that, at least up to $\delta = 100$, it appears that κ_{exp} is roughly typical, hanging around or under the mode of the distribution of κ's – so it appears that the κ's are concentrated around δ^2 – whereas κ_{flat} is consistently well under the bulk of the distribution. This last result isn't surprising, since a quick calculation shows that, for γ_i drawn i.i.d. from the folded normal distribution with $\mu = 0, \sigma = 1$, we have

$$\begin{aligned}
\mathbb{E}[\kappa] &= \mathbb{E}\left[\frac{\sum_{i=1}^{\delta}\gamma_i^2}{\sqrt{d}\min_{(j,m)\in[d]\times[\delta]_0}\left|f_j^{d*}(\gamma \circ S^{-m}\gamma)\right|}\right] \\
&\geq \mathbb{E}\left[\frac{\sum_{i=1}^{\delta}\gamma_i^2}{\gamma_1\gamma_\delta}\right] \\
&= \mathbb{E}\left[\sum_{i=2}^{\delta-1}\frac{\gamma_i^2}{\gamma_1\gamma_\delta} + \frac{\gamma_1}{\gamma_\delta} + \frac{\gamma_\delta}{\gamma_1}\right] \\
&= (\delta - 2)\frac{\pi}{2} + 2 \geq \delta,
\end{aligned}$$

where we have used independence of γ_i and

$$\mathbb{E}[\gamma_i] = \sqrt{\frac{2}{\pi}} \int_0^\infty x \mathrm{e}^{-x^2/2} \, \mathrm{d}x = -\sqrt{\frac{2}{\pi}} \mathrm{e}^{-x^2/2} \Big|_{x=0}^{\infty} = \sqrt{\frac{2}{\pi}}.$$

At the very least, this ensures that $\kappa_{\text{flat}} \leq \frac{7}{8}\delta < \mathbb{E}[\kappa]$, but empirically, we see that the second line of this inequality is very wasteful and indeed the near-flat masks are exceptionally well-conditioned.

Besides the relative suitability of the exponential and near-flat masks, we see that the distribution is qualitatively steady over several instances of d and δ: positively-skewed with a peak at δ^2, and little to no probability mass towards δ. This description holds between Fig. 4.7a and Fig. 4.7b, where we fix $d = 64$ and move δ from the somewhat small support of 5 to nearly complete support $\delta = 31$, where $T_\delta(\mathcal{H}^d)$ almost covers all of $\mathcal{H}^d$. Between Fig. 4.7b and Fig. 4.7c, we maintained this ratio of δ/d at almost $1/2$, and qualitatively, there appears to be little difference between the shapes of the histograms and the relative positions of κ_{flat} and κ_{exp} within them. Between Fig. 4.7c and Fig. 4.7d, we fixed δ and increased d to 999, and there is an interesting, if slight, impact on the distribution: we notice that the conditioning tends to get worse for the random masks, pushing the peak of the distribution to the right and tightening it. One reason for the general worsening of condition numbers when d increases is that it roughly increases the set over which the denominator of $\kappa = \frac{\|\gamma\|_2^2/\sqrt{d}}{\min_{j,m}|f_j^{d*}(\gamma \circ S^{-m}\gamma)|}$ is minimized, but this effect may be upper bounded by writing

$$\kappa \leq \frac{\|\gamma\|_2^2}{\min\limits_{m \in [\delta]_0, \theta \in \mathbb{R}} \left| \sum_{j=1}^{\delta-m} \mathrm{e}^{\mathrm{i}\theta(j-1)} \gamma_j \gamma_{j+m} \right|},$$

which we have already considered in the proofs of the bounds for κ_{flat} and κ_{exp}.

In Fig. 4.6, we look at the distribution of per-entry variance of propagated noise among randomly-distributed masks γ. Specifically, we consider the result of Proposition 10, which says that the variance of X_{ij}, when $X = \mathcal{D}_\delta^{-1} A^{-1} y$ for $y \sim \mathcal{N}(0, I_{dD})$, is given by $|s_m|_1$, where $m = i - j \bmod d$. Since s_m is a function of γ (as we may recall from (4.42)), $|s_m|_1$ is a random variable, and visualizing its distribution would be useful in ascertaining

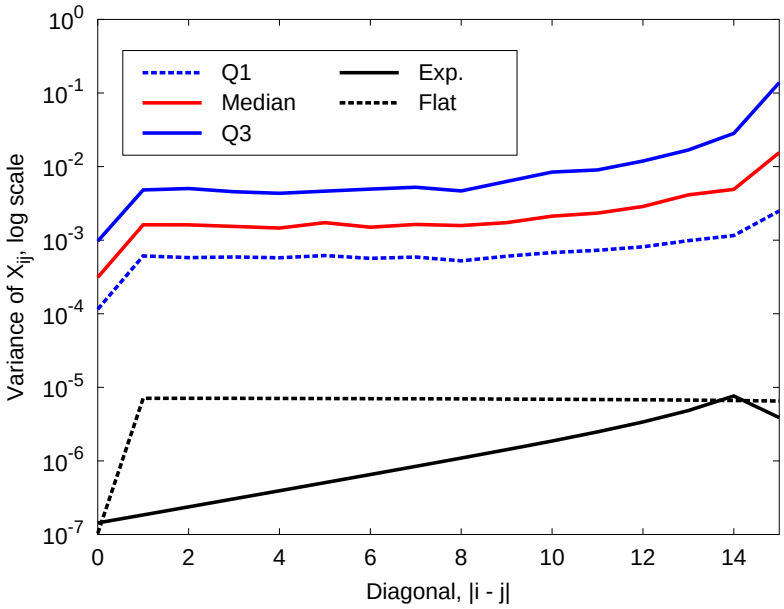
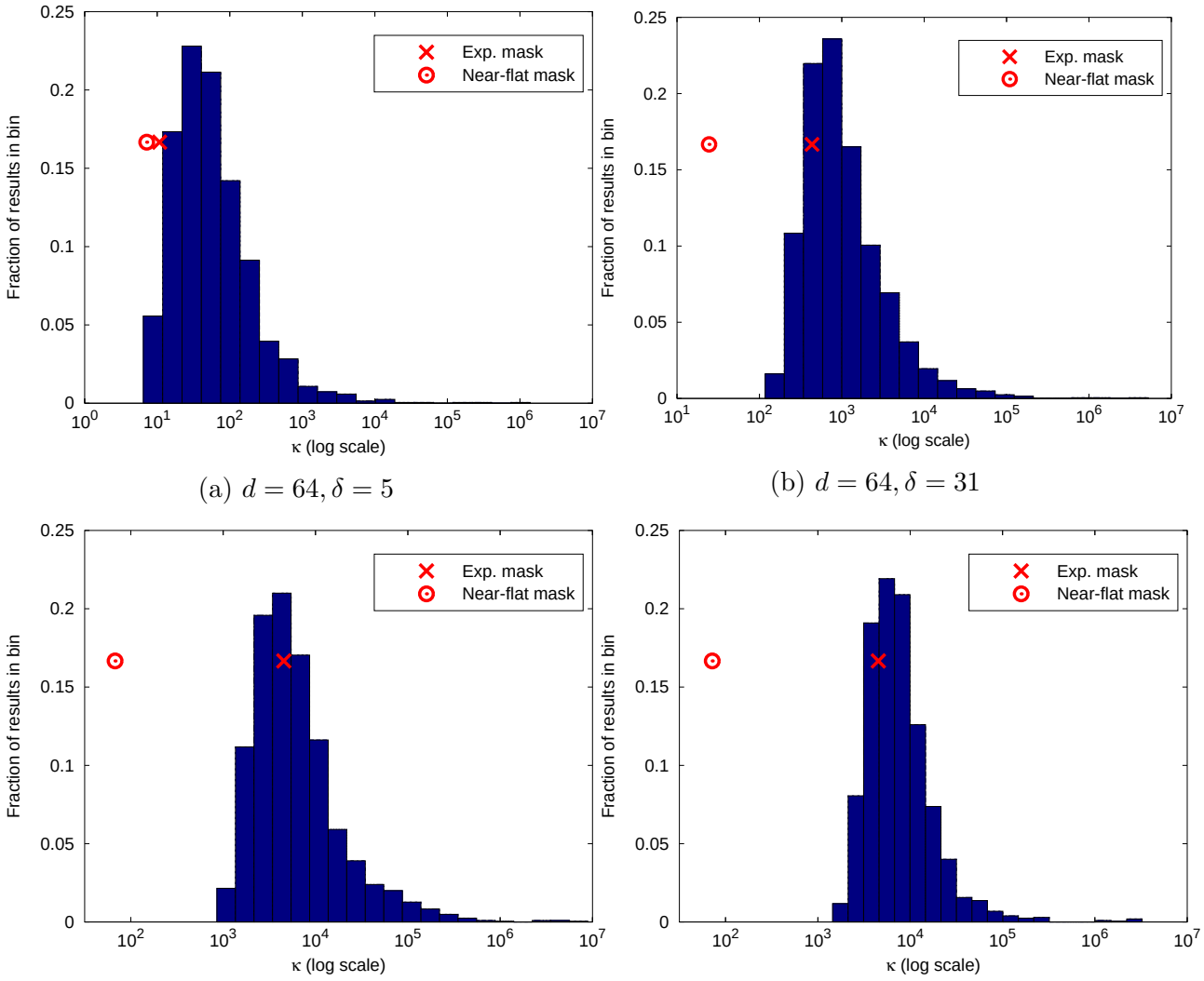


Figure 4.6: Distribution of per-entry variance among randomly generated local Fourier measurement systems. $d = 64, \delta = 16$.

the "typical factor" by which noise is magnified by $\mathcal{A}^{-1}$. To do this, we simply took 1024 pseudo-random instances of $\gamma \sim |\mathcal{N}(0, I_\delta)|$, manually calculated $\frac{|s_m|_1}{\|\gamma\|_2^2}, m \in [\delta]_0$ for each,[2] and plotted the median and quartiles of these results as a function of m in Fig. 4.6. For comparison, we repeated this calculation for γ_{exp} and γ_{flat}. The purpose of the normalization $\frac{s_m}{\|\gamma\|_2^2}$ is to make the result invariant under scalar multiplication: we want to understand the size of noise magnified by $\mathcal{A}^{-1}(y)$ relative to what we might expect of the magnitudes of the measurements $\mathcal{A}(X)$. Taking note of the logarithmic scale, we see that γ_{exp} and γ_{flat} actually perform *very* well by this metric: the worst-case diagonals for both γ_{exp} and γ_{flat} are over an order of magnitude better than the lowest quartile of random γ, and the randomized γ show a far greater tendency to degenerate towards the edges (when $|i - j| = 15$), whereas γ_{flat} especially renders an equal distribution of noise over the off-diagonals of $\mathcal{A}^{-1}(y)$.

We make a brief comparison of the conditioning of random local Fourier measurement systems versus *general* local measurement systems in Fig. 4.8. Here, we drew 1024 random $m_j \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, I_\delta), j \in [2\delta - 1]$ and 1024 random $\gamma \sim \mathcal{N}(0, I_\delta)$, manually computed the

[2]Recall that this does *not* necessitate generating $y \sim \mathcal{N}(0, I_{dD})$ and sampling the distribution of $\mathcal{A}^{-1}(y)$, since the distribution of $\mathcal{A}^{-1}(y)$ is known explicitly from Proposition 10.



(a) $d = 64, \delta = 5$

(b) $d = 64, \delta = 31$

(c) $d = 202, \delta = 100$

(d) $d = 999, \delta = 100$

Figure 4.7: Distribution of condition numbers for $\gamma \sim |\mathcal{N}(0, I_\delta)|$ for different values of d and δ

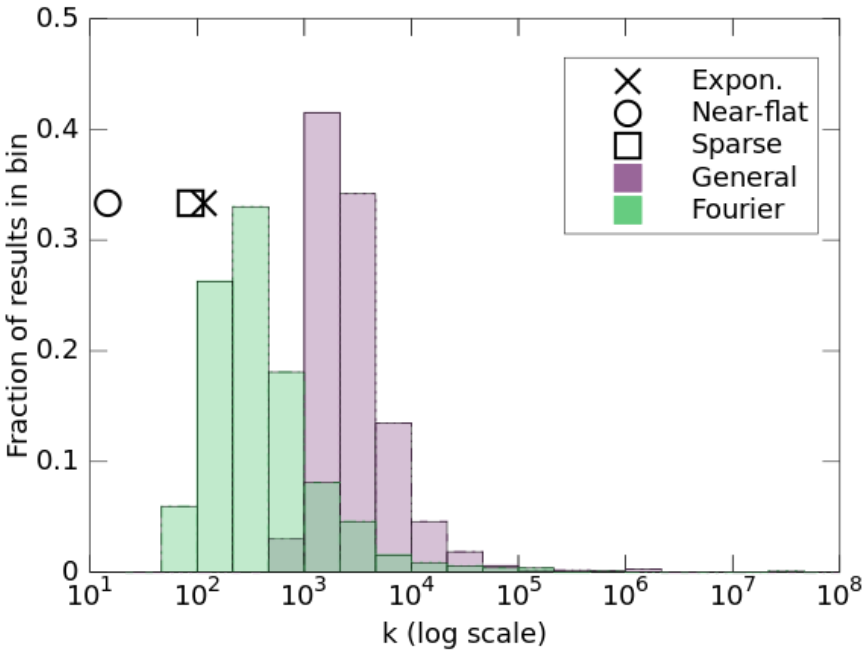


Figure 4.8: Comparison of distribution of condition numbers for random general and Fourier families of masks. $d = 64, \delta = 16$

condition numbers according to the results of Theorem 7 and Proposition 2, and produced a histogram of the results. Again, we plot κ_{exp} and κ_{flat} for reference, but this time we also include the "sparse construction" of Example 2 in Section 3.2. The comparison between general and Fourier local measurement systems is clear: the distributions have roughly the same shape, but the distribution of the Fourier measurement systems' condition numbers is translated to the left by about half an order of magnitude. This is a somewhat surprising result, since the conventional wisdom in the phase retrieval literature, especially concerning PhaseLift and Wirtinger Flow, suggests that random measurements tend to be highly well-behaved on average [23, 31, 63, 64], but we have no theoretical result concerning how these distributions ought to appear, so we cannot substantiate this intuition.

Overall, the results in Figs. 4.6–4.8 are very pleasing. They clearly substantiate our claim that the γ which produce spanning families are generic, and also show that the two examples we have provided and studied in Section 4.3 are respectively typical, in the case of γ_{exp}, and outstanding, in the case of γ_{flat}, by the metrics of condition number and variance in $\mathcal{A}^{-1}(y)$ for random $y \sim \mathcal{N}(0, I_{dD})$.

Chapter 5

Angular Synchronization

5.1 Definition and Previous Work

In this section, we consider the problem of angular synchronization, which appears as a subproblem in many approaches to phase retrieval, including ours. In particular, we make a study of it in this section in order to improve the results in Sections 3.4 and 3.5 and to supply a crucial lemma for Section 6.3.3.

Angular synchronization is the problem of recovering a vector of complex units $x_i \in \mathbb{S}^1, i \in [n]$, or $x \in (\mathbb{S}^1)^n$ from estimates of their relative phases

$$\widetilde{X}_{ij} = x_j^* x_i \eta_{ij}, (i,j) \in E$$

where $\eta_{ij} \in \mathbb{S}^1$ are the "noise terms" and $E \subseteq [n]^2$ is a list of the pairs of indices for which we have such estimates. This immediately suggests associating a graph to this problem; namely, taking the vertex set to be $V = [n]$, we set $G = (V, E)$. We also remark that this problem invites the same global phase ambiguity as phase retrieval: indeed $(\mathrm{e}^{\mathrm{i}\theta} x_i)^* (\mathrm{e}^{\mathrm{i}\theta} x_j) \eta_{ij} = x_i^* x_j \eta_{ij}$ for any $\theta \in [0, 2\pi)$. Obviously, then, our goal is to get an estimate $\tilde{x}$ that is guaranteeably close to the ground truth x in some chosen metric, say our usual $\min_{\theta \in \mathbb{R}} ||x - \mathrm{e}^{\mathrm{i}\theta} y||_2$, and

to acquire this estimate with the least computational cost. Naturally, x is not known, so we instead attempt to minimize some cost function corresponding to how well our estimate explains the measurement data, usually taking a form similar to the frustration function stated in (3.26). Often, we more simply take only the numerator of this expression,

$$\sum_{(i,j)\in E} w_{ij}|x_i - X_{ij}x_j|^2 = x^*(D - W \circ X)x, \tag{5.1}$$

where D and W are the degree and weight matrices specified in (3.4). For convenience, we define

$$L = D - W \circ X,\ \underline{L} = D - W \circ xx^*,\ \text{and}\ L_G = D - W. \tag{5.2}$$

The approach to angular synchronization that we consider in this dissertation, and the approach most studied in the literature, then is to attempt the non-convex optimization problem

$$\begin{aligned} \min_{z\in\mathbb{C}^n} \quad & z^*Lz \\ \text{s.t.} \quad & |z_i| = 1 \end{aligned} \tag{5.3}$$

The modern study of eigenvector-based methods for angular synchronization appears to have begun with a 2011 paper by Amit Singer [84], in which he proposed two ways of solving this problem which remain as the basis of the state of the art. The first is almost identical to the eigenvector-based approach that we have used in the algorithm proposed in chapter 3, and the second is a semidefinite relaxation of the same. Namely, using the unweighted adjacency matrix $A_{ij} = X_{ij}\chi_E(i,j)$, his eigenvector method solves

$$\max_{||z||_2^2=n} z^*Az$$

to find the largest eigenvector $\hat{z}$ of A and rounds to a vector of units by taking $\widetilde{\mathbf{x}} = \text{sgn}(\hat{z})$. The SDP method solves

$$\begin{aligned} \max_{Z\in\mathcal{H}^d} \quad & \text{Tr}(AZ) \\ \text{s.t.} \quad & Z \succeq 0 \\ & Z_{ii} = 1 \end{aligned} \tag{5.4}$$

In this paper, he studies the problem under a noise model where the disturbances η_{ij} are distributed according to

$$\eta_{ij} = \begin{cases} 1, & \text{with probability } p \\ \text{Unif}(\mathbb{S}^1), & \text{with probability } 1-p \end{cases},$$

such that the measurement $\widetilde{X}_{ij}$ is exact with probability p and is completely meaningless – being drawn from the uniform distribution on $\mathbb{S}^1$ – with probability $1-p$. He proves the robustness of this method in the sense that there is a probability p_c, dependent on the spectral gap and size of the graph G, for which parameter values $p > p_c$ guarantee "better than random" approximations of x with high probability. Moreover, he shows that, experimentally, both of these recovery algorithms work acceptably, if not extremely, well, with little to be gained by transferring from the eigenvector problem to the computationally more expensive semidefinite program.

The literature on angular synchronization since this paper has largely consisted of analyzing generalizations and variations of these methods. One major generalization has been to apply these methods to larger classes of group synchronization problems such as synchronization over the orthogonal groups $O(d)$ or the special Euclidean groups $SE(d)$ [8, 16, 80]. Naturally, much of the interest in this subject has been the treatment of synchronization over $SO(3)$ and $SE(3)$, as these correspond to pose estimation problems fundamental to computer vision, as in [36, 37, 41, 47]. Significant results giving guarantees of robustness, as well as proofs that these relaxations are solved *exactly* in certain cases may be found in [2, 9, 37, 80].

5.2 Tightness of SDP Relaxation

5.2.1 Introduction and Main Result

We have already presented one angular synchronization result in Section 3.4, which drew largely on [2]. We remark that this theorem lacks some generality in that the graph G is not permitted to be weighted, which restricts us from applying some knowledge that we may have about the problem. For example, suppose our relative phase measurements $\{X_{ij}\}_{(i,j)\in E}$ are disturbed by noise drawn from a fixed, phase-invariant probability distribution, say

$$X_{ij} = \operatorname{sgn}(x_i^* x_j + \epsilon_{ij}), \epsilon_{ij} = a_{ij} + \mathrm{i}\, b_{ij} \quad \text{with} \quad a_{ij}, b_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

then we will have more confidence in the relative phases represented by the larger magnitude entries of $X = \mathcal{A}^{-1}(\mathbf{y})$. It would be intuitive to use this knowledge to privilege some edges of the graph over others in the frustration function (3.26) that we are trying to minimize, say by using $w_{ij} = |X_{ij}|$. Unfortunately, theorem 4 assumes an unweighted graph and its proof technique does not readily admit a satisfactory adjustment towards weighted edges, though we shall impose one later. Therefore, we will take a distinct approach, drawing upon recent results in the literature that consider certain convex relaxations of (5.3) [9, 18, 80].

To begin this discussion, we gather our notation: $G = (V = [n], E)$ is a connected graph with a weighted adjacency matrix $W = [w_{ij}] \in \mathcal{S}^n$ satisfying $w_{ij} \geq 0$ and $w_{ij} \neq 0$ only if $(i, j) \in E$. We take $D = \operatorname{diag}(W\mathbb{1})$ to be the degree matrix. $\underline{x} \in (\mathbb{S}^1)^n$ is the ground truth vector, which we attempt to recover, and $X, \underline{X} \in \mathcal{H}^n$ are our noisy and ground truth edge data matrices, satisfying

$$X_{ij} = \begin{cases} \eta_{ij} x_i x_j^*, & (i,j) \in E \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \underline{X}_{ij} = \begin{cases} x_i x_j^*, & (i,j) \in E \\ 0, & \text{otherwise} \end{cases},$$

where $\eta_{ij} = \eta_{ji}^* \in \mathbb{S}^1$ for each $(i, j) \in E$. We then define $L, \underline{L}$, and L_G as in (5.2). Finally, given a graph $G = (V = [n], E)$ with unweighted adjacency matrix A_G, for any set $0 \in S \subseteq \mathbb{C}$,

by S^E we mean the set of matrices X satisfying $X \in \mathcal{H}^n \cap S^{n\times n}$ and $X = X \circ A_G$ (so that $X_{ij} = 0$ for all non-adjacent pairs $(i,j) \notin E$). Specifically, $\mathbb{R}^E_{++}$ will denote the set of matrices representing valid positive weightings of G and $(\mathbb{S}^1)^E$ will denote matrices containing valid assignments of relative phases to edges of G.

Towards formulating the appropriate SDP relaxations, we recall the transformation $\mathfrak{R} : \mathbb{C} \to \mathbb{R}^{2\times 2}$ defined by

$$\mathfrak{R}(a + \mathrm{i}\, b) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

It is well-known that $\mathfrak{R}$ is the canonical isomorphism from $\mathbb{C}$ into $\mathbb{R}^{2\times 2}$, and indeed if we extend it to matrices by taking, for $A \in \mathbb{C}^{m\times n}$, $\mathfrak{R}(A) \in \mathbb{R}^{2m\times 2n}$ to be a block matrix with $\mathfrak{R}(A)_{ij} = \mathfrak{R}(A_{ij})$, then it remains an isomorphism on $\mathbb{C}^{m\times n}$, and indeed it preserves the eigenvalues and eigenvectors of Hermitian matrices (see, e.g. [101, p. 101]). In particular, a Hermitian matrix $A \in \mathcal{H}^n$ is semi-definite if and only if $\mathfrak{R}(A)$ is semi-definite; we notice that the multiplicities of its eigenvalues are all doubled, as $Av = \lambda v$ implies $\mathfrak{R}(A)\mathfrak{R}(v) = \lambda\mathfrak{R}(v)$, giving that the two columns of $\mathfrak{R}(v)$ are each eigenvectors of $\mathfrak{R}(A)$ with eigenvalue λ. For convenience, we extend the inverse $\mathfrak{R}^{-1} : \bigcup_{m,n\in\mathbb{N}} \mathbb{R}^{2m\times 2n} \to \bigcup_{m,n\in\mathbb{N}} \mathbb{C}^{m\times n}$ to $\bigcup_{n\in\mathbb{N}} \mathbb{R}^{2n}$ in the obvious way, by letting

$$\mathfrak{R}^{-1} \begin{bmatrix} a \\ b \end{bmatrix} = a + bi.$$

With this, we consider SDP relaxations of (5.3). Specifically, we observe that $z^*Lz = \mathrm{Tr}(Lzz^*)$, so an equivalent optimization problem will be

$$\begin{array}{ll} \min\limits_{Z\in\mathcal{H}^n} & \mathrm{Tr}(LZ) \\ \text{s.t.} & Z_{ii} = 1 \\ & \mathrm{rank}(Z) = 1 \\ & Z \succeq 0 \end{array}, \tag{5.5}$$

where the optimizer $\hat{z}$ of (5.3) is recovered from the optimal matrix $\hat{Z}$ by merely factoring $\hat{Z} = \hat{z}\hat{z}^*$. To get a convex relaxation, we simply omit the non-convex rank one constraint,

yielding

$$\begin{aligned} \min_{Z \in \mathcal{H}^n} \quad & \mathrm{Tr}(LZ) \\ \text{s.t.} \quad & Z_{ii} = 1 \\ & Z \succeq 0 \end{aligned} \quad \cdot \tag{5.6}$$

We remark that if an optimizer $\hat{Z}$ of (5.6) is rank one, then it is also an optimizer of (5.5) since the feasible set of (5.6) is strictly larger than that of (5.5); in this case, then, factoring $\hat{Z}$ gives the global minimizer of (5.3). Considering that many results in the optimization literature (and many of the software libraries) are written for real-valued SDPs, we take advantage of these by further relaxing the feasible set, casting into the real domain with

$$\begin{aligned} \min_{Z \in \mathcal{S}^{2n \times 2n}} \quad & \mathrm{Tr}(\mathfrak{R}(L)Z) \\ \text{s.t.} \quad & Z_{ii} = I_2 \\ & Z \succeq 0 \end{aligned} \quad , \tag{5.7}$$

where $Z_{ii} = [e_{2i-1}\ e_{2i}]^* Z [e_{2i-1}\ e_{2i}]$ in this case refers to the i^{th} 2×2 diagonal block of Z. With this in mind, we state two of the algorithms for angular synchronization that will be considered in this dissertation.[1]

Algorithm 2 Angular Synchronization by Eigenvectors

Input: Connected graph $G = (V = [n], E)$, weight matrix $W \in \mathbb{R}^E_{++}$, and relative phase data $X \in (\mathbb{S}^1)^E$.

Output: A vector $x \in (\mathbb{S}^1)^n$ of phases.

1: Let $L = \mathrm{diag}(Wone) - W \circ X$ be the Laplacian of G.
2: Let $u \in \mathrm{argmin}_{\|z\|_2 = 1} z^* L z$ be an eigenvector corresponding to the smallest eigenvalue of L.
3: Return $x = \mathrm{sgn}(u)$.

At this point, we recognize the previous work existing on this problem. Namely, in [9], Bandeira, Boumal, and Singer prove that the optimizer $\hat{Z}$ of (5.6) is rank one (and therefore yields a minimizer of (5.3)) when L is sufficiently close to $\underline{L}$. Unfortunately for our purposes, this paper only considers the case when $G = K_n$ is the complete graph and the

[1] For Algorithm 3, we will actually implement the optimization with (5.7), but since all the SDPs involved are strictly feasible – we may take $Z = I_n$ for (5.6) and $\Lambda = -(\|L\|_2 + \epsilon) I_n$ for (5.9) – optimizers exist for each [95, Thm. 3.1], and these coincide when the opimization "works," as we shall see in Lemma 15.

Algorithm 3 Angular Synchronization by SDP Relaxation

Input: Connected graph $G = (V = [n], E)$, weight matrix $W \in \mathbb{R}^E_{++}$, and relative phase data $X \in (\mathbb{S}^1)^E$.

Output: A vector $x \in (\mathbb{S}^1)^n$ of phases.

1: Let $L = \operatorname{diag}(W\mathbb{1}) - W \circ X$ be the connection Laplacian of G.
2: Let $\hat{Z}\mathbb{C}^{n\times n}$ be a minimizer of (5.6) with this data.
3: Let $u \in \mathbb{C}^n$ be an eigenvector of $\hat{Z}$ corresponding to its largest eigenvalue.
4: Return $x = \operatorname{sgn}(u)$.

weights $W = \mathbb{1}\mathbb{1}^* - I_n$ are constant. A more general result appears in [80], where Rosen, Carlone, Bandeira, and Leonard prove a similar result for synchronization over $SE(d)$, of which angular synchronization is a special case. Moreover, these results allow for a weighted graph, and include a bound on $\min_{\theta\in[0,2\pi)}\|\hat{z} - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2$ in terms of $\|L - \underline{L}\|_2$ and the spectral gap of the graph. Nonetheless, we find that narrowing to the case of $SO(2)$ (equivalent to angular synchronization) allows for a tighter error bound compared to their original result. In [18], Calafiore, Carlone, and Dellaert use methods similar to those in [80] to analyze $SE(2)$ synchronization. Furthermore, and pertinent to the present work, the authors exchange the rotational components in $SO(2)$ for complex units, but they do not admit weighted graphs, nor do they supply explicit bounds on the error of their estimate or on what level of noise may be tolerated and still guarantee that their convex relaxation solves (5.3) exactly. Significantly, all three of these works supply an *a posteriori*-certifiable condition that can verify whether the solution obtained is indeed optimal for Eqs. (5.3) and (5.5)–(5.7).

The results in this chapter, particularly Theorems 8 and 9 are heavily based on this previous work, although they are original, having not appeared in this form in the literature. In particular, our result is more general than those in [9], since we consider all possible graphs rather than K_n alone. It is slightly more general than [18], since it admits weighted graphs. Although [80] applies to a far more general problem – that of $SE(d)$ synchronization – the result of Theorem 8 is slightly sharper than what is given by applying their result to angular synchronization as a special case. The present result is achieved by synthesizing proof techniques from all three papers.

Additionally, Appendix D demonstrates a mistake that appeared in the original ar-

gument of a proposition necessary to the main results of [80]. Fortunately, we were able to produce an alternative proof that establishes the same statement. The same appendix also brought in ideas from [9] that extend our "special case" bound to the general setting of $SE(d)$. With this improvement, for which the author of this dissertation is acknowledged in an erratum to be issued for [80], Theorem 8 is almost a corollary of their Theorem 12, except that we explicitly prove a basin of attraction, inside which the SDP (5.6) produces a solution $\hat{Z} = \hat{z}\hat{z}^*$ that may be factorized to obtain a minimizer of (5.3), and our bound is better by a constant factor of 2 than the bound produced by the improved result of [80].

As they appear in the original paper, Proposition 2 and Theorem 12 of [80] give us the following.

Proposition 12 (Proposition 2 and Theorem 12 in [80])**.** *There exists a constant $\beta > 0$, depending on $\underline{L}$, such that, if $\|L - \underline{L}\|_2 < \beta$, then* (5.7) *has a unique solution $\hat{Z}$ which may be factored as $\hat{Z} = RR^*$, with $R = \mathfrak{R}(\hat{z})$ where $\hat{z} \in (\mathbb{S}^1)^n$ is a global optimizer of* (5.3)*. Furthermore,*

$$\min_{\theta \in [0,2\pi)} \|\hat{z} - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 \leq 2\sqrt{\frac{n\|\underline{L} - L\|_2}{\lambda_2(L_G)}},$$

where $\lambda_2(L_G)$ is the second smallest eigenvalue of L_G.

We strengthen this result in Theorem 8 by giving β explicitly and by increasing the exponent of $\|\underline{L} - L\|_2$ in the error bound, which improves the convergence rate as $L \to \underline{L}$.

Theorem 8. *Given a connected, weighted graph $G = (V = [n], E)$, $W \in \mathbb{R}^E_{++}$ with spectral gap $\tau = \lambda_2(D - W)$ and rotational data $X \in (\mathbb{S}^1)^E$, suppose that $\hat{z}$ is a minimizer of* (5.3)*, where $L = D - W \circ X$. By $\underline{x}$ we denote the ground truth, and we take $\underline{L} = D - W \circ \underline{x}\underline{x}^*$ and $\hat{L} = D - W \circ \hat{z}\hat{z}^*$. Then if $\|L - \hat{L}\|_2 < \frac{\tau}{1+\sqrt{n}}$, $\hat{Z} = \hat{z}\hat{z}^*$ and $\mathfrak{R}(\hat{Z})$ are the unique minimizers of* (5.6) *and* (5.7)*. In any case, we have*

$$\min_{\theta \in [0,2\pi)} \|\hat{z} - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 \leq \frac{2\sqrt{2n}\|\underline{L} - L\|_2}{\tau}. \tag{5.8}$$

5.2.2 Dual Problems

To prove theorem 8, we introduce dual problems for (5.3) and (5.7). Specifically, we give the Lagrangian function $\mathcal{L} : \mathbb{C}^n \times \mathbb{R}^n \to \mathbb{R}$ of (5.3),

$$\mathcal{L}(z, \lambda) = z^* L z + \sum_{i=1}^{n} \lambda_i (1 - z_i^* z_i) = z^* (L - \mathrm{diag}(\lambda)) z + \sum_{i=1}^{n} \lambda_i,$$

and the dual function $q : \mathbb{R}^n \to \mathbb{R}$

$$q(\lambda) = \inf_{z \in \mathbb{C}^n} \mathcal{L}(z, \lambda),$$

which have the properties that for any $\lambda \in \mathbb{R}^n, z \in (\mathbb{S}^1)^n$, we have

$$q(\lambda) \leq \mathcal{L}(z, \lambda) = z^* L z.$$

In particular, $\sup_{\lambda \in \mathbb{R}^n} q(\lambda) \leq \min_{z \in (\mathbb{S}^1)^n} z^* L z$. Additionally, for any $\lambda \in \mathbb{R}^n$ such that $L - \mathrm{diag}(\lambda) \not\succeq 0$, we have $q(\lambda) = -\infty$. Indeed, if $v^*(L - \mathrm{diag}(\lambda))v < 0$, then we may take

$$q(\lambda) \leq \lim_{t \to \infty} (tv)^*(L - \mathrm{diag}(\lambda))(tv) = \lim_{t \to \infty} t^2 v^*(L - \mathrm{diag}(\lambda))v = -\infty.$$

Otherwise, in the case that $L - \mathrm{diag}(\lambda) \succeq 0$, the quadratic form $z^*(L - \mathrm{diag}(\lambda))z$ is minimized when $z = 0$, so

$$q(\lambda) = \mathcal{L}(0, \lambda) = \sum_{i=1}^{n} \lambda_i.$$

Together, this gives $q(\lambda)$ as

$$q(\lambda) = \begin{cases} \sum_{i=1}^{n} \lambda_i, & L - \mathrm{diag}(\lambda) \succeq 0 \\ -\infty, & L - \mathrm{diag}(\lambda) \not\succeq 0 \end{cases}$$

Writing $\Lambda = \operatorname{diag}(\lambda)$, it is clear that the supremum of $\sup q = \sup_{L-\Lambda \succeq 0} \operatorname{Tr}(\Lambda)$, which is an SDP. Recalling $\sup q \leq \min_{z \in (\mathbb{S}^1)^n} z^* L z$, we declare the dual problem of (5.3) to be

$$\begin{aligned} \max_{\Lambda \in \mathbb{R}^{n \times n}} \quad & \operatorname{Tr}(\Lambda) \\ \text{s.t.} \quad & L - \Lambda \succeq 0 \\ & \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \end{aligned} \quad , \tag{5.9}$$

where it is "dual" in the sense that, if Λ^* optimizes (5.9) and $\hat{z}$ optimizes (5.3), then $\operatorname{Tr}(\Lambda^*) \leq \hat{z}^* L \hat{z}$.

To find the dual of (5.6), we define $\mathcal{L}_{\text{SDP}} : \mathcal{H}_+^n \times \mathbb{R}^n \to \mathbb{R}$ and $d_{\text{SDP}} : \mathbb{R}^n \to \mathbb{R}$ by

$$\mathcal{L}_{\text{SDP}}(Z, \lambda) = \operatorname{Tr}(LZ) + \sum_{i=1}^{n} \lambda_i (1 - Z_{ii}) = \operatorname{Tr}((L - \operatorname{diag}(\lambda))Z) + \operatorname{Tr}(\operatorname{diag}(\lambda)), \text{ and}$$

$$d_{\text{SDP}}(\lambda) = \inf_{Z \in \mathcal{H}_+^n} \mathcal{L}_{\text{SDP}}(Z, \lambda) = \begin{cases} \operatorname{Tr}(\operatorname{diag}(\lambda)), & L - \operatorname{diag}(\lambda) \succeq 0 \\ -\infty, & \text{otherwise} \end{cases}$$

As before, we see that $\sup_{\lambda \in \mathbb{R}^n} d_{\text{SDP}}(\lambda) \leq \mathcal{L}_{\text{SDP}}(Z, \lambda) = \operatorname{Tr}(LZ)$ for any Z which is feasible to (5.6); therefore, if Λ^* and $\hat{Z}$ optimize (5.9) and (5.6), we will have $\operatorname{Tr}(\Lambda^*) \leq \operatorname{Tr}(L\hat{Z})$.

To prove the uniqueness of the solution to (5.6), we will need to quote a result from [3]. They use the primal-dual format with the primal problem

$$\begin{aligned} \min_{X \in \mathcal{S}^n} \quad & \operatorname{Tr}(CX) \\ \text{s.t.} \quad & \operatorname{Tr}(A_k X) = b_k, \quad k \in [m] \\ & X \succeq 0 \end{aligned} \tag{5.10}$$

where $C, A_k \in \mathcal{S}^n$ and $b \in \mathbb{R}^m$ are fixed. The dual problem is

$$\begin{aligned} \max_{y \in \mathbb{R}^m} \quad & b^T y \\ \text{s.t.} \quad & C - \textstyle\sum_{k=1}^m y_k A_k \succeq 0 \end{aligned} \tag{5.11}$$

where $b \in \mathbb{R}^m$ and A_k are as in the primal. Among other results, they prove the following:

Proposition 13 (Theorems 9 and 10 in [3]). *Suppose that $y \in \mathbb{R}^m$ is dual feasible and optimal, with* $\operatorname{rank}(C - \sum_{k=1}^m y_k A_k) = s$. *Let the columns of $Q \in \mathbb{R}^{n \times n-s}$ be an orthonormal basis for* $\operatorname{Nul}(C - \sum_{k=1}^m y_k A_k)$, *such that*

$$\operatorname{Col}(Q) = \operatorname{Nul}(C - \sum_{k=1}^m y_k A_k) \quad \textit{and} \quad Q^T Q = I.$$

then if

$$\operatorname{span}\{Q^T A_k Q\}_{k \in [m]} = \mathcal{S}^{n-s}, \tag{5.12}$$

there is a unqiue optimal primal solution X.

In order to use this result, we will need the dual of (5.7). Following Eqs. (5.10) and (5.11), this gives

$$\begin{aligned} \max_{\Lambda \in \mathcal{S}^{2n}} \quad & \operatorname{Tr}(\Lambda) \\ \text{s.t.} \quad & \mathfrak{R}(L) - \Lambda \succeq 0 \\ & \Lambda = \operatorname{diag}(\Lambda_1, \ldots, \Lambda_n) \\ & \Lambda_i \in \mathcal{S}^2 \end{aligned} \tag{5.13}$$

5.2.3 Proof of Theorem 8

With this in mind, we state a few lemmas whose proofs will constitute a proof of Theorem 8. Each of these assumes the notation and hypotheses of Theorem 8.

Lemma 15 (Sufficient conditions for strong duality). *If $\hat{z} \in (\mathbb{S}^1)^n$ satisfies*

$$L - \operatorname{diag} \operatorname{Re}(\hat{z}\hat{z}^* L) \succeq 0, \ \textit{and} \tag{5.14}$$

$$\operatorname{Nul}(L - \operatorname{diag} \operatorname{Re}(\hat{z}\hat{z}^* L)) = \operatorname{span}(\hat{z}), \tag{5.15}$$

then $\hat{Z} = \hat{z}\hat{z}^*$ is the unique optimizer of (5.6) *and $\mathfrak{R}(\hat{Z})$ is the unique optimizer of* (5.7).

Lemma 16. *If $\|L - \hat{L}\|_2 < \frac{\tau}{1+\sqrt{n}}$, then $\hat{z}$ meets the conditions of Lemma 15 and $\hat{Z} = \hat{z}\hat{z}^*$ is the unique optimizer of* (5.6) *and $\mathfrak{R}(\hat{Z})$ is the unique optimizer of* (5.7).

Lemma 17. *Suppose that $\hat{z}$ minimizes* (5.3)*. Then*

$$\min_{\theta\in[0,2\pi)} \|\hat{z} - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 \leq \frac{2\sqrt{2n}\|\underline{L} - L\|_2}{\tau}.$$

Before proving these, we begin with a further lemma that explains the recurrent $L - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^* L)$ term.

Lemma 18. *Suppose $A \succeq 0$. Then if $\hat{z}$ is an optimizer of*

$$\min_{z\in(\mathbb{S}^1)^n} \; z^* A z,$$

we have

$$\hat{z} \in \operatorname{Nul}(A - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^* A)).$$

Proof of Lemma 18. We define $f : (\mathbb{S}^1)^n \to \mathbb{R}$ by $f(z) = z^* A z$. We observe that $(\mathbb{S}^1)^n$ is an n-dimensional manifold with charts given by

$$\phi_U : U \to (\mathbb{S}^1)^n, \quad \phi_U(\theta_1, \ldots, \theta_n)_i = \mathrm{e}^{\mathrm{i}\theta_i},$$

where $U \subseteq \mathbb{R}^n$ is any open set satisfying $\|x - y\|_\infty < 2\pi$ for all $x, y \in U$ (or, more generally, $x - y \notin 2\pi(\mathbb{Z}^n) \setminus \{0\}$ for all $x, y \in U$). Assume $\hat{z}$ is a minimizer of (5.3). Because $z^* A z$ is smooth on $(\mathbb{S}^1)^n$, for any chart ϕ_U with $\hat{z} \in \phi_U(U)$, we must have $\nabla(f \circ \phi_U)|_{\phi_U^{-1}(\hat{z})} = 0$. In particular, if $\theta \in \mathbb{R}^n$ is such that $\hat{z}_i = \mathrm{e}^{\mathrm{i}\theta_i}$, then

$$\nabla(f \circ \phi_U)|_{\phi_U^{-1}(\hat{z})} = \nabla_\theta \hat{z}^* A \hat{z}.$$

We remark that

$$\hat{z}^* A \hat{z} = \sum_{ij} \mathrm{e}^{\mathrm{i}(\theta_j - \theta_i)} A_{ij} = \sum_i \mathrm{e}^{\mathrm{i}\theta_i} \hat{z}^* A_i$$

where $A_i = Ae_i$ and $A_{ij} = (A_i)_j$. This gives

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\hat{z}^*A\hat{z} &= \mathrm{i}\mathrm{e}^{\mathrm{i}\theta_j}\hat{z}^*A_j + \sum_{i=1}^{n}\mathrm{e}^{\mathrm{i}\theta_i}\frac{\partial}{\partial\theta_j}\hat{z}^*A_i \\ &= \mathrm{i}\mathrm{e}^{\mathrm{i}\theta_j}\hat{z}^*A_j - \mathrm{i}\mathrm{e}^{\mathrm{i}\theta_j}\sum_{i=1}^{n}\mathrm{e}^{-\mathrm{i}\theta_i}A_{ji} \\ &= \mathrm{i}\mathrm{e}^{\mathrm{i}\theta_j}\hat{z}^*A_j - \mathrm{i}\mathrm{e}^{-\mathrm{i}\theta_j}A_j^*\hat{z} \\ &= -2\,\mathrm{Im}(\mathrm{e}^{\mathrm{i}\theta_j}\hat{z}^*A_j) = -2\,\mathrm{Im}(\hat{z}\hat{z}^*A)_{jj}\end{aligned}$$

Therefore, $\nabla(f\circ\phi_U)|_{\phi_U^{-1}(\hat{z})} = 0$ if and only if $\operatorname{diag}\mathrm{Im}(\hat{z}\hat{z}^*A) = 0$. On the other hand, we observe that

$$\begin{aligned}((A - \operatorname{diag}\mathrm{Re}(\hat{z}\hat{z}^*A))\hat{z})_i &= A_i^*\hat{z} - \mathrm{Re}(\hat{z}_i\hat{z}^*A_i)\hat{z}_i \\ &= A_i^*\hat{z} - \mathrm{Re}(\overline{\hat{z}}_iA_i^*\hat{z})\hat{z}_i,\end{aligned}$$

such that $\hat{z}\in\mathrm{Nul}(A - \operatorname{diag}\mathrm{Re}(\hat{z}\hat{z}^*A))$ iff

$$\begin{aligned} & A_i^*\hat{z} &=& \mathrm{Re}(\overline{\hat{z}}_iA_i^*\hat{z})\hat{z}_i, & \text{for all } i \\ \iff \quad & \overline{\hat{z}}_iA_i^*\hat{z} &=& \mathrm{Re}(\overline{\hat{z}}_iA_i^*\hat{z}), & \text{for all } i \\ \iff \quad & \overline{\hat{z}}_iA_i^*\hat{z} &\in& \mathbb{R} & \text{for all } i \\ \iff \quad & \operatorname{diag}\mathrm{Im}(\hat{z}\hat{z}^*A) &=& 0 & \end{aligned}$$

This may be summarized as

$$z\in\mathrm{Nul}(A - \operatorname{diag}\mathrm{Re}(\hat{z}\hat{z}^*A)) \iff \overline{\hat{z}}_iA_i^*\hat{z}\in\mathbb{R} \text{ for all } i \tag{5.16}$$

so that $\hat{z}$ is a minimizer of (5.3) only if $\hat{z}\in\mathrm{Nul}(A - \operatorname{diag}\mathrm{Re}(\hat{z}\hat{z}^*A))$. □

Beyond being a useful result, this suggests that the matrix $L - \operatorname{diag}\mathrm{Re}(\hat{z}\hat{z}^*L)$ can play a role in certifying minima of these optimization problems, as we shall see in the proofs of Lemmas 15–17. We remark, however, that this condition is far from being *sufficient* to show that $\hat{z}$ is an optimizer of $\min_{z\in(\mathbb{S}^1)^n} z^*Az$. Indeed, even the stronger condition $\mathrm{Nul}(A -$

$\operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*A)) = \operatorname{span}(\hat{z})$ is insufficient. For example, if we take $A \in \mathcal{S}^2$ to be $A = \mathbb{1}\mathbb{1}^*$, then with $z = \mathbb{1}$ we have

$$A - \operatorname{diag}\operatorname{Re}(\mathbb{1}\mathbb{1}^*A) = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix},$$

so that $\operatorname{Nul}(A - \operatorname{diag}\operatorname{Re}(\mathbb{1}\mathbb{1}^*A)) = \operatorname{span}(\mathbb{1})$, but

$$\begin{bmatrix} 1 \\ \mathrm{e}^{\mathrm{i}\theta} \end{bmatrix}^* A \begin{bmatrix} 1 \\ \mathrm{e}^{\mathrm{i}\theta} \end{bmatrix} = 2 + 2\cos\theta,$$

which is *maximized* at $z = \mathbb{1}$ and minimized at $\hat{z} = (1, -1)^T$. We now prove Lemmas 15–17.

Proof of Lemma 15. We begin by showing that

$$\hat{z} \in \operatorname{Nul}(L - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L)) \text{ and } L - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L) \succeq 0$$

suffices to show that $\hat{Z} = \hat{z}\hat{z}^*$ and $\mathfrak{R}(\hat{Z})$ are optimizers of (5.6) and (5.7), respectively. We then show that, if $\operatorname{Nul}(L - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L)) = \operatorname{span}(\hat{z})$ holds in addition to these, then they are the *unique* optimizers of their problems.

Suppose that $\hat{z} \in (\mathbb{S}^1)^n$ satisfies $L - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L) \succeq 0$ and $(L - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L))\hat{z} = 0$, and set $\hat{Z} = \hat{z}\hat{z}^*$. This means $\operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L)$ is feasible for (5.9), so we set $\hat{\Lambda} = \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L)$ to obtain

$$\begin{array}{rclcll} \operatorname{Tr}(\hat{\Lambda}) & = & \operatorname{Tr}(\operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L)) & = & \operatorname{Tr}(\operatorname{Re}(\hat{z}\hat{z}^*L)) & \\ & = & \sum_{i=1}^n \operatorname{Re}(\hat{z}_i\hat{z}^*L_i) & = & \sum_{i=1}^n \hat{z}_i\hat{z}^*L_i & \text{(from (5.16))} \\ & = & \hat{z}^*L\hat{z} & & & \end{array}$$

Therefore $(\hat{z}, \hat{\Lambda})$ is a primal-dual pair for (5.3) and (5.9) with equal objective function values, showing that they are both optimizers for their respective problems. Since (5.9) is also the dual problem for (5.6), and since $\operatorname{Tr}(L\hat{Z}) = \operatorname{Tr}(L\hat{z}\hat{z}^*) = \hat{z}^*L\hat{z}$, $\hat{\Lambda}$ also certifies optimality of $\hat{Z}$ for *its* problem, and similarly also for $\mathfrak{R}(\hat{Z})$ by observing that $\operatorname{Tr}(\mathfrak{R}(A)) = 2\operatorname{Tr}(A)$ for

general $A \in \mathcal{H}^n$.

To show uniqueness, we will show that, under the additional assumption that $\mathrm{Nul}(L - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^* L)) = \operatorname{span}(\hat{z})$, then $\mathfrak{R}(\hat{\Lambda})$ satisfies the hypotheses of Proposition 13. To this end, observe that $\mathrm{Nul}(L - \hat{\Lambda}) = \operatorname{span}(\hat{z})$ implies

$$\mathrm{Nul}(\mathfrak{R}(L) - \mathfrak{R}(\hat{\Lambda})) = \operatorname{Col}\mathfrak{R}(\hat{z}).$$

The two columns of $\mathfrak{R}(\hat{z})$ are trivially orthogonal, so we set $Z_i = \mathfrak{R}(\hat{z}_i)$ and

$$Q = \frac{1}{n}\mathfrak{R}(\hat{z}) = \frac{1}{n}\begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix}.$$

Towards proving that Q satisfies (5.12), we note that $Z_i^T Z_i = Z_i Z_i^T = I_2$ and claim that, for $A \in \mathcal{S}^2$, one block-diagonal pre-image of A is $n^2 \operatorname{diag}(Z_1 A Z_1^T, 0, \ldots, 0)$. Indeed,

$$Q^T n^2 \operatorname{diag}(Z_1 A Z_1^T, 0, \ldots, 0) Q = Z_1^T Z_1 A Z_1^T Z_1 = A.$$

This establishes that $\operatorname{span}\{Q^T \operatorname{diag}(\Lambda_i) Q\}_{\Lambda_i \in \mathcal{S}^2} = \mathcal{S}^2$, which, by Proposition 13, gives us that (5.7) has a unique solution. Since we have already established optimality of $\mathfrak{R}(\hat{Z})$, this unique solution is $\mathfrak{R}(\hat{Z})$.

Label the feasible sets of (5.6) and (5.7) F_1 and F_2, respectively. Then $\mathfrak{R}(F_1) \subseteq F_2$, such that if $\mathfrak{R}(\hat{Z})$ is the unique minimizer of $\operatorname{Tr}(\mathfrak{R}(L)Z)$ over F_2, then $\hat{Z}$ is the unique minimizer of $\operatorname{Tr}(LZ)$ over F_1. This completes the proof. □

Proof of Lemma 16. Accepting the notation of Theorem 8, suppose that $\hat{z}$ is an optimizer of (5.3). For convenience, we set $Y = L - \operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^* L)$. Then, by Lemma 18 we have $Y\hat{z} = 0$. Therefore, by Lemma 15, to show that the solution to (5.6) is unique, it suffices to have $P^* Y P \succ 0$, where the columns of $P \in \mathbb{C}^{n \times n-1}$ form an orthonormal basis for $\hat{z}^\perp$.

To this end, we introduce $\hat{L} = D - W \circ \hat{z}\hat{z}^*$, which has that $\mathrm{Nul}(\hat{L}) = \operatorname{span}(\hat{z})$ and

$P^*\hat{L}P \succ 0$ when G is connected (see, e.g., lemma 1.7 of [26]), so that

$$\hat{L} - \operatorname{diag}\operatorname{Re}\hat{z}\hat{z}^*\hat{L} = \hat{L}.$$

By Weyl's inequalities (Theorem 4.3.1 in [55]), the smallest eigenvalue $\lambda_1(P^*YP)$ of P^*YP satisfies

$$\lambda_1(P^*YP) \geq \lambda_1(P^*\hat{L}P) - \|P^*(Y-\hat{L})P\|_2 = \lambda_2(\hat{L}) - \|Y-\hat{L}\|_2,$$

so it suffices to have $\lambda_2(\hat{L}) = \tau > \|Y-\hat{L}\|_2$. The lemma follows by observing that

$$\begin{aligned}
\|Y-\hat{L}\|_2 &\leq \|L-\hat{L}\|_2 + \|\operatorname{diag}\operatorname{Re}(\hat{z}\hat{z}^*L)\|_2 \\
&\leq \|L-\hat{L}\|_2 + \|L\hat{z}\|_\infty \\
&\leq \|L-\hat{L}\|_2 + \|L\hat{z}\|_2 \\
&= \|L-\hat{L}\|_2 + \|(L-\hat{L})\hat{z}\|_2 \\
&\leq \|L-\hat{L}\|_2 + \sqrt{n}\|L-\hat{L}\|_2
\end{aligned}$$

□

Proof of Lemma 17. We begin by assuming that $\underline{x}^*\hat{z} = \hat{z}^*\underline{x} = |\underline{x}^*\hat{z}|$, which is accomplished by taking appropriate representatives of $\underline{x}$ and $\hat{z}$ in $(\mathbb{S}^1)^n/\mathbb{S}^1$. This gives that

$$\min_{\theta\in[0,2\pi)} \|\underline{x} - \mathrm{e}^{\mathrm{i}\theta}\hat{z}\|_2 = \|\underline{x}-\hat{z}\|_2.$$

Writing $\Delta L = L - \underline{L}$, we have, by optimality of $\hat{z}$, that $\hat{z}^*L\hat{z} \leq \underline{x}^*L\underline{x}$. Since $\underline{x} \in \operatorname{Nul}(\underline{L})$, this gives

$$\hat{z}^*\underline{L}\hat{z} + \hat{z}^*\Delta L\hat{z} = \hat{z}^*L\hat{z} \leq \underline{x}^*L\underline{x} = \underline{x}^*\underline{L}\underline{x} + \underline{x}^*\Delta L\underline{x} = \underline{x}^*\Delta L\underline{x},$$

which yields

$$\begin{aligned}\hat{z}^*\underline{L}\hat{z} &\leq \underline{x}^*\Delta L\underline{x} - \hat{z}^*\Delta L\hat{z} \\ &= (\underline{x}-\hat{z})^*\Delta L(\underline{x}+\hat{z}) \\ &\leq \|\underline{x}-\hat{z}\|_2\|\Delta L\|_2\sqrt{2n}\end{aligned} \tag{5.17}$$

We then lower-bound the left-hand side of (5.17) by setting

$$y = \mathrm{Proj}_{\underline{x}^\perp}\hat{z} = \hat{z} - \frac{1}{n}\underline{x}\underline{x}^*\hat{z},$$

so that

$$\hat{z}^*\underline{L}\hat{z} = y^*\underline{L}y.$$

At this point, we remark that $\|\underline{x}\underline{x}^* - \hat{z}\hat{z}^*\|_F^2 = 2n^2 - 2|\underline{x}^*\hat{z}|^2$ and $\|\underline{x}-\hat{z}\|_2^2 = 2n - 2|\underline{x}^*\hat{z}|$, such that

$$\begin{aligned}n\|\underline{x}-\hat{z}\|_2^2 \leq \|\underline{x}\underline{x}^* - \hat{z}\hat{z}^*\|_F^2 &= 2(n-|\underline{x}^*\hat{z}|)(n+|\underline{x}^*\hat{z}|) \\ &= \|\underline{x}-\hat{z}\|_2^2(n+|\underline{x}^*\hat{z}|) \leq 2n\|\underline{x}-\hat{z}\|_2^2.\end{aligned}$$

Therefore, we may observe that

$$\|y\|_2^2 = \|\hat{z}\|_2^2 - \frac{1}{n^2}|\underline{x}^*\hat{z}|^2\|\underline{x}\|_2^2 = n - \frac{1}{n}|\underline{x}^*\hat{z}|^2 = \frac{\|\underline{x}\underline{x}^* - \hat{z}\hat{z}^2\|_F^2}{2n},$$

giving $\frac{1}{2}\|\underline{x}-\hat{z}\|_2^2 \leq \|y\|_2^2 \leq \|\underline{x}-\hat{z}\|_2^2$. In this way, since y is orthogonal to the null space of $\underline{L}$, we have

$$\hat{z}^*\underline{L}\hat{z} = y^*\underline{L}y \geq \lambda_2(\underline{L})\|y\|_2^2 \geq \frac{\lambda_2(\underline{L})}{2}\|\underline{x}-\hat{z}\|_2^2. \tag{5.18}$$

Combining this with (5.17) completes the proof. □

5.2.4 Spanning Tree Strategies

By way of practical interest, in this section we briefly consider a strategy for solving the angular synchronization that is cheap to compute and theoretically tight. Namely, given

an angular synchronization problem (5.3) with data $W \circ X \in \mathcal{H}^n, W = |W \circ X|, X = \operatorname{sgn} W \circ X$ such that W is the weighted adjacency matrix of a connected graph G, we will solve (5.3) by taking $G' \subseteq G$ to be a spanning tree of G and deciding the phases of $\hat{z}_i$ by arbitrarily fixing the phase of one vertex and directly propagating the relative phases along the edges. This works by leveraging the property that, for any two vertices v, w in a tree $G' = (V', E')$, there exists a unique path (see, e.g., Theorem 1.5.1 of [30]) $p(v, w) = vv_1 \cdots v_{k-1}w$ from v to w. Therefore, to build an appropriate vector of phases z, we may simply fix a root vertex r, and set $z_r = 1$. For each vertex v, if $e_1, \ldots, e_k$ are the edges in the path $p(r, v)$ from r to v, we set $z_v = \prod_{i=1}^{k} X_{e_i}$.[2] Algorithm 4 makes this procedure precise, and Fig. 5.1 illustrates a simple example.

This method is quite popular in the literature, precisely because of its quick implentation and amenability to simple analysis. In fact, in a paper that preceded the joint work presented in Chapter 3, this was the strategy analyzed for phase retrieval with local measurements [57]; Proposition 14 somewhat generalizes the bound achieved there. Of more broad interest to the community is the canonical generalization of Algorithm 4 to arbitrary group synchronization over a graph, and several authors working on various group synchronization problems – mostly pose and position estimation – have used and studied spanning tree-based techniques, including [18, 24, 36, 47]. Notably, because the solution of group synchronization over a spanning tree is so cheap, it admits stochastic consensus methods such as that proposed in [41], and this is the strategy followed in [47].

Before stating Algorithm 4, we say that, given a tree $G' = (V' = [n], E')$, $(k_1, \ldots, k_n)$ is a *rooted ordering* of V' if $(k_i)_{i=1}^n$ is a permutation of $[n]$ such that, for each $i \geq 2$, k_i has exactly one neighbor in $\{k_1, \ldots, k_{i-1}\}$. Having fixed a rooted ordering, we call k_1 the *root*, and the preceding neighbor (the unique k_j such that $j < i$ and $(k_i, k_j) \in E'$) of k_i is called the *parent* of k_i, and neighbors that suceed it are called its *children* or *descendants*. Indeed, by a breadth-first search [86], we may find, from an adjacency list,[3] a rooted ordering with

[2]Here, note that we have used the ordered pair $e_i = (v_i, v_i')$, representing the edge between $v_i, v_i' \in [n]$, as a subscript denoting for X; specifically, X_{e_i} in this instance means $X_{v_i v_i'}$.

[3]This is equivalent to an adjacency matrix stored sparsely. Using the Yale sparse format, which is still standard for everyday sparse computation [10], vertex n is adjacent to $JA(k)$ for $k \in [IA(n), IA(n+1))$. IA and JA are defined in [10] and stored such that this block is continuous in memory.

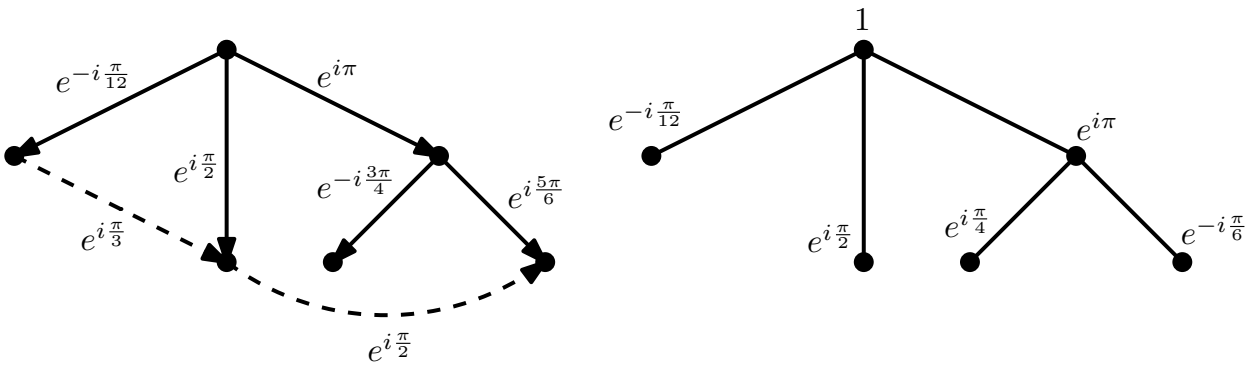


Figure 5.1: Example of angular synchronization on a spanning tree

any vertex as the root with time complexity $\mathcal{O}(|V| + |E|)$. With this, we state Algorithm 4, which also has a runtime of $\mathcal{O}(|V| + |E|)$.

Algorithm 4 Angular Synchronization on a Spanning Tree

Input: A spanning tree $G' = (V = [n], E') \subseteq G = (V, E)$ of a connected graph. Edge phase data $X_{ij} \in \mathbb{S}^1$ for $(i, j) \in E$ with $X \in \mathcal{H}^n$.

Output: A vector of phases $\hat{z} \in (\mathbb{S}^1)^n$.

1: Choose $(k_1, \ldots, k_n)$ to be any rooted ordering of V.
2: Set $\hat{z}_1 = 1$.
3: For $i = 2, \ldots, n$, set $\hat{z}_{k_i} = \hat{z}_{k_j} X_{k_i k_j}$ where k_j is the parent of k_i.

We remark that Algorithm 4 makes no mention of a weight matrix W. This is because the output of Algorithm 4 uniquely (up to scaling by $\mathrm{e}^{\mathrm{i}\theta}$) satisfies $\hat{z}^* L_{G'} \hat{z} = 0$, regardless of W, such that the solution solves angular synchronization exactly over the tree. Additionally, since $L_{G'} \succeq 0$, $\hat{z}$ is an optimum for (5.3) and therefore coincides with the results produced by factorizing the solution to (5.6) or by taking $z = \mathrm{sgn}(u)$ where u is the eigenvector for $L_{G'}$'s smallest eigenvalue. We prove this in Proposition 14.

Proposition 14. *Let $G' = ([n], E') \subseteq G = ([n], E)$ be a spanning tree for some connected graph G. Suppose further that $X, \underline{X} \in \mathcal{H}^n$ satisfy $X_{ij}, \underline{X}_{ij} \in \mathbb{S}^1$ if $(i, j) \in E'$ and $X_{ij}, \underline{X}_{ij} = 0$ otherwise. Let $W \in \mathbb{R}_+^{n \times n} \cap \mathcal{S}^n$ have $\mathrm{supp}(W) = \mathrm{supp}(X)$. Then setting $L_{G'} = D - W \circ X$ and $\underline{L}_{G'} = D - W \circ \underline{X}$, where $D = \mathrm{diag}(W \mathbb{1}_n)$, the output $\hat{z}$ of Algorithm 4 for X satisfies $\mathrm{span}(\hat{z}) = \mathrm{Nul}(L_{G'})$. Furthermore, if $\underline{z}$ is the output for $\underline{X}$, we have*

$$\min_{\theta \in [0, 2\pi]} \|\hat{z} - \mathrm{e}^{\mathrm{i}\theta} \underline{z}\|_2 \leq \sqrt{\frac{2}{\tau_{G'}}} \|X - \underline{X}\|_F, \tag{5.19}$$

where $\tau_{G'} = \lambda_2(\mathrm{diag}(|X|\mathbb{1}) - X)$ *is the unweighted spectral gap of* G'.

Proof of Proposition 14. Recalling (5.1), we have

$$\hat{z}^* L_{G'} \hat{z} = \sum_{(i,j)\in E'} w_{ij} |\hat{z}_i - X_{ij}\hat{z}_j|^2,$$

which immediately equals zero, since if $(i,j) \in E'$, we have set $\hat{z}_j^* \hat{z}_i = X_{ij}$ in line 3 of Algorithm 4. Since G' is connected, $\dim \mathrm{Nul}(L_{G'}) = 1$, so $\mathrm{Nul}(L_{G'}) = \mathrm{span}(\hat{z})$.

To get the error bound, we remark that Algorithm 4 does not depend on W, so, having found $\hat{z}$ from X alone, $\mathrm{span}(\hat{z}) = \mathrm{Nul}(L_{G'})$ for any weight matrix $W = |W|, \mathrm{supp}(W) = \mathrm{supp}(X)$. In particular, taking $W = |X|$, so that $L_{G'} = \mathrm{diag}(|X|\mathbb{1}) - X$, we have

$$\hat{z}^* \underline{L}_{G'} \hat{z} = \sum_{(i,j)\in E'} |\hat{z}_i - \underline{X}_{ij}\hat{z}_j|^2 = \sum_{(i,j)\in E'} |X_{ij} - \underline{X}_{ij}|^2 = \|X - \underline{X}\|_F^2.$$

The proof is completed by finding, as in (5.18), that $\hat{z}^* \underline{L}_{G'} \hat{z} \geq \min_\theta \frac{\tau_{G'}}{2} \|\hat{z} - \mathrm{e}^{\mathrm{i}\theta}\underline{z}\|_2^2$. □

This result demonstrates the real appeal of this method: the solution is exact, and marvelously cheap to compute. When G is a tree, we can get an exact solution to the SDP of (5.6) in $\mathcal{O}(n)$! The drawback, however, is that trees suffer from dismal spectral gaps, and this weakens the bound proposed in Proposition 14. From Theorem 4.1 in [28] and Theorem 4.2 in [74] we have that, for a tree G' on n vertices,

$$\frac{4}{n\,\mathrm{diam}(G')} \leq \tau_{G'} \leq 2\left(1 - \cos\left(\frac{\pi}{\mathrm{diam}(G')+1}\right)\right) \leq \frac{\pi^2}{(\mathrm{diam}(G')+1)^2}, \tag{5.20}$$

where $\mathrm{diam}(G')$ is the diameter of G'. To make $\tau_{G'}$ large, then, we will want $\mathrm{diam}(G')$ as small as possible, but in the case of the graph $G_{d,\delta}$ associated with[4] $T_\delta(\mathbb{C}^{d\times d})$, the path from 1 to $\lceil d/2 \rceil$ will have length at least

$$\mathrm{diam}(G') \geq \left\lceil \frac{\lceil d/2 \rceil - 1}{\delta - 1} \right\rceil \geq \frac{d}{3\delta},$$

[4]Specifically, $G_{d,\delta} = ([d], E)$ where $(i,j) \in E$ if $i \neq j$ and $|i-j| \bmod d < \delta$.

as long as $d \geq 6$. However, a spanning tree $G' \subseteq G_{d,\delta}$ with diameter of this asymptotic order is always achievable by setting $G' = ([d], E')$ by setting

$$p = p(d, \delta - \delta + 1) = \delta, 2\delta - 1, \ldots, (d - \delta + 1)$$

to be the increasing path of length $\ell = \lceil \frac{d-2\delta+1}{\delta-1} \rceil$ from δ to $d - \delta + 1$ and connecting each vertex $i \in [d] \setminus p$ to any choice $j \in p$ such that $|i - j| \bmod d < \delta$. Then

$$\operatorname{dist}(i, j) \leq 2 + \ell \leq \frac{d-1}{\delta - 1} + 1 \leq \frac{3d}{\delta},$$

so $\operatorname{diam}(G') \leq \frac{3d}{\delta}$. These considerations give us that we may always construct a spanning tree of $G_{d,\delta}$ with

$$\frac{4\delta}{3d^2} \leq \tau_{G'} \leq \frac{100\delta^2}{d^2}.$$

Of course, the lower bound is the only one that may be used in the result of Proposition 14, and in any event, we know from Corollary 3.2 of [38] that $\tau_{G'} \leq \tau_G$ for any spanning tree $G' \subseteq G$. Nonetheless, we formalize this lower bound in Corollary 6.

Corollary 6. *If $G = G_{d,\delta}$, then there exists a spanning tree $G' \subseteq G$ with $\tau_{G'} \geq \frac{4\delta}{3d^2}$. Using the notation of Proposition 14, the output of Algorithm 4 on G' achieves*

$$\min_{\theta \in [0, 2\pi]} \|\hat{z} - \mathrm{e}^{\mathrm{i}\theta} \underline{z}\|_2 \leq \sqrt{3/2} \left(\frac{d}{\delta^{1/2}} \right) \|X - \underline{X}\|_F.$$

5.3 Refined Error Guarantees

5.3.1 Main Result

In this section, we consider how the theory developed in Section 5.2 can improve the results of Sections 3.4 and 3.5, specifically seeking improvements over the main error bounds proven in Theorems 4 and 5 and Corollaries 2 and 3. Since the main contribution of this dissertation is to provide theoretical upper bounds on the reconstruction error of the phase

retrieval algorithm proposed in Chapter 3 and its derivatives, these results are among the most significant of this chapter. In particular, we will find in Corollary 8 that the main recovery results of Section 3.5 may be reduced by as much as pulling a factor of d/δ out of the "phase error" terms by using exact angular synchronization through SDP. Considering that we tend to assume $\delta \ll d$, this constitutes a significant sharpening of these bounds.

We begin by noticing a few different strategies we could take to improve Theorem 4. First of all, if the noise level is low enough, by Theorem 8 we can use SDP to obtain the *actual* minimizer of $\min_{y\in\mathbb{C}^d} \eta_{\widetilde{X}}(\mathrm{sgn}(y))$. This will spare us the inequalities of (3.28), which ultimately cost us a factor of τ in the denominator of the error bound. Doing so would immediately improve the guarantee of Corollary 2. In more specific terms, we prove the following:

Theorem 9 (See Theorem 4)**.** *Suppose that $G = (V = [d], E)$ is an undirected, connected, and unweighted graph (so that $W_{ij} = \chi_{E(i,j)}$) with $\tau_G = \lambda_2(D - W)$. Let $\underline{x} \in (\mathbb{S}^1)^d$ be the ground truth vector, and set $\underline{\widetilde{X}} = W \circ \underline{x}\underline{x}^*$. Let $L = D - W \circ \widetilde{X}$ with $\widetilde{X} \in \mathcal{H}^d, |\widetilde{X}_{ij}| = 1$ for all $(i,j) \in E$, and suppose $x \in (\mathbb{S}^1)^d$ is an optimizer of (5.3). If $\|\widetilde{X} - \underline{\widetilde{X}}\|_2 < \frac{\tau_G}{1+\sqrt{d}}$, then xx^* is the unique solution to (5.6). In any case, x satisfies:*

$$\min_{\theta\in\mathbb{R}} \|x - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 \leq \frac{2\sqrt{2}\|\widetilde{X} - \underline{\widetilde{X}}\|_F}{\sqrt{\tau_G}} \tag{5.21}$$

5.3.2 τ_G vs. $\tau_N \min_{i\in V} \deg(i)$

Before proving Theorem 9, we remark that the spectral gap used here is different from the τ used in Theorem 4. There, we used $\tau_N = \lambda_2(I - D^{-1/2}WD^{-1/2})$; if we were to merely "delete" a $\sqrt{\tau_N}$ factor from the bound given in Theorem 4, we would get something slightly different from (5.21); namely, we would arrive at

$$\min_{\theta\in\mathbb{R}} \|x - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 \leq \frac{C\|\widetilde{X} - \underline{\widetilde{X}}\|_F}{\sqrt{\tau_N \min_{i\in V}(\deg(i))}}. \tag{5.22}$$

We notice that this differs from (5.21) only on which "spectral gap" appears in the denominator: $\tau_N \min_{i \in V} \deg(i)$ or τ_G. By inspection, when G is k-regular, $D = kI$ and these two coincide, since $D - W = kI(I - D^{-1/2} W D^{-1/2})$, giving $\tau_G = k\tau_N = \min_{i \in V} \deg(i) \tau_N$. However, when the vertices of G do *not* have constant degree, it is possible that $\tau_N \min_{i \in V} \deg(i) < \tau_G$; this makes sense, since the $\min_{i \in V} \deg(i)$ factor comes from lower-bounding the left-hand side in Lemma 4 quite wastefully by

$$\sum_{i \in V} \deg(i) |g_i - \mathrm{e}^{\mathrm{i}\theta}|^2 \geq \min_{i \in V} \deg(i) \sum_{i \in V} |g_i - \mathrm{e}^{\mathrm{i}\theta}|^2.$$

By contrast, using τ_G "evenly mixes" the potentially varying degrees into the eigenvector problem. To make this precise, Proposition 15 establishes that (5.21) is never worse than (5.22), and we follow it with an example to show a case where the improvement is strict.

Proposition 15. *Given a weighted graph $G = (V = [d], E)$ with weight matrix $W \in \mathcal{S}^d$ satisfying $W_{ij} \geq 0$ and $W_{ij} = 0$ iff $(i, j) \notin E$ and degree matrix D, set*

$$\tau_G = \lambda_2(D - W) \text{ and } \tau_N = \lambda_2(I - D^{-1/2} W D^{-1/2}).$$

Then $\tau_G \geq \tau_N \min_{i \in V} \deg(i)$.

Proof of Proposition 15. We rely multiple times on the identity, for $A \in \mathcal{H}^d$,

$$\inf_{v \perp \mathrm{Nul}(A)} \frac{v^* A v}{v^* v} = \inf_v \sup_{w \in \mathrm{Nul}(A)} \frac{v^* A v}{(v - w)^* (v - w)},$$

which holds since $v^* A v = (v - w)^* A (v - w)$ and the denominator is minimized by taking $w = \mathrm{Proj}_{\mathrm{Nul}(A)} v$. When $\mathrm{Nul}(A) = \mathrm{span}(w)$, this reduces to

$$\inf_{v \perp w} \frac{v^* A v}{v^* v} = \inf_v \sup_t \frac{v^* A v}{(v - tw)^* (v - tw)}.$$

With this in mind, we set $L = D - W$ and $\mathcal{L} = D^{-1/2} L D^{-1/2}$ and use a change of variables

$w = D^{-1/2}v$ to obtain

$$\begin{aligned}
\tau_N &= \inf_{v \perp D^{1/2}\mathbb{1}} \frac{v^*\mathcal{L}v}{v^*v} \\
&= \inf_v \sup_t \frac{v^*\mathcal{L}v}{(v - tD^{1/2}\mathbb{1})(v - tD^{1/2})} \\
&= \inf_w \sup_t \frac{(D^{1/2}w)^*\mathcal{L}D^{1/2}w}{(D^{1/2}w - tD^{1/2}\mathbb{1})^*(D^{1/2}w - tD^{1/2}\mathbb{1})} \\
&= \inf_w \sup_t \frac{w^*Lw}{(w - t\mathbb{1})^*D(w - t\mathbb{1})} \\
&\leq \frac{1}{\min_{i \in V} \deg(i)} \inf_w \sup_t \frac{w^*Lw}{(w - t\mathbb{1})^*(w - t\mathbb{1})} \\
&= \frac{1}{\min_{i \in V} \deg(i)} \inf_{w \perp \mathbb{1}} \frac{w^*Lw}{w^*w} \\
&= \frac{\tau_G}{\min_{i \in V} \deg(i)}
\end{aligned}$$

□

We now present a sequence of unweighted graphs G_n such that $\tau_G(G_n) > \tau_N(G_n)$. G_n will be the complete graph on n vertices, K_n, with an extra vertex having exactly one edge. Namely, we may set $G_n = (V_n, E_n)$ with

$$\begin{gathered}
V_n = [n+1] \text{ and} \\
E_n = \{(x, y) \in [n]^2 : x \neq y\} \cup \{(n, n+1), (n+1, n)\}.
\end{gathered} \tag{5.23}$$

Then, considering that $D = \operatorname{diag}((n-1)\mathbb{1}_{n-1}^*, n, 1)$, the graph Laplacians $L, \mathcal{L} \in \mathbb{R}^{(n+1)\times(n+1)}$

of G_n become

$$L = D - W = \begin{bmatrix} nI_{n-1} - \mathbb{1}_{n-1}\mathbb{1}_{n-1}^* & -\mathbb{1}_{n-1} & 0_{n-1} \\ -\mathbb{1}_{n-1}^* & n & -1 \\ 0_{n-1}^* & -1 & 1 \end{bmatrix} \quad \text{and}$$

$$\mathcal{L} = I - D^{-1/2}WD^{-1/2} = \begin{bmatrix} \frac{n}{n-1}I_{n-1} - \frac{1}{n-1}\mathbb{1}_{n-1}\mathbb{1}_{n-1}^* & -\frac{1}{\sqrt{n(n-1)}}\mathbb{1}_{n-1} & 0_{n-1} \\ -\frac{1}{\sqrt{n(n-1)}}\mathbb{1}_{n-1}^* & 1 & -\frac{1}{\sqrt{n}} \\ 0_{n-1}^* & -\frac{1}{\sqrt{n}} & 1 \end{bmatrix}$$

We finish the example by proving Proposition 16.

Proposition 16. *With $n > 3$ and $G_n = (V_n, E_n)$ defined as in (5.23), $\tau_N = \lambda_2(\mathcal{L}) < 1$ and $\tau_G = \lambda_2(L) = 1$. In particular, since $\min_{i \in V} \deg(i) = \deg(n+1) = 1$, we have $\tau_G > \tau_N \min_{i \in V} \deg(i)$.*

Proof. Take $c_1, \ldots, c_{n-1} \in \mathbb{R}$ such that $\sum_{i=1}^{n-1} c_i = 0$ and set $c = \sum_{i=1}^{n-1} c_i e_i^{n-1} \in \mathbb{R}^{n-1}$. Then $c \perp \mathbb{1}_{n-1}$ and

$$L \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} (nI_{n-1} - \mathbb{1}_{n-1}\mathbb{1}_{n-1}^*)c \\ -\mathbb{1}_{n-1}^* c \\ 0 \end{bmatrix} = \begin{bmatrix} nc \\ 0 \\ 0 \end{bmatrix} = n \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix}, \text{ while}$$

$$\mathcal{L} \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} (\frac{n}{n-1}I_{n-1} - \frac{1}{n-1}\mathbb{1}_{n-1}\mathbb{1}_{n-1}^*)c \\ -\frac{1}{\sqrt{n(n-1)}}\mathbb{1}_{n-1}^* c \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{n}{n-1}c \\ 0 \\ 0 \end{bmatrix} = \frac{n}{n-1} \begin{bmatrix} c \\ 0 \\ 0 \end{bmatrix},$$

which identifies $\{\mathbb{1}_{[n-1]}^{n+1}, e_n, e_{n+1}\}^\perp$ as an $n-2$-dimensional eigenspace of both L and $\mathcal{L}$, with eigenvalues n and $\frac{n}{n-1}$, respectively (recall $\mathbb{1}_{[n-1]}^{n+1} = (1, \ldots, 1, 0, 0)^T$). Therefore, the other three eigenvectors (including the nullspace vectors $\mathbb{1}_{n+1}$ and $D^{1/2}\mathbb{1}_{n+1}$ of L and $\mathcal{L}$) are in $W = \operatorname{span}\{\mathbb{1}_{[n-1]}^{n+1}, e_n, e_{n+1}\}$.

We state the remaining eigenvalues of L directly by giving their eigenvectors. Clearly

$L\mathbb{1}_{n+1} = 0_{n+1}$. We observe additionally that

$$L\left(\begin{bmatrix}\mathbb{1}_{n-1}\\0\\0\end{bmatrix}+\begin{bmatrix}0_{n-1}\\0\\-(n-1)\end{bmatrix}\right)=\begin{bmatrix}\mathbb{1}_{n-1}\\-(n-1)\\0\end{bmatrix}+\begin{bmatrix}0_{n-1}\\n-1\\1-n\end{bmatrix}=\begin{bmatrix}\mathbb{1}_{n-1}\\0\\-(n-1)\end{bmatrix},$$

such that $(1,\ldots,1,0,1-n)^T$ has an eigenvalue of 1, and

$$L\begin{bmatrix}\mathbb{1}_{n-1}\\-n\\1\end{bmatrix}=L(\mathbb{1}_{n+1}-(n+1)e_n)=-(n+1)Le_n=(n+1)\begin{bmatrix}\mathbb{1}_{n-1}\\-n\\1\end{bmatrix},$$

such that $(1,\ldots,1,-n,1)^T$ has an eigenvalue of $n+1$. Therefore, the spectrum of L is n with a multiplicity of $n-2$, and $0, 1$, and $n+1$, each with multiplicity 1, so the spectral gap is $\tau_G = 1$ as stated.

To get the last three eigenvalues of $\mathcal{L}$, we take

$$M=\begin{bmatrix}\frac{1}{\sqrt{n-1}}\mathbb{1}_{[n-1]}^{n+1} & e_n & e_{n+1}\end{bmatrix}$$

as an orthogonal basis of W, such that the remaining eigenvalues of $\mathcal{L}$ are also the eigenvalues of $M^*\mathcal{L}M$, which we calculate to be

$$M^*\mathcal{L}M=\begin{bmatrix}\frac{1}{n-1} & -\frac{1}{\sqrt{n}} & 0\\-\frac{1}{\sqrt{n}} & 1 & -\frac{1}{\sqrt{n}}\\0 & -\frac{1}{\sqrt{n}} & 1\end{bmatrix}.$$

We calculate the characteristic polynomial of $M^*\mathcal{L}M$ directly:

$$\det(M^*\mathcal{L}M-\lambda I)=-\lambda\left(\lambda^2-\left(\frac{2n-1}{n-1}\right)\lambda+\frac{n^2-n+2}{n(n-1)}\right).$$

The two remaining nonzero eigenvalues may be obtained by the quadratic formula:

$$\lambda = \frac{\frac{2n-1}{n-1} \pm \sqrt{\left(\frac{2n-1}{n-1}\right)^2 - 4\frac{n^2-n+2}{n(n-1)}}}{2}.$$

τ_N is obtained by taking the negative square root, and we reduce its argument to $\frac{4n^2-11n+8}{n(n-1)^2}$. Now, to prove $\tau_N < 1$, it suffices to show

$$\begin{aligned} & \frac{2n-1}{n-1} - \frac{1}{n-1}\left(\frac{4n^2-11n+8}{n}\right)^{1/2} < 2 \\ \iff & 1 < \left(\frac{4n^2-11n+8}{n}\right)^{1/2} \\ \iff & 0 < n^2-3n+2, \end{aligned}$$

which factors into $(n-2)(n-1)$ and is trivially positive when $n \geq 3$. □

5.3.3 Proof of Theorem 9

We proceed to the proof of Theorem 9, which uses the following variation of Lemma 5.

Lemma 19 (See Lemma 5)**.** *Under the hypotheses of Theorem 9, set $g \in \mathbb{C}^d, \Lambda \in \mathbb{C}^{d\times d}$ by*

$$g_i = (\underline{x})_i^* x_i \quad \textit{and} \quad \Lambda_{ij} = (\underline{\widetilde{X}})_{ij}^* \widetilde{X}_{ij}.$$

Then

$$\sum_{(i,j)\in E} |g_i - g_j|^2 \leq 4\|\widetilde{X} - \underline{\widetilde{X}}\|_F^2.$$

Proof of Lemma 19. By an argument identical to that used in (3.27) of Lemma 5, we have

$$\begin{aligned} \sum_{(i,j)\in E}\left(\frac{1}{2}|g_i-g_j|^2 - |\Lambda_{ij}-1|^2\right) &\leq \sum_{(i,j)\in E}|x_i - \widetilde{X}_{ij}x_j|^2 = x^*Lx \\ &\leq \underline{x}^*L\underline{x} \;=\; \sum_{(i,j)\in E}|\underline{x}_i - \widetilde{X}_{ij}\underline{x}_j|^2, \end{aligned}$$

where the second inequality comes from optimality of x. Additionally, we observe that, just as in (3.29),

$$\sum_{(i,j)\in E} |\Lambda_{ij} - 1|^2 = \sum_{(i,j)\in E} |\underline{x}_i - \widetilde{X}_{ij}\underline{x}_j|^2 = \left\|\widetilde{X} - \underline{\widetilde{X}}\right\|_F^2.$$

Combining these two immediately gives the lemma. □

Proof of Theorem 9. We take $\underline{L} = D - W \circ \underline{\widetilde{X}}$ as usual and recognize that, as in (5.18),

$$x^* \underline{L} x \geq \frac{\tau_G}{2} \min_{\theta\in\mathbb{R}} \|x - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2^2.$$

Since we additionally have, as in (3.26), that

$$x^* \underline{L} x = \sum_{(i,j)\in E} |x_i - \underline{x}_i \underline{x}_j^* x_j|^2 = \sum_{(i,j)\in E} |g_i - g_j|^2,$$

this gives

$$\frac{\tau_G}{2} \min_{\theta\in\mathbb{R}} \|x - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2^2 \leq \sum_{(i,j)\in E} |g_i - g_j|^2,$$

which suffices with Lemma 19 to complete the proof. □

Substituting these results into the proof of Corollary 2, we gain an immediate improvement.

Corollary 7 (See Corollary 2)**.** *Let $\widetilde{X}_0$ be the matrix in* (3.7)*, $\widetilde{\mathbf{x}}_0$ be the vector of true phases* (3.7)*, and $\widetilde{X}$ be as in line 3 of Algorithm 1 with $\widetilde{\mathbf{x}}$ the optimizer of* (5.3) *using $L = I - \frac{1}{2\delta-1}\widetilde{X}$. Suppose that $\|\widetilde{X}_0 - \widetilde{X}\|_F \leq \eta \|\widetilde{X}_0\|_F$ for some $\eta > 0$. Then*

$$\min_{\theta\in[0,2\pi]} \|\widetilde{\mathbf{x}}_0 - \mathrm{e}^{\mathrm{i}\theta}\widetilde{\mathbf{x}}\|_2 \leq 3\frac{\eta d^{\frac{3}{2}}}{\delta}.$$

Proof of Corollary 7. We apply (5.21) with

$$\tau_G = \tau_N(2\delta - 1) \geq \frac{\pi^2\delta^2}{6d^2}(2\delta - 1) \text{ and } \|X - X_0\|_F \leq \eta\sqrt{d(2\delta - 1)},$$

and reduce the constant by observing $2\sqrt{2}/(\pi/\sqrt{6}) \leq 3$. □

This leads us to construct a variant of Algorithm 1, replacing the eigenvector-based angular synchronization stage with an exact solution of the angular synchronization problem; this variant is stated in Algorithm 5. With this in hand, we replace Corollary 2 with Corollary 7 in the proofs of the final error bounds, Theorem 5 and Corollary 3, and immediately obtain the following improvement over Corollary 3.

Algorithm 5 Phase Retrieval from Local Correlation Measurements, with SDP

Input: Measurements $\mathbf{y} \in \mathbb{R}^D$ as per (3.8)
Output: $\mathbf{x} \in \mathbb{C}^d$ with $\mathbf{x} \approx \mathrm{e}^{-\mathrm{i}\theta}\mathbf{x}_0$ for some $\theta \in [0, 2\pi]$

1: Compute the Hermitian matrix $X = \left((\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})})^{-1}\mathbf{y}\right)/2 + \left((\mathcal{A}|_{T_\delta(\mathbb{C}^{d\times d})})^{-1}\mathbf{y}\right)^*/2 \in T_\delta(\mathbb{C}^{d\times d})$ as an estimate of $T_\delta(\mathbf{x}_0\mathbf{x}_0^*)$
2: Form the banded matrix of phases, $\tilde{X} \in T_\delta(\mathbb{C}^{d\times d})$, by normalizing the non-zero entries of X
3: Compute $\hat{Z}$, the solution to (5.6) with $L = (2\delta - 1)I - \widetilde{X}$, and take $\widetilde{\mathbf{x}} = \mathrm{sgn}(u)$, where u is the top eigenvector of $\hat{Z}$.
4: Set $x_j = \sqrt{X_{j,j}} \cdot (\tilde{x})_j$ for all $j \in [d]$ to form $\mathbf{x} \in \mathbb{C}^d$

Corollary 8 (See Corollary 3)**.** *Using the notation of Corollary 7, let* $(x_0)_{\min} := \min_j |(x_0)_j|$ *be the smallest magnitude of any entry in* $\mathbf{x}_0$ *and suppose* $\|\widetilde{X} - \widetilde{X}_0\|_2 < \delta^2/d^{5/2}$*. Then the estimate* $\mathbf{x}$ *produced by Algorithm 5 satisfies*

$$\begin{aligned}
\min_{\theta\in[0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 &\leq 3\left(\frac{\|\mathbf{x}_0\|_\infty}{(x_0)^2_{\min}}\right)\left(\frac{d}{\delta^{3/2}}\right)\sigma_{\min}^{-1}\|\mathbf{n}\|_2 + d^{\frac{1}{4}}\sqrt{\sigma_{\min}^{-1}\|\mathbf{n}\|_2}, \\
\min_{\theta\in[0,2\pi]} \|\mathbf{x}_0 - \mathrm{e}^{\mathrm{i}\theta}\mathbf{x}\|_2 &\leq 3\left(\frac{\|\mathbf{x}_0\|_\infty}{(x_0)^2_{\min}}\right)\left(\frac{d}{\delta^{3/2}}\right)\kappa\frac{\|X_0\|_F}{\mathrm{SNR}} + d^{\frac{1}{4}}\sqrt{\kappa\frac{\|X_0\|_F}{\mathrm{SNR}}}.
\end{aligned} \tag{5.24}$$

5.3.4 Results with Weighted Graphs

While the extrication of a d/δ factor from this inequality is considerable, we remark first of all that Theorem 8, besides establishing a guaranteed basin wherein (5.3) may be solved exactly by semidefinite programming, plays a somewhat muted role in these new bounds. Specifically, (5.8) is not quoted in this section at all, which means that we have no clear statement of what the admission of weighted graphs in Theorem 8 accomplishes for us.

This defies intuition, since weighting our cost function should push the angular synchronization solver to focus on getting the phases of the large magnitude entries of x correct,

and to trust their phase measurements more (as discussed in Section 5.2). However, the proof techniques we have used so far are not sophisticated enough to quantify these benefits, which is not surprising, seeing that they derive primarily from model-specific intuition. In particular, bounding $\||\mathbf{x}_0| \circ (\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}_0)\|_2$ by $\|\mathbf{x}_0\|_\infty \|\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}_0\|_2$ wastes any potential "cooperation" between the magnitudes of $\mathbf{x}_0$ the accuracy of the phases in $\widetilde{\mathbf{x}}$, but it is not straightforward to show how the weight matrix W would guarantee such cooperation.

In [57], the authors approach this a bit differently by bounding the same "phase error" term with $\|\mathbf{x}_0\|_2 \|\widetilde{\mathbf{x}}_0 - \widetilde{\mathbf{x}}\|_\infty$. The angular synchronization algorithm they use involves finding a spanning tree of G, which admits a fairly clean analysis of $\widetilde{\mathbf{x}}$, but the assumptions required to bound $\|\widetilde{\mathbf{x}}_0 - \widetilde{\mathbf{x}}\|_\infty$ are heavy. In Section 5.2.4, we will take a closer look at such spanning tree methods and make an attempt to improve upon the results of this type as found in [57].

For the time being, to satiate these intuitions, we state the results that may be compiled from the existing theory in Corollary 9, and perform in Section 5.4 a numerical study that compares weighted vs. unweighted angular synchronization.

Corollary 9. *Given $\underline{x} \in \mathbb{C}^d$, suppose $\underline{X} = \underline{x}\underline{x}^* \circ (I + A_G)$, where A_G is the unweighted adjacency matrix of the connected graph $G = (V = [d], E)$. Suppose further that $X \in \mathcal{H}^d$ shares the sparsity structure of $\underline{X}$ (namely, $X = X \circ (I + A_G)$). We then define W, D, and L by*

$$W_{ij} = \begin{cases} 0, & i = j \\ |X_{ij}|, & \textit{otherwise} \end{cases}, \quad D = \operatorname{diag}(W\mathbb{1}),$$

$$L = D - W \circ \operatorname{sgn}(X), \quad \textit{and} \quad \underline{L} = D - W \circ \operatorname{sgn}(\underline{X}).$$

Then, setting $x_i = |X_{ii}|^{1/2} \hat{z}_i$, where

$$\hat{z} \in \operatorname*{argmin}_{z \in (\mathbb{S}^1)^d} z^* L z,$$

we have

$$\min_{\theta \in [0, 2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta} \underline{x}\|_2 \leq \|\underline{x}\|_\infty \frac{4\sqrt{2d} \|X - \underline{X}\|_2}{\tau_W} + d^{1/4} \sqrt{\|X - \underline{X}\|_F}, \tag{5.25}$$

where $\tau_W = \lambda_2(D - W)$.

Proof of Corollary 9. We split the norm into

$$\begin{aligned}\min_{\theta\in[0,2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 &\leq \||\underline{x}| \circ (\operatorname{sgn}(x) - \operatorname{sgn}(\underline{x}))\|_2 + \||\underline{x}| - |x|\|_2 \\ &\leq \|\underline{x}\|_\infty \|\operatorname{sgn}(x) - \operatorname{sgn}(\underline{x})\|_2 + \||\underline{x}| - |x|\|_2 \\ &\leq \|\underline{x}\|_\infty \frac{2\sqrt{2d}\|L - \underline{L}\|_2}{\tau_W} + d^{1/4}\sqrt{\|X - \underline{X}\|_F},\end{aligned}$$

where the last inequality comes from Theorem 8 and Lemma 7. To complete the proof, we bound $\|L - \underline{L}\|_2$ by setting

$$\underline{W} = \begin{cases} 0, & i = j \\ |\underline{X}_{ij}|, & \text{otherwise} \end{cases}$$

and taking

$$\begin{aligned}\|L - \underline{L}\|_2 &= \|W \circ (\operatorname{sgn} X - \operatorname{sgn} \underline{X})\|_2 \\ &\leq \|W \circ \operatorname{sgn} X - \underline{W} \circ \operatorname{sgn} \underline{X}\|_2 + \|(W - \underline{W}) \circ \operatorname{sgn} \underline{X}\|_2 \\ &\leq \|X - \underline{X}\|_2 + \|W - \underline{W}\|_2 \\ &\leq 2\|X - \underline{X}\|_2\end{aligned}$$

The second inequality in this series comes from considering that $W \circ \operatorname{sgn} X - \underline{W} \circ \operatorname{sgn} \underline{X}$ is simply $X - \underline{X}$ without its diagonal entries, and the third comes from $||a| - |b|| \leq |a - b|$ for any $a, b \in \mathbb{C}$. □

Comparing Eqs. (5.24) and (5.25), we notice that (5.25) uses yet another variation on the spectral gap: $\tau_W = \lambda_2(D - W)$, the spectral gap of the graph weighted by the entries of X. Not only do we see τ_W in the denominator rather than $\sqrt{\tau_W}$ (we fear, but do not necessarily expect, spectral gaps to be small), but τ_W is a massively unpredictable quantity, possibly having very little to do with the unweighted graph sitting beneath it. Indeed, even

if $G = K_n$ is a complete graph, by taking another graph $G' = ([n], E')$ and weighting G with

$$W_{ij} = \begin{cases} 1, & (i,j) \in E' \\ \epsilon, & \text{otherwise} \end{cases},$$

then $\lim_{\epsilon \to 0} \tau_W(\epsilon) = \tau_{G'}$, where $\tau_{G'} = D_{G'} - A_{G'}$ is the unweighted spectral gap of G'. In other words, weak entries (and, therefore, small weights) of X have the potential to effectively *disconnect* – meaning "drastically reduce the spectral gap of" – the graph on which we are depending for our angular synchronization. In principle, (5.24) is punished for small entries of X in the term $(x_0)^2_{\min}$ appearing in the denominator of the phase error bound, but it is hard to gain a theoretical statement telling us whether τ_W is a more efficient means of quantifying the extent to which noise and small entries of $\underline{x}$ "disconnect" our angular synchronization graph. For the moment, we relegate this question to the numerical study of Section 5.4.

5.4 Numerical Experiments

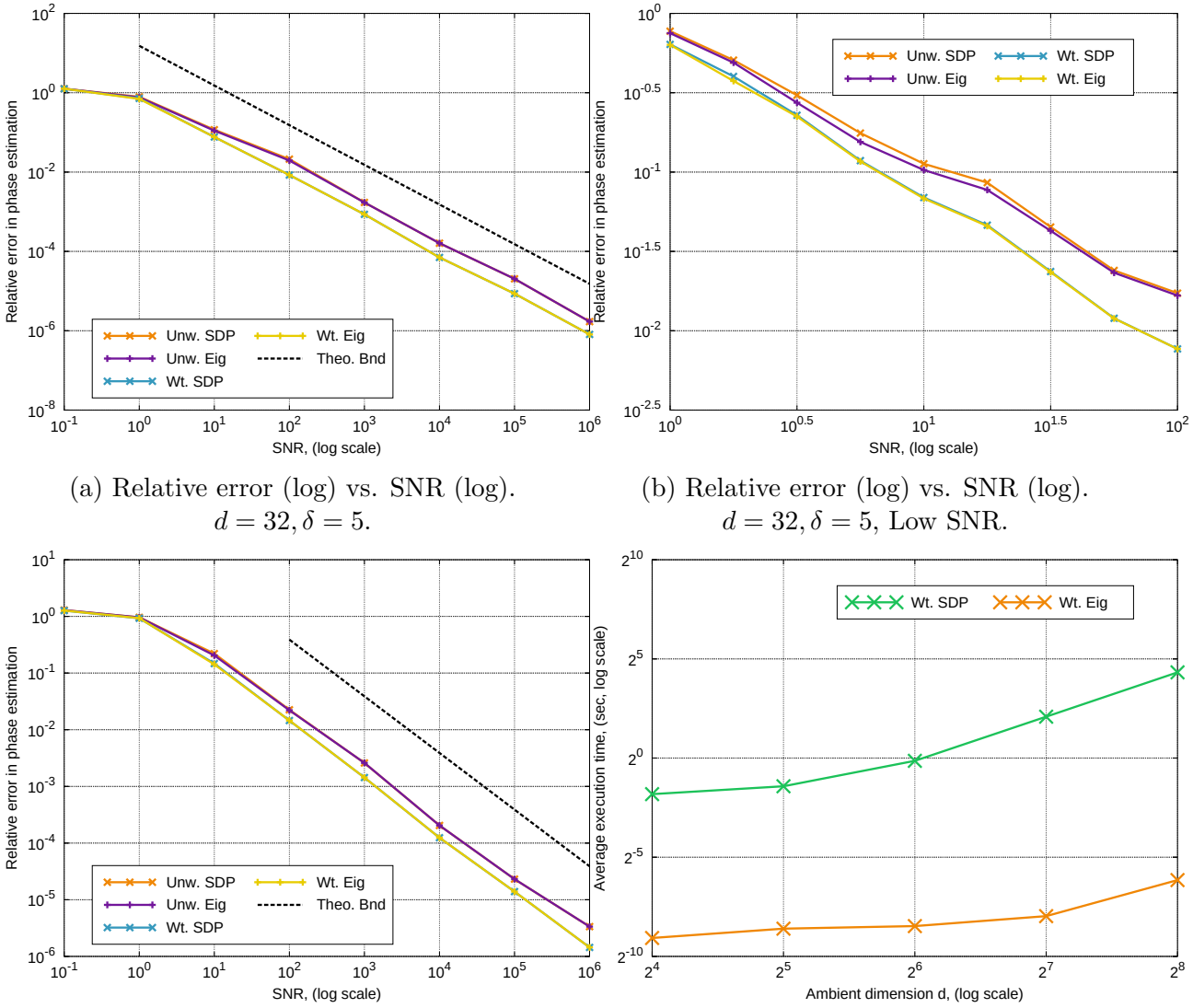
In evaluating the angular synchronization methods studied in this chapter, our only considerations are reconstruction error and execution time. As we shall see, there is an extremely cogent tradeoff to be navigated here: the SDP recovery algorithm of Algorithm 3 is significantly more costly than the simple eigenvector solve of Algorithm 2, whereas the tighter theoretical results are only proven for the SDP method. Thankfully, the conventional wisdom of the angular synchronization literature is that, while SDP methods are more easily analyzed (having decades of convex optimization theory behind them), the eigenvector method, which solves a convex problem and then imposes an egregiously non-linear transformation on this result, performs just as well numerically. Singer found this in his seminal paper [84], and similar results were replicated in [18], in which SDP was actually the *worst* performer on noise compared to three other methods (an eigenvector method similar to ours, Gauss-Newton minimization, and a third heuristic method not discussed here). Therefore, our goal in this section is to demonstrate numerically that the cheaper Algorithms 2 and 4 gain enough in speed that their often negligible disadvantages in reconstruction error are justified.

To this end, Fig. 5.2 makes a comparison of the reconstruction performance of the SDP method described in Algorithm 3 versus that of the eigenvector method of Algorithm 2. The two main takeaways from these results are that the SDP appears to confer negligible accuracy to the reconstruction, whereas retaining the weights in the Laplacian (as discussed, largely without proof, in Section 5.3.4) gives a substantial improvement to signal recovery at no computational cost. In Figs. 5.2a and 5.2b, we used $d = 32, \delta = 5$. Here, we randomly generated 32 examples $x \sim \mathcal{CN}(0, I_n)$ and wrote, for the *weighted* experiments, $Y = T_\delta(xx^*) + N$, where N is a member of $T_\delta(\mathcal{H}^d)$ with Gaussian entries, scaled appropriately for SNR.[5] We then set $X = Y - \mathrm{diag}(Y)$ and devised the graph Laplacian by taking $W = |X|$ such that $L = \mathrm{diag}(W\mathbb{1}) - W \circ \mathrm{sgn}(X)$. For the unweighted experiments, we simply normalized the entries of Y, to obtain $\widetilde{X} = \mathrm{sgn}(Y - \mathrm{diag}(Y))$ and $\tilde{L} = \mathrm{diag}(|\widetilde{X}|\mathbb{1}) - \widetilde{X}$. In Fig. 5.2c, we implemented the same idea, but with a different truncation $T_{\delta,s}(\mathcal{H}^d)$, to be introduced in Section 6.1.1. Figures 5.2a and 5.2c take a wide range of SNRs, while Fig. 5.2b zooms into a closer view of the low SNR setting.

The verdict is clear: as observed elsewhere in the angular synchronization literature, the eigenvector-based recovery method of Algorithm 2 works just as well as the SDP method, even though the theoretical bounds for SDP are more attractive (and more straightforwardly provable). As a matter of fact, Fig. 5.2b even shows that, at extremely low SNRs, the eigenvector method works *better* in some instances, although this could be a statistical artifact that fails to appear under other noise models. Nicely, our intuition in Section 5.3.4 of weighting the edges according to the (pairwise products of the) magnitudes of the nodes notably boosts performance – far more than does switching from eigenvector recovery to SDP. We have also included the theoretical bounds of Theorem 8 (for the unweighted graph) for reference, illustrating the theorem and showing that, in practice, average recovery is stronger than the worst-case by about an order of magnitude. To underscore the appeal of this numerical result, we compare the computational cost of SDP optimization versus eigenvector recovery in Fig. 5.2d – as expected, the eigenvector method is massively faster,

[5]Specifically, we take $N = \frac{N\|T_\delta(xx^*)\|_F}{\|N'\|_F \,\mathrm{SNR}}$, where $N' \in T_\delta(\mathcal{H}^d)$ is such that $N'_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0,1)$ for $i < j$ and $N'_{ii} \overset{\text{i.i.d.}}{\sim} \mathbb{N}(0,1)$, with the lower triangular part determined by Hermitianity.

performing in less than one hundredth of a second at $d = 64$, when the convex relaxation is already taking several seconds to complete. Considering that the eigenvector method is tied with the SDP for accuracy, and that both are empirically beating the theoretical bound by about a factor of 10, we take this as substantial evidence that we may safely continue using the eigenvector-based angular synchronization strategy in deployable implementations.



(a) Relative error (log) vs. SNR (log). $d = 32, \delta = 5$.

(b) Relative error (log) vs. SNR (log). $d = 32, \delta = 5$, Low SNR.

(c) Relative error on $T_{\delta,s}$ (log) vs. SNR (log). $d = 32, \delta = 5, s = 4$.

(d) Execution time vs. Problem size. $\delta = 4$ throughout.

Figure 5.2: Angular synchronization over $T_\delta(\mathcal{H}^d)$ by SDP relaxation and eigenvector recovery. Weighted vs. Unweighted graphs. Reconstruction Accuracy and Execution Time

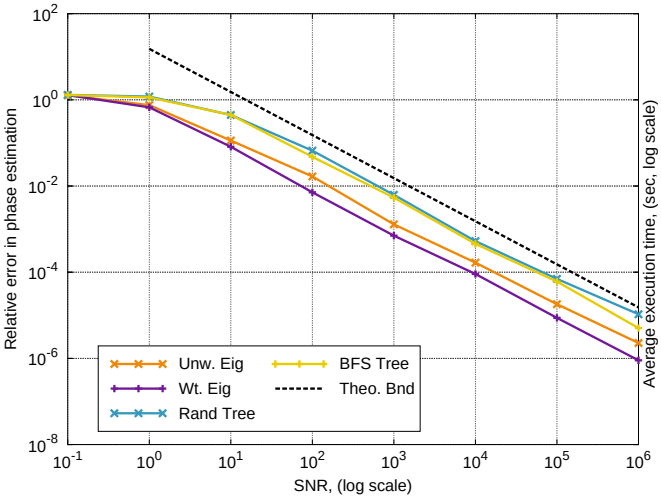
Taking Algorithm 2 as an acceptable yardstick, a brief comparison is made between this method and a couple of variants of the tree-based algorithm defined in Section 5.2.4,

and we see that, overall, the loss in accuracy is not justified by the gain in runtime achieved by the tree-based methods. We run an experiment similar in form to that in Fig. 5.2 for $d = 32$, $\delta = 5$ (and $s = 1$) in Fig. 5.3a, remarking that we are comparing two methods of finding spanning trees of the graph G (which, in this case, is $G_{d,\delta}$). One, marked "BFS Tree," refers to simply taking a breadth-first search beginning from vertex 1 (the noise is iid across vertices, and the graph is cyclical, so this is a perfectly "symmetric" choice), whereas the other uses an algorithm described in [104] to draw a spanning tree uniformly from the set of all spanning trees of $G_{\delta,d}$. Considering again that $G_{\delta,d}$ is cyclical, a breadth-first search will always find a minimum-diameter spanning tree, which, according to (5.20), is favorable for angular synchronization. Indeed, Fig. 5.3 shows that these minimum-diameter trees unmistakably outperform the random trees, in both reconstruction accuracy and runtime. While the random tree algorithm may hold interest in cases where a minimum diameter tree is not obviously obtainable from the structure of the problem, for the moment it has yielded no practical benefit, besides to justify the intuition that supports the BFS method.

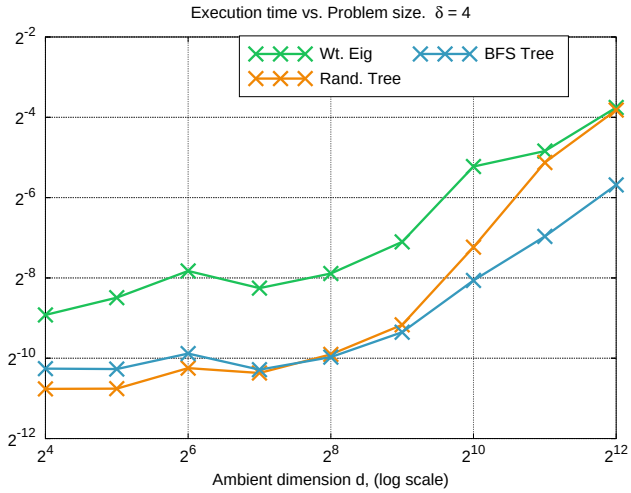
At any rate, in comparing tree-based to eigenvector synchronization, the conspicuous loss in reconstruction accuracy compared to both the weighted and unweighted eigenvector methods completely offsets the trees' meager improvement in runtime. Figures 5.3b and 5.3c show that the tree-based methods enjoy only a factor of 8 or so faster execution, which only becomes relevant in the extremely large problem sizes of $d \approx 1000$ or greater (at which point a $2^{10} \times 2^{10}$ SDP is all but impossible). To emphasize this observation, Fig. 5.3c looks at solve time vs problem size when $\delta = d/4$, such that the measurement complexity $d(2\delta - 1)$ – and therefore the edge count, which slows the tree methods, and the sparsity constant, which slows the matrix multiplies in the eigenvector solve – grows quadratically with d. Here, the $\mathcal{O}(d^2)$ execution time growth is observed clearly for all three synchronization algorithms, and the 8x speedup afforded by the tree methods is certainly not enough to justify paying the 3-10x penalty in accuracy.

On the other hand, we remark that, for extremely large problem sizes (such as the 2 dimensional case of Chapter 7) *and* with our highly structured adjacency graph, it may be possible that the BFS tree method is a practical choice. In a setup such as this, where

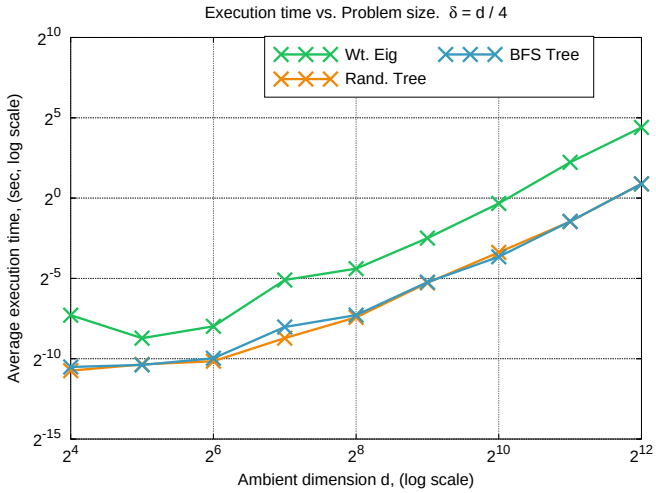
the eigenvector solve can easily take half a minute to complete, it may be that pre-storing a known spanning tree (such as the one described in Section 5.2.4) and propagating phases along it in $\mathcal{O}(d)$ time becomes the only practical implementation. This would hold especially true at high SNRs when these dB losses in reconstruction accuracy are inconsequential to the application.



(a) Relative error (log) vs. SNR (log). $d = 32, \delta = 5$.



(b) Execution time vs. Problem size. $\delta = 4$ throughout.



(c) Execution time vs. Problem size. $\delta = d/4$.

Figure 5.3: Execution time vs. problem size and reconstruction error vs. SNR, Angular synchronization over $T_\delta(\mathcal{H}^d)$ by eigenvector recovery and tree-based propagation.

Chapter 6

Ptychographic Model

6.1 Setting and Notation

In our model for the ptychographic setup of (3.2), we have so far assumed that measurements are taken corresponding to all shifts $\ell \in [d]_0$; in the notation of (3.2), this is equivalent to taking $P = [d]_0$. This chapter analysis a useful generalization to the case where $P = s[d/s]_0$, where $s \in \mathbb{N}$ is a divisor of d.

The motivation for studying this case is that, unfortunately, in practice, taking $P = [d]_0$ is usually an impossibility, since in many cases an illumination of the sample can cause damage to the sample [87], and applying the illumination beam (which can be highly irradiative) repeatedly at a single point can destroy it. In ptychography as it is usually performed in the lab, the beam is shifted by a far larger distance than the width of a single pixel[1] – instead of overlapping on $\delta - 1$ of δ pixels, adjacent illumination regions will typically overlap on a percentage of their support on the order of 50% or even less [25, 83]. Considering the risks to the sample and the costs of operating the measurement equipment, there are strong incentives to reduce the number of illuminations applied to any object, and therefore our theory ought to address a model that reflects this concern.

[1]The difficulty of moving the illumination apparatus at a scale equal to the desired optical resolution is another reason taking $P = [d]_0$ is a cumbersome assumption.

6.1.1 Measurement Operator and Its Domain

Towards this model, instead of using all shifts in our lifted measurement system, we instead fix a shift size $s \in \mathbb{N}$ where $d = \overline{d}s$ with $\overline{d} \in \mathbb{N}$ and use $S^{s\ell} m_j m_j^* S^{-s\ell}$ for $\ell \in [\overline{d}]_0$. In this light, we introduce the following generalization of the lifted measurement system: given a family of masks of support δ, $\{m_j\}_{j\in[D]} \subseteq \mathbb{C}^d$, and $s, \overline{d} \in \mathbb{N}$ with $\overline{d} = d/s$, the associated *lifted measurement system of shift* s is the set

$$\mathcal{L}^s_{\{m_j\}} := \{S^{s\ell} m_j m_j^* S^{-s\ell}\}_{(\ell,j)\in[\overline{d}]_0\times[D]} \subseteq \mathbb{C}^{d\times d}. \tag{6.1}$$

This leads to an obvious redefinition of the measurement operator, now $\mathcal{A} : \mathbb{C}^{d\times d} \to \mathbb{R}^{[\overline{d}]_0\times[D]}$:

$$\mathcal{A}(X)_{(\ell,j)} = \langle S^{s\ell} m_j m_j^* S^{-s\ell}, X\rangle, \quad (\ell, j) \in [\overline{d}]_0 \times [D]. \tag{6.2}$$

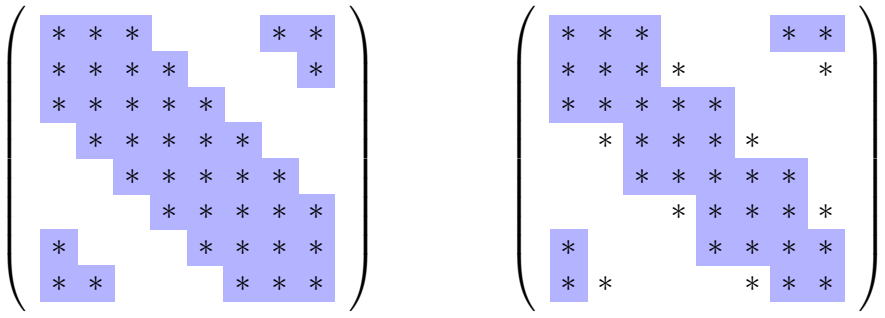
This will also force us to reconsider the subspace of $\mathbb{C}^{d\times d}$ with which we are working in the domain of $\mathcal{A}$, since it is clearly impossible, by inspection of Fig. 6.1, for $\mathcal{L}^s$ to span $T_\delta(\mathbb{C}^{d\times d})$ with a shift size $s > 1$. In an effort to define the subspace analogous to $T_\delta(\mathbb{C}^{d\times d})$ in the ptychographic case, we let $\mathcal{J}_{\delta,s} = \bigcup_{\ell\in[\overline{d}]_0} \operatorname{supp}(S^{s\ell}\mathbb{1}_{[\delta]}\mathbb{1}^*_{[\delta]}S^{-s\ell})$ be the set of indices "reached" by this system, and we let

$$T_{\delta,s}(X) = \begin{cases} X_{ij}, & (i,j) \in \mathcal{J}_{\delta,s} \\ 0, & \text{otherwise} \end{cases} \tag{6.3}$$

be the projection onto the associated subspace of $\mathbb{C}^{d\times d}$. $T_{\delta,s}$ is visualized in Fig. 6.1. To get a feel for $\mathcal{J}_{\delta,s}$, we observe that

$$\left(S^{s\ell} m_k m_k^* S^{-s\ell}\right)_{ij} = (S^{s\ell} m_k)_i (\overline{S^{s\ell} m_k})_j = (m_k)_{i-s\ell} (\overline{m_k})_{j-s\ell},$$

so $\left(S^{s\ell} m_k m_k^* S^{-s\ell}\right)_{ij} = 0$ when $(i - s\ell, j - s\ell) \notin [\delta]^2$, i.e. when $(i,j) \notin [\delta]^2_{s\ell+1}$. Hence the indices onto which we are projecting are those in $\mathcal{J}_{\delta,s} = \bigcup_{\ell\in[\overline{d}]_0} [\delta]^2_{s\ell+1}$. This set may be revisualized by calculating which j's are admissible for each i; for a fixed i, we look at all



(a) $T_3(\mathbb{C}^{8\times 8})$ (b) $T_{3,2}(\mathbb{C}^{8\times 8})$

Figure 6.1: $T_\delta(\mathbb{C}^{d\times d})$ vs. $T_{\delta,s}(\mathbb{C}^{d\times d})$ for $d = 8, \delta = 3, s = 2$

shifts ℓ such that $i \in [\delta]_{s\ell+1}$, and j is allowed to be in their union.

In the (pathological) case where $s \geq \delta$, obviously any given index can only appear in one of the $[\delta]_{s\ell+1}$, namely $i \in [\delta]_{s\ell+1}$ iff $i \bmod s \leq \delta$ and $\lfloor i/s \rfloor = \ell$, so in this case we would have

$$\mathcal{J}_{\delta,s} = \{(i,j) : \lfloor i/s \rfloor = \lfloor j/s \rfloor \text{ and } i \bmod s, j \bmod s \leq \delta\}.$$

However, this case is not typical, since $T_{\delta,s}(\mathbb{1}\mathbb{1}^*)$ will be the adjacency matrix of an unconnected graph, and there will be groups of coordinates whose relative phases are completely undetermined by $T_{\delta,s}(xx^*)$, which makes such an arrangement unfeasible from a phase retrieval point of view. For example, for any $\theta \in \mathbb{R}$, we have

$$T_{2,2}\left(\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\begin{bmatrix}1\\1\\1\\1\end{bmatrix}^*\right) = T_{2,2}\left(\begin{bmatrix}1\\1\\e^{i\theta}\\e^{i\theta}\end{bmatrix}\begin{bmatrix}1\\1\\e^{i\theta}\\e^{i\theta}\end{bmatrix}^*\right) = \begin{bmatrix}1&1&0&0\\1&1&0&0\\0&0&1&1\\0&0&1&1\end{bmatrix}$$

6.2 Conditioning of $\mathcal{A}$ for Ptychography

With this setup in hand, we begin our analysis of the linear system $\mathcal{A}(X) = y$ with a number of lemmas that unravel the structure of this operator. Our goal will be to proceed similarly to Section 4.2 by rewriting $\mathcal{A}$ as a product of block-circulant matrix with certain

permutations, at which point we will be able to cite Corollary 5, which renders a convenient expression for the condition number.

6.2.1 New Operators

In service of this strategy, in this section we introduce a few new operators that are useful in the analysis of $\mathcal{A}$. Of these, $\mathcal{T}_N$ and $\mathcal{P}$ are crucial in generalizing and adding to the results on circulant matrices found in Section 4.2.1, and $J_{(k_1,k_2,n)}$ is necessary to construct a full-rank vectorization of $T_{\delta,s}$.[2] For $N \in \mathbb{N}$, we define $\mathcal{T}_N : \bigcup_{\ell\in\mathbb{N}} \mathbb{C}^{\ell N\times m} \to \bigcup_{\ell\in\mathbb{N}} \mathbb{C}^{\ell m\times N}$, the blockwise transpose operator, defined by

$$\mathcal{T}_N\left(\begin{bmatrix} V_1 \\ \vdots \\ V_\ell \end{bmatrix}\right) = \begin{bmatrix} V_1^* \\ \vdots \\ V_\ell^* \end{bmatrix} \tag{6.4}$$

for $V_1, \ldots, V_\ell \in \mathbb{C}^{N\times m}$.

We also define, for $(k_j)_{j=1}^n$ and permutation $P \in \{0,1\}^{n\times n}$, the *blockwise permutation operator* $\mathcal{P}(P,(k_j)) : \mathbb{C}^{K\times K} \to \mathbb{C}^{K\times K}$, where $K = \sum_{j=1}^n k_j$. Our intention will be to permute the blocks of a block vector $\begin{bmatrix} v_1^T & \cdots & v_n^T \end{bmatrix}^T$, where $v_j \in \mathbb{C}^{k_j}$. In order to specify $\mathcal{P}(P,(k_j))$ precisely, we permit an overloading of notation on permutations: namely, if $P \in \{0,1\}^{m\times m}$ is a permutation, then we identify P with the mapping $\pi : [m] \to [m]$ where $\pi(i) = j$ whenever $Pe_i = e_j$. In particular, if we write $P(i)$, we mean "j such that $Pe_i = e_j$." With this in mind, $\mathcal{P}(P,(k_j))$ is defined, for $v_j \in \mathbb{C}^{k_j}$, by

$$\mathcal{P}(P,(k_j)) \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} v_{P(1)} \\ \vdots \\ v_{P(n)} \end{bmatrix}. \tag{6.5}$$

Another way of stating $\mathcal{P}(P,(k_j))$ is that it takes P and replaces its j^{th} (from left to right)

[2] That is, we need $J_{(k_1,k_2,n)}$ to avoid the unused entries, as visualized in Fig. 6.1.

1 with I_{k_j}, resizing as necessary. For example, if

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad \text{then} \quad \mathcal{P}(P,(1,2,3)) = \begin{bmatrix} 0 & 0 & I_3 \\ 0 & I_2 & 0 \\ I_1 & 0 & 0 \end{bmatrix} = \left[\begin{array}{c|cc|ccc} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

Defined by blocks, we have $\mathcal{P}(P,(1,2,3)) \begin{bmatrix} v_1^T & v_2^T & v_3^T \end{bmatrix}^T = \begin{bmatrix} v_3^T & v_2^T & v_1^T \end{bmatrix}$, or, for example,

$$\mathcal{P}(P,(1,2,3)) \left[\begin{array}{c} 1 \\ \hline 2 \\ 3 \\ \hline 4 \\ 5 \\ 6 \end{array}\right] = \left[\begin{array}{c} 4 \\ 5 \\ 6 \\ \hline 2 \\ 3 \\ \hline 1 \end{array}\right].$$

We also introduce the indexed family of matrices $J_{(k_1,k_2,n)}$, for $k_1, k_2, n \in \mathbb{N}$ with $k_1 \leq k_2$. The goal here is to have a matrix that includes the columns of I_n corresponding to the interval $[k_1, k_2) \bmod n$. Recalling the convention of taking indices modulo n, we set

$$J_{(k_1,k_2,n)} = \begin{cases} \begin{bmatrix} e^n_{k_1} & \cdots & e^n_{k_2-1} \end{bmatrix}, & k_2 - k_1 < n \\ I_n, & \text{otherwise} \end{cases} \tag{6.6}$$

Then $J_{(k_1,k_2,n)} \in \mathbb{C}^{n \times \min\{n, k_2-k_1\}}$ represents an orthogonal basis for $\text{span}\{e_i : i \bmod n \in [k_1, k_2)\}$.[3]

[3]To be precise, when $S \subseteq \mathcal{S}$, where $\mathcal{S}$ is any set with an equivalence relation $\sim$ and equivalence classes $\{[x] : x \in \mathcal{S}\}$, by $[x] \in S$, $x \in S/\sim$, and $[x] \in S/\sim$, we mean $x \sim y$ for some $y \in S$.

6.2.2 Lemmas on Block Circulant Structure

We begin with Lemma 20, which describes the transposes of block circulant matrices. For this lemma and the remainder of this section, the reader is advised to recall the definitions of R_k and $P^{(d,N)}$ from Section 1.3 and (4.7), as well as $\mathcal{P}(P, \{k_i\})$ and $\mathcal{T}_N$ from Section 6.2.1.

Lemma 20. *Given $k, N, m \in \mathbb{N}$ and $V \in \mathbb{C}^{kN \times m}$, we have*

$$\operatorname{circ}^N(V)^* = \operatorname{circ}^m\left((R_k \otimes I_m)\mathcal{T}_N(V)\right).$$

Proof of Lemma 20. Suppose V_i are the $N \times m$ blocks of V, such that $V = \begin{bmatrix} V_1^T \cdots V_k^T \end{bmatrix}^T$. Indexing blockwise, we have $\operatorname{circ}^N(V)_{[ij]} = V_{i-j+1}$, so that $\operatorname{circ}^N(V)^*_{[ij]} = V^*_{j-i+1}$. In other words,

$$\operatorname{circ}^N(V)^* = \begin{bmatrix} V_1^* & V_2^* & \cdots & V_N^* \\ V_N^* & V_1^* & \cdots & V_{N-1}^* \\ \vdots & & \ddots & \vdots \\ V_2^* & V_3^* & \cdots & V_1^* \end{bmatrix} = \operatorname{circ}^m((R_k \otimes I_m)\mathcal{T}_N(V))$$

as claimed. □

Lemmas 21 and 22 provide identities for a few block matrix structures that will be of interest.

Lemma 21. *Given $N_1, N_2, k, m \in \mathbb{N}$ and $V_i \in \mathbb{C}^{kN_1 \times m}$ for $i \in [N_2]$, we have*

$$\begin{bmatrix} \operatorname{circ}^{N_1}(V_1) & \cdots & \operatorname{circ}^{N_1}(V_{N_2}) \end{bmatrix} (P^{(k,N_2)} \otimes I_m)^* = \operatorname{circ}^{N_1}\left(\begin{bmatrix} V_1 & \cdots & V_{N_2} \end{bmatrix}\right).$$

Proof of Lemma 21. We quote (4.9) from Lemma 11 and consider that $P^{(k,N_2)} \otimes I_m$ is a permutation that changes the blockwise indices of $m \times p$ blocks (or, acting from the right, $p \times m$ blocks) exactly the way that $P^{(k,N_2)}$ changes the indices of a vector. □

Lemma 22. *Given $k, n \in \mathbb{N}$ and $V_j \in \mathbb{C}^{m_j \times n_j}$ for $j \in [n]$ and setting $M = \sum_{j=1}^n m_j, N =$*

$\sum_{j=1}^{N} n_j$, *and* $D = \mathrm{diag}(I_k \otimes V_j)_{j=1}^{n} \in \mathbb{C}^{kM \times kN}$, *we have*

$$D = \begin{bmatrix} I_k \otimes V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I_k \otimes V_n \end{bmatrix} = P_1(I_k \otimes \mathrm{diag}(V_j)_{j=1}^{n})P_2^*$$

where $P_1 = \mathcal{P}(P^{(n,k)}, (m_{j \bmod_1 n})_{j=1}^{kn})$ *and* $P_2 = \mathcal{P}(P^{(n,k)}, (n_{j \bmod_1 n})_{j=1}^{kn})$.

Proof of Lemma 22. We immediately reduce to the case $m_j = n_j = 1$ for all j by observing that P_1 and P_2 will act on blockwise indices precisely as $P^{(n,k)}$ acts on individual indices. Here, we replace V_j with $v_j \in \mathbb{C}$, and note that $\mathrm{diag}(V_j)_{j=1}^{n} = \mathrm{diag}(v)$. Hence, we need only remark that

$$(\mathrm{diag}(I_k \otimes v_\ell)_{\ell=1}^{n})_{((i_1-1)k+i_2)((j_1-1)k+j_2)} = \begin{cases} v_{i_1}, & i_1 = j_1 \text{ and } i_2 = j_2 \\ 0, & \text{otherwise} \end{cases},$$

while

$$\begin{aligned} &(P^{(n,k)}(I_k \otimes \mathrm{diag}(v))P^{(n,k)*})_{((i_1-1)k+i_2)((j_1-1)k+j_2)} \\ &\quad = (I_k \otimes \mathrm{diag}(v))_{((i_2-1)n+i_1)((j_2-1)n+j_1)} \\ &\quad = \begin{cases} v_{i_1}, & i_1 = j_1 \text{ and } i_2 = j_2 \\ 0, & \text{otherwise} \end{cases}. \end{aligned}$$

□

6.2.3 Zero Columns in Matrix Representation of $\mathcal{A}$

To begin the discussion of the matrix representation of $\mathcal{A}$, we refresh our notation: $d, \delta, \overline{d}, s \in \mathbb{N}$ satisfy $d = \overline{d}s$ and $2\delta - 1 \leq d$. We have $D \in \mathbb{N}$ (arbitrary for now) measurement vectors $\{m_j\}_{j \in [D]} \in \mathbb{C}^d$ satisfying $1 \in \mathrm{supp}(m_j) \subseteq [\delta]$, and we set $g_m^k = \mathrm{diag}(m_k m_k^*, m)$ for all $1 - \delta \leq m \leq \delta - 1$ and all $k \in [D]$. The expressions $\mathcal{L}_{\{m_j\}}^s, T_{\delta,s}$, and $\mathcal{A}$ are defined in

Eqs. (6.1)–(6.3).

We now consider the question of when $\operatorname{span}\mathcal{L}^s_{\{m_j\}} = T_{\delta,s}$ and what the condition number of $\mathcal{A}$ will be. As in (4.4), we vectorize X by its diagonals[4] with $\mathcal{D}_\delta(X) \in \mathbb{C}^{d(2\delta-1)}$ and write $A \in \mathbb{C}^{\overline{d}D\times(2\delta-1)d}$ such that

$$(A\mathcal{D}_\delta(X))_{(j-1)\overline{d}+\ell} = \left(A \begin{bmatrix} \operatorname{diag}(X, 1-\delta) \\ \vdots \\ \operatorname{diag}(X, \delta-1) \end{bmatrix} \right)_{(j-1)\overline{d}+\ell} = \mathcal{A}(X)_{(\ell-1,j)}$$

$$= \left\langle S^{s(\ell-1)} m_j m_j^* S^{s(\ell-1)}, X \right\rangle = \sum_{m=1-\delta}^{\delta-1} S^{s(\ell-1)} g_m^{j*} \operatorname{diag}(X, m),$$

which gives the $(j-1)\overline{d}+\ell^{\text{th}}$ row of A as

$$\begin{bmatrix} S^{s(\ell-1)} g_{1-\delta}^j \\ \vdots \\ S^{s(\ell-1)} g_{\delta-1}^j \end{bmatrix}^*$$

so that, by Lemma 20, we have[5]

$$A = \begin{bmatrix} \operatorname{circ}^s(g_{1-\delta}^1) & \cdots & \operatorname{circ}^s(g_{1-\delta}^D) \\ \vdots & \ddots & \vdots \\ \operatorname{circ}^s(g_{\delta-1}^1) & \cdots & \operatorname{circ}^s(g_{\delta-1}^D) \end{bmatrix}^* = \begin{bmatrix} \operatorname{circ}(R_{\overline{d}}\mathcal{T}_s g_{1-\delta}^1) & \cdots & \operatorname{circ}(R_{\overline{d}}\mathcal{T}_s g_{\delta-1}^1) \\ \vdots & \ddots & \vdots \\ \operatorname{circ}(R_{\overline{d}}\mathcal{T}_s g_{1-\delta}^D) & \cdots & \operatorname{circ}(R_{\overline{d}}\mathcal{T}_s g_{\delta-1}^D) \end{bmatrix}. \tag{6.7}$$

However, because $T_{\delta,s} \subsetneq T_\delta$ when $s > 1$, this operator can never be invertible in such a case. In fact, when $s > 1$, A has several *completely zero* columns, corresponding precisely to the coordinates of entries in $T_\delta/T_{\delta,s}$,[6] in the sense that $\mathcal{D}_\delta(T_\delta/T_{\delta,s}) \subseteq \operatorname{Nul}(A)$. The remainder of this section is dedicated to explicitly stating an orthonormal basis for $\mathcal{D}_\delta(T_{\delta,s})$ by enumerating the indices of these zero columns. This is achieved in Proposition 17, which provides this basis as the columns of a matrix N, such that $N^*\mathcal{D}_\delta(X)$ will be a convenient vectorization

[4]Notice that this will force A, the matrix representing $\mathcal{A}$, to be singular. We expand on this later.

[5]For reference, we remark that $\operatorname{circ}^s(g_m^k) \in \mathbb{C}^{d\times\overline{d}}$ and $\operatorname{circ}(R_{\overline{d}}\mathcal{T}_s g_m^k) \in \mathbb{C}^{\overline{d}\times d}$.

[6]Here, by the quotient V/W of nested subspaces $W \subseteq V$, we mean $V \cap W^\perp$.

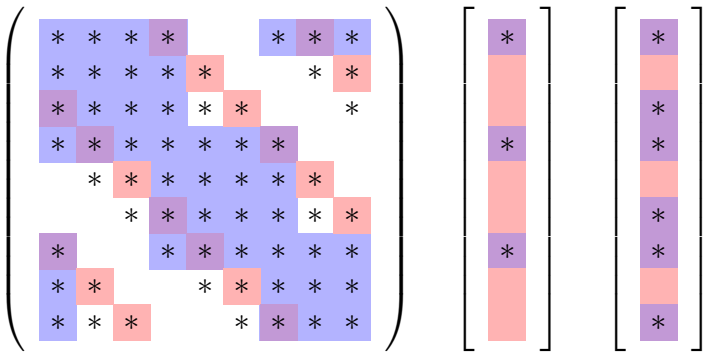


Figure 6.2: $T_{4,3}(\mathbb{C}^{9\times 9})$, and $\mathrm{supp}(g_3^j), \mathrm{supp}(g_{-2}^j)$

of $T_{\delta,s}(X)$ and AN will represent $\mathcal{A}|_{T_{\delta,s}(\mathbb{C}^{d\times d})}$. The proof is rather technical, and consists of constructing and verifying the proposed construction. For a first read, Fig. 6.2 may be a useful visual guide that stands to illuminate this otherwise opaquely stated result: here, we have given examples of the supports of the off-diagonals of $T_{\delta,s}(\mathbb{1}\mathbb{1}^*)$. These pictures may be corresponded with the expressions in (6.9) and (6.11).

Proposition 17. *Fix $d = \overline{d}s$ and δ satisfying $s, \overline{d}, \delta \in \mathbb{N}, 2\delta - 1 \leq d$. We let*

$$N = \mathrm{diag}(I_{\overline{d}} \otimes N_m)_{m=1-\delta}^{\delta-1}, \text{ where } N_m = \begin{cases} J_{(|m|+1,\delta+1,s)}, & m < 0 \\ J_{(1,\delta-m+1,s)}, & \text{otherwise} \end{cases}, \tag{6.8}$$

and we note that $N_m \in \mathbb{C}^{s\times \min\{s,\delta-|m|\}}$, so observing $\sum_{m=1-\delta}^{\delta-1} \min\{s, \delta - |m|\} = s(2\delta - s)$, we have $N \in \mathbb{C}^{d(2\delta-1)\times d(2\delta-s)}$. Then N is an orthogonal basis for $\mathcal{D}_\delta(T_{\delta,s}(\mathbb{C}^{d\times d}))$ and therefore $AN \in \mathbb{C}^{\overline{d}D\times d(2\delta-s)}$ is the matrix representation of $\mathcal{A}|_{T_{\delta,s}(\mathbb{C}^{d\times d})}$ with respect to the basis $N^*\mathcal{D}_\delta(T_{\delta,s}) \in \mathbb{C}^{d(2\delta-s)}$, in the sense that

$$(ANN^*\mathcal{D}_\delta(X))_{(j-1)\overline{d}+\ell} = (A\mathcal{D}_\delta(T_{\delta,s}(X)))_{(j-1)\overline{d}+\ell} = \mathcal{A}(X)_{(\ell-1,j)}.$$

Proof of Proposition 17. We begin by fixing a diagonal $m \geq 0$ and considering which indices along this diagonal will be encountered by the shifting masks $S^\ell m_j m_j^* S^{-\ell}$. We consider that

g_m^j, the m^{th} diagonal of $m_j m_j^*$, has support[7]

$$\operatorname{supp}(g_m^j) \subseteq [1, \delta - m] = [1, \delta - m + 1). \tag{6.9}$$

Therefore, the set of all indices in $[d]$ hit by shifts of g_m^j is given by

$$\bigcup_{\ell \in [\overline{d}]_0} [1, \delta - m + 1) + \ell s = \{i \in [d] : i \bmod s \in [1, \delta - m + 1)\}.$$

Recalling (6.6), it is clear that $N_m = J_{(1,\delta-m+1,s)} \in \mathbb{C}^{d \times \min\{s, \delta - m\}}$, repeated $\overline{d}$ times, will span exactly these indices, in the sense

$$(I_{\overline{d}} \otimes N_m)(I_{\overline{d}} \otimes N_m)^* \operatorname{diag}(\mathbb{C}^{d \times d}, m) = \operatorname{diag}(T_{\delta,s}(\mathbb{C}^{d \times d}), m). \tag{6.10}$$

Similarly, for $m < 0$, we have

$$\operatorname{supp}(g_m^j) \subseteq [|m| + 1, \delta + 1), \tag{6.11}$$

so we will simply use $N_m = I_{\overline{d}} \otimes J_{(|m|+1,\delta+1,s)}$ in this instance, and these N_m satisfy (6.10) for negative m. Synthesizing (6.10) for $m = 1 - \delta, \ldots, \delta - 1$, we have

$$\mathcal{D}_\delta(T_{\delta,s}(X)) = \begin{bmatrix} \operatorname{diag}(T_{\delta,s}(X), 1 - \delta) \\ \vdots \\ \operatorname{diag}(T_{\delta,s}(X), \delta - 1) \end{bmatrix} = \begin{bmatrix} N_{1-\delta} N_{1-\delta}^* \operatorname{diag}(X, 1 - \delta) \\ \vdots \\ N_{\delta-1} N_{\delta-1}^* \operatorname{diag}(X, \delta - 1) \end{bmatrix} = N N^* \mathcal{D}_\delta(X),$$

which completes the proof. □

[7] It is worth emphasizing that equality here, namely the case that $\operatorname{supp}(g_m^j) = [1, \delta - m]$, is not only feasible, but common.

6.2.4 Main Result

To prove a condition number result analogous to that of Theorem 7 for AN, we will need to show that the restriction operator N commutes well with the permutations used in the condition number analysis of Section 4.2, preserving the block-circulant structures that made the analysis possible. Thankfully it does; following the intuition of (4.19), referring to our expression of A in (6.7), and making use of Lemmas 11 and 21, we can arrive at

$$\begin{aligned} A' := P^{(\overline{d},D)} A \left(P^{(\overline{d},2\delta-1)} \otimes I_s\right)^* &= \mathrm{circ}^D \left(P^{(\overline{d},D)} \begin{bmatrix} R_{\overline{d}}\mathcal{T}_s g^1_{1-\delta} & \cdots & R_{\overline{d}}\mathcal{T}_s g^1_{\delta-1} \\ \vdots & \ddots & \vdots \\ R_{\overline{d}}\mathcal{T}_s g^D_{1-\delta} & \cdots & R_{\overline{d}}\mathcal{T}_s g^D_{\delta-1} \end{bmatrix} \right). \\ &= \mathrm{circ}^D \left(P^{(\overline{d},D)} (I_D \otimes R_{\overline{d}}) \begin{bmatrix} \mathcal{T}_s g^1_{1-\delta} & \cdots & \mathcal{T}_s g^1_{\delta-1} \\ \vdots & \ddots & \vdots \\ \mathcal{T}_s g^D_{1-\delta} & \cdots & \mathcal{T}_s g^D_{\delta-1} \end{bmatrix} \right). \end{aligned} \tag{6.12}$$

This may be reduced further by applying Lemma 22, which gives us that, setting

$$\begin{aligned} P_1 &= \mathcal{P}(P^{(2\delta-1,\overline{d})}, (s)_{j=1}^{\overline{d}(2\delta-1)}) = P^{(2\delta-1,\overline{d})} \otimes I_s \\ P_2 &= \mathcal{P}(P^{(2\delta-1,\overline{d})}, (\min\{s, \delta - |m|\})_{m=1-\delta}^{\delta-1}) \\ N' &= \mathrm{diag}(N_m)_{m=1-\delta}^{\delta-1}, \end{aligned}$$

we will have $N = \mathrm{diag}(I_{\overline{d}} \otimes N_m)_{m=1-\delta}^{\delta-1} = P_1(I_{\overline{d}} \otimes N')P_2^*$. This gives

$$A'(I_{\overline{d}} \otimes N') = P^{(\overline{d},D)} A P_1 (I_{\overline{d}} \otimes N') = P^{(\overline{d},D)} A N P_2. \tag{6.13}$$

Considering that the circ^D in (6.12) will repeat the inner matrix, which is of size $\overline{d}D \times s(2\delta - 1)$, $\overline{d}$ times, we have

$$A'(I_{\overline{d}} \otimes N') = \mathrm{circ}^D \left(P^{(\overline{d},D)} (I_D \otimes R_{\overline{d}}) \begin{bmatrix} \mathcal{T}_s g^1_{1-\delta} & \cdots & \mathcal{T}_s g^1_{\delta-1} \\ \vdots & \ddots & \vdots \\ \mathcal{T}_s g^D_{1-\delta} & \cdots & \mathcal{T}_s g^D_{\delta-1} \end{bmatrix} N' \right),$$

which, along with (6.13) and Corollary 5, gives us Theorem 10.

Theorem 10. *Taking A as in* (6.7)*, N and N_m as in* (6.8)*, and setting*

$$H = P^{(\overline{d},D)}(I_D \otimes R_{\overline{d}}) \begin{bmatrix} \mathcal{T}_s g^1_{1-\delta} & \cdots & \mathcal{T}_s g^1_{\delta-1} \\ \vdots & \ddots & \vdots \\ \mathcal{T}_s g^D_{1-\delta} & \cdots & \mathcal{T}_s g^D_{\delta-1} \end{bmatrix} \operatorname{diag}(N_m)_{m=1-\delta}^{\delta-1}$$

and $M_j = \sqrt{\overline{d}}(f_j^{\overline{d}} \otimes I_D)^* H$ for $j \in [\overline{d}]$, the condition number of AN is given by

$$\frac{\max\limits_{i\in[\overline{d}]} \sigma_{\max}(M_i)}{\min\limits_{i\in[\overline{d}]} \sigma_{\min}(M_i)}.$$

In particular, $\mathcal{A}|_{T_{\delta,s}(\mathbb{C}^{d\times d})}$ is invertible if and only if each of the M_i are of full rank.

6.3 Recovery Algorithm

In this section, we discuss a number of algorithms by which we can recover an estimate of x_0 from $T_{\delta,s}(\mathcal{A}^{-1}(y))$. We begin with Section 6.3.1, which discusses some improvements that can be made to the magnitude estimation step of Algorithm 1 (the $\sqrt{X_{j,j}}$ of line 4). These improvements, which we call "blockwise" vector or magnitude estimation, were first implemented in the numerical study of Section 3.6.1, and while they empirically delivered better results, no proof was yet discovered to guarantee by how much they improve the estimate of magnitudes. Section 6.3.2 describes and proves the robustness bounds for an algorithm almost completely analogous to Algorithm 1, taking advantage of the results in Section 6.3.1 for the magnitude estimation. Finally, in Section 6.3.3, we use blockwise vector estimation to devise a completely new recovery algorithm that is quite general with respect to the sets of shifts it permits. Although its recovery guarantees are not attractive, we will find in Section 6.4.3 that these theoretical results belie strong empirical performance.

6.3.1 Blockwise Magnitude and Vector Estimation

For the moment, we restrict our discussion to the special case of dense shifts in our measurements, where $s = 1$ as in the work of Chapters 3 and 4. In Section 3.6.1, it was noted that the technique used to calculate the magnitudes of the entries of $\underline{x}$ from $X \approx T_\delta(\underline{x}\underline{x}^*)$ was functional, but rudimentary. Recall from line 4 of Algorithm 1 that we have $X = \mathcal{A}^{-1}(\mathcal{A}(x_0 x_0^*)+n)$, where $\underline{x} \in \mathbb{C}^d$ is the ground truth objective vector, $\mathcal{A} : \mathbb{C}^{d\times d} \to \mathbb{R}^{dD}$ is the linear measurement operator determined by the masks $\{m_j\}_{j\in[D]}$, as defined, for example, in (4.1) of Section 4.1, and $n \in \mathbb{R}^{dD}$ is arbitrary noise. Then $X = \underline{X} + \mathcal{A}^{-1}(n)$, where $\underline{X} = T_\delta(\underline{x}\underline{x}^*)$, and we estimate the magnitude of $\underline{x}_i$ by simply taking $|x_i| = \sqrt{X_{ii}} \approx \sqrt{\underline{X}_{ii}} = |\underline{x}_i|$. This technique works, as we are able to show in Lemma 7, but empirically, we found that a slightly more sophisticated technique does a much better job here. While this comparison was not made explicit in Section 3.6, we briefly illustrate it in Section 6.4.3.

We notice that taking $|x_i| = X_{ii}$ is equivalent to taking $|x_i|$ to be the rank-1 approximation of the $1\times 1, i^{\text{th}}$ diagonal block matrix of X, namely $\left[X_{ii}\right]$, since these diagonal blocks are equal to the diagonal blocks of the *untruncated* $\underline{x}\underline{x}^*$ when there is no noise. However, given the width of the diagonal band in $T_\delta(\mathbb{C}^{d\times d})$, we could just as easily take blocks of size up to $\delta \times \delta$ and calculate their top eigenvectors; this would give us $2\delta - 1$ estimates for each entry's magnitude, so we can combine them by averaging them together. To denote these blocks, we will set $|X|^{(\ell)} = \operatorname{diag}(\mathbb{1}_{[\delta]_\ell})|X|\operatorname{diag}(\mathbb{1}_{[\delta]_\ell})$[8] and $u^{(\ell)}$ to be the top eigenvector of $|X|^{(\ell)}$, normalized such that $\|u^{(\ell)}\|_2 = \||X|^{(\ell)}\|_2$. We then produce our estimate of the magnitudes by taking $|x| = \frac{1}{\delta}\sum_{\ell=1}^{d} u^{(\ell)}$.

Before formally stating this algorithm, we observe a few ways in which it may be generalized. Firstly, we notice that this method can easily handle arbitrary block sizes m for the blockwise, eigenvector-based magnitude estimations by simply taking $|X|^{(\ell,m)} = \operatorname{diag}(\mathbb{1}_{[m]_\ell})|X|\operatorname{diag}(\mathbb{1}_{[m]_\ell})$, as long as $m \le \delta$ – although this would require us to change the denominator in the averaging step, using $\frac{1}{m}\sum_{\ell=1}^{d} u^{(\ell,m)}$. Proceeding even further, we can

[8]Here, and in the remainder of this section, we emphasize that all indices of objects in $\mathbb{C}^d$ and $\mathbb{C}^{d\times d}$ are taken modulo d.

generalize this technique to use any collection $\{J_i\}_{i=1}^N$, $J_i \subseteq [d]$ satisfying

$$\begin{aligned} [d] &\subseteq \bigcup_{i=1}^{N} J_i \\ \mathbb{1}_{J_i}\mathbb{1}_{J_i}^* &\in T_\delta(\mathbb{C}^{d\times d}) \end{aligned} \tag{6.14}$$

We will call any collection satisfying (6.14) a (T_δ, d)-*covering* (or just *covering* when T_δ and d are clear from context), and the process of estimating magnitudes of $\underline{x}$ from X with respect to a (T_δ, d)-covering is described in Algorithm 6.[9] It is worth remarking that the "averaging step," specified in line 3, is optimal in the least-squares sense. We set $P_{J_i} = \mathrm{diag}(\mathbb{1}_{J_i})$ to be the orthogonal projections onto the coordinate subspace associated with J_i and consider that the vectors $u^{(J_i)}$ (in line 2) represent estimates of the projections $P_{J_i}|\underline{x}|$. The least squares solution to $P_{J_i}u = u^{(J_i)}$, or

$$\begin{bmatrix} P_{J_1} \\ \vdots \\ P_{J_N} \end{bmatrix} u = \begin{bmatrix} u^{(J_1)} \\ \vdots \\ u^{(J_N)} \end{bmatrix}$$

is obtained by taking the pseudoinverse. In this case, we have

$$\begin{bmatrix} P_{J_1} \\ \vdots \\ P_{J_N} \end{bmatrix}^* \begin{bmatrix} P_{J_1} \\ \vdots \\ P_{J_N} \end{bmatrix} = \sum_{i=1}^{N} P_{J_i}^* P_{J_i} = \sum_{i=1}^{N} P_{J_i} = \mathrm{diag}(\mu),$$

with $\mu_j = |\{i : j \in J_i\}|$ as in line 1 of Algorithm 6. Considering that $P_{J_i}u^{(J_i)} = u^{(J_i)}$, we have

$$u = \begin{bmatrix} P_{J_1} \\ \vdots \\ P_{J_N} \end{bmatrix}^\dagger \begin{bmatrix} u^{(J_1)} \\ \vdots \\ u^{(J_N)} \end{bmatrix} = \mathrm{diag}(\mu)^{-1} \begin{bmatrix} P_{J_1} \\ \vdots \\ P_{J_N} \end{bmatrix}^* \begin{bmatrix} u^{(J_1)} \\ \vdots \\ u^{(J_N)} \end{bmatrix} = D_\mu^{-1} \left(\sum_{i=1}^{N} u^{(J_i)} \right).$$

We will denote the output of Algorithm 6 by $|x| = \mathrm{BlockMag}(X, \{J_i\})$. In an overloading of notation, when the covering consists of intervals of length m, in the sense that $J_i = [m]_i$

[9] We remark that this definition and the recovery algorithm are very obviously extensible to the use of $T_{\delta,s}$ instead of T_δ. In fact, this is a *restriction*, if we consider in (6.14) that $T_{\delta,s} \subseteq T_\delta$. The definition, therefore, of a $(T_{\delta,s}, d)$-covering, is made by analogy to (6.14).

Algorithm 6 Blockwise Magnitude Estimation

Input: $X \in T_{\delta,s}(\mathcal{H}^d)$, typically assumed to be an approximation $X \approx T_{\delta,s}(\underline{x}\underline{x}^*)$. A $(T_{\delta,s}, d)$-covering $\{J_i\}_{i\in[N]}$.

Output: An estimate $|x|$ of $|\underline{x}|$.

1: For $j \in [d]$, set $\mu_j = |\{i : j \in J_i\}|$ to be the number of appearences the index j makes in $\{J_i\}$.

2: For $i \in [N]$, set $u^{(J_i)}$ to be the leading eigenvector of $|X^{(J_i)}| = \operatorname{diag}(\mathbb{1}_{J_i})|X|\operatorname{diag}(\mathbb{1}_{J_i})$, normalized such that $\|u^{(J_i)}\|_2 = \sqrt{\|X^{(J_i)}\|_2}$.

3: Return $|x| = D_\mu^{-1}\left(\sum_{i=1}^N u^{(J_i)}\right)$.

for $i = 1, \ldots, d$, we will also write it as $\operatorname{BlockMag}(X, \{[m]_i\}_{i\in[d]}) = \operatorname{BlockMag}(X, m)$. To include the case of ptychography, where these intervals are shifted by more than 1, we write

$$\begin{aligned} \operatorname{BlockMag}(X, m) &= \operatorname{BlockMag}(X, \{[m]_i\}_{i\in[d]}), \text{ and} \\ \operatorname{BlockMag}(X, (m, s)) &= \operatorname{BlockMag}(X, \{[m]_{1+s(\ell-1)}\}_{\ell\in[\overline{d}]}), \end{aligned} \tag{6.15}$$

where we require $\overline{d} = \frac{d}{s}$ to be an integer. In this way, the magnitude estimation technique used in Section 3.6 is simply $|x| = \operatorname{BlockMag}(X, \delta)$, whereas the $|x_i| = \sqrt{X_{ii}}$ technique stated in Algorithm 1 is equivalent to $|x| = \operatorname{BlockMag}(X, 1)$.

Using Lemma 9, we are able to quickly prove a bound on the error of the estimate produced by this method.

Proposition 18. *Let $\{J_i\}_{i\in[N]}$ be a $(T_{\delta,s}, d)$-covering, and suppose $\underline{X} = T_{\delta,s}(\underline{x}\underline{x}^*)$ for some $\underline{x} \in \mathbb{C}^d$. Using the notation of Algorithm 6 (in particular, μ_j is as in line 1, and $\underline{u}^{(J_i)} = |\operatorname{diag}(\mathbb{1}_{J_i})\underline{x}|$), given $X \in T_{\delta,s}(\mathcal{H}^d)$, we have that the output $|x|$ satisfies*

$$\begin{aligned} \operatorname{BlockMag}(\underline{X}, \{J_i\}) &= |\underline{x}| \\ \|\operatorname{BlockMag}(X, \{J_i\}) - |\underline{x}|\|_2 &\le \frac{\max_j \mu_j}{\min_j \mu_j} \frac{1+2\sqrt{2}}{\min_i \|\underline{u}^{(J_i)}\|_2} \|\underline{X} - X\|_F. \end{aligned} \tag{6.16}$$

As special cases for $T_\delta(\mathbb{C}^{d\times d})$, we have

$$\begin{aligned}
\|\mathrm{BlockMag}(X,m)-\underline{x}\|_2 &\le \frac{1+2\sqrt{2}}{\min_i\|\underline{u}^{[m]_i}\|_2}\|X-\underline{X}\|_F\\
\|\mathrm{BlockMag}(X,\delta)-\underline{x}\|_2 &\le \frac{1+2\sqrt{2}}{\min_i\|\underline{u}^{[\delta]_i}\|_2}\|X-\underline{X}\|_F\\
\|\mathrm{BlockMag}(X,1)-\underline{x}\|_2 &\le \frac{\|\mathrm{diag}(X-\underline{X})\|_F}{\min_i|\underline{x}_i|}
\end{aligned} \tag{6.17}$$

Proof of Proposition 18. The first inequality of (6.16) is clear, since line 2 of Algorithm 6 will always return $\underline{u}^{(J_i)} = \mathbb{1}_{J_i}\circ|\underline{x}|$, so line 3 will give

$$\left(D_\mu^{-1}\left(\sum_{i=1}^N \underline{u}^{(J_i)}\right)\right)_j = \frac{1}{\mu_j}\sum_{i=1}^N |\underline{x}|_j \mathbb{1}_{j\in J_i} = |\underline{x}|_j.$$

The second comes by writing

$$\left\|D_\mu^{-1}(\sum_{i=1}^N u^{(J_i)}-\underline{u}^{(J_i)})\right\|_2^2 \le \left(\frac{1}{\min_j\mu_j}\right)^2\left\|\sum_{i=1}^N(u^{(J_i)}-\underline{u}^{(J_i)})\right\|_2^2. \tag{6.18}$$

From there, we consider that the j^{th} term in the summation

$$\left\|\sum_{i=1}^N(u^{(J_i)}-\underline{u}^{(J_i)})\right\|_2^2 = \sum_{j=1}^d\left(\sum_{i=1}^N u^{(J_i)}-\underline{u}^{(J_i)}\right)_j^2 \tag{6.19}$$

has at most $\max_k\mu_k$ nonzero summands, so, by $(\sum_{i=1}^n a_i)^2 \le n\sum_{i=1}^n a_i^2$, we have

$$\left\|\sum_{i=1}^N(u^{(J_i)}-\underline{u}^{(J_i)})\right\|_2^2 \le \max_j\mu_j\sum_{i=1}^N(u^{(J_i)}-\underline{u}^{(J_i)})^2. \tag{6.20}$$

We then apply Lemma 9[10] to get

$$\sum_{i=1}^N(u^{(J_i)}-\underline{u}^{(J_i)})^2 \le (1+2\sqrt{2})\frac{\|\underline{u}^{(J_i)}\underline{u}^{(J_i)*}-X^{(J_i)}\|_F^2}{\|\underline{u}^{(J_i)}\|_2^2} \tag{6.21}$$

[10]Here, we use the substitution $\eta\|\mathbf{x}_0\|_2 = \frac{\|X-X_0\|_F}{\|\mathbf{x}_0\|_2}$.

In this expression, we consider that, in the summation $\sum_{i=1}^{N}\|\underline{u}^{(J_i)}\underline{u}^{(J_i)*} - X^{(J_i)}\|_F^2$, the term $(\underline{X}_{ij} - X_{ij})^2$ appears at most $\max\{\mu_i, \mu_j\}$ times, such that $\sum_{i=1}^{N}\|\underline{u}^{(J_i)}\underline{u}^{(J_i)*} - X^{(J_i)}\|_F^2 \leq \max_j \mu_j \|\underline{X} - X\|_F^2$. Using this substitution, combining (6.18)–(6.21), and taking the square root of both sides gives

$$\left\| D_\mu^{-1}\left(\sum_{i=1}^{N} u^{(J_i)} - \underline{u}^{(J_i)}\right)\right\|_2 \leq \frac{\max_j \mu_j}{\min \mu_j}\frac{1+2\sqrt{2}}{\min_i \|\underline{u}^{(J_i)}\|_2}\|\underline{X} - X\|_F$$

as desired.

The first two inequalities of (6.17) are immediate by observing that $\mu_j = m$ when the covering is $\{[m]_i\}_{i\in[d]}$. The third comes from setting $\epsilon_i = X_{ii} - \underline{X}_{ii}$ and writing

$$\begin{aligned}
\||x| - |\underline{x}|\|_2^2 &= \sum_{i=1}^{d}\left(\sqrt{X_{ii}} - |\underline{x}|_i\right)^2 = \sum_{i=1}^{d}\left(\sqrt{|\underline{x}|_i^2 + \epsilon_i} - \sqrt{|\underline{x}|_i^2}\right)^2 \\
&= \sum_{i=1}^{d}\left(\frac{((|\underline{x}|_i^2 + \epsilon_i) - |\underline{x}|_i^2)^2}{\sqrt{|\underline{x}|_i^2 + \epsilon_i} + |\underline{x}|_i}\right)^2 \leq \sum_{i=1}^{d}\frac{\epsilon_i^2}{\min_i |\underline{x}|_i^2} \\
&= \frac{\|\mathrm{diag}(X - \underline{X})\|_F^2}{\min_i |\underline{x}|_i^2}
\end{aligned}$$

□

We end with a few remarks on the results of Proposition 18. One immediate benefit from Eq. (6.17) is that the estimation error from $\mathrm{BlockMag}(X, 1)$ no longer scales poorly with $d^{1/4}$ as in Lemma 7. The $\min_i |\underline{x}_i|$ factor in the denominator is also not a problem, since in the recovery results of, say, Theorem 5 or Corollary 8, the same term (to a higher power) already appears in the phase-error expression.

In the error bound for $\mathrm{BlockMag}(X, m)$ in Eq. (6.17), we notice that the bound is *strictly decreasing* with m, since, for $m_1 > m_2$, we have

$$\min_i \|\underline{x}^{[m_1]_i}\|_2^2 \geq (m_1 - m_2)\min_i |\underline{x}_i|^2 + \min_i \|\underline{x}^{[m_2]_i}\|_2^2 \geq m_1 \min_i |\underline{x}_i|^2.$$

Also, considering (6.16), it is clear that $\text{BlockMag}(X, \delta)$ gives the absolute best bound over all $\{J_i\}$, since any (T_δ, d)-covering $\{J_i\}_{i\in[N]}$ satisfies $J_i \subseteq [\delta]_\ell$ for some ℓ. This gives that $\min_i \|\underline{u}^{(J_i)}\|_2 \leq \min_i \|\underline{u}^{[\delta]_i}\|_2$, and obviously $\frac{\max_j \mu_j}{\min_j \mu_j} \geq 1$, so the bound for $\text{BlockMag}(X, \{J_i\})$ in (6.16) can never be better than that for $\text{BlockMag}(X, \delta)$ in (6.17).

We also remark that two easy ways to ensure $\frac{\max_j \mu_j}{\min_j \mu_j} = 1$ is minimized are to take some fixed $J_0 \subseteq [\delta]_0$ and let $J_\ell = J_0 + \ell$ be a "cyclic" covering, or to let $\{J_i\}$ be a partition of $[d]$. These strategies will be relevant in Section 6.3.2, but in the case of $s = 1$, $\text{BlockMag}(X, \delta)$ always has the optimal bound for magnitude estimation error.

6.3.2 Standard Recovery Algorithm for Ptychography

To recover x, an estimate of $\mathrm{e}^{\mathrm{i}\theta}\underline{x}$ from $X = \mathcal{A}^{-1}(y)$, we will not have to develop, nor even prove, any more technology. The contents of Chapter 5 and Sections 6.2.4 and 6.3.1 are sufficient to develop an algorithm, stated in Algorithm 7 that is proven in Theorem 11 to stably produce an estimate of $\mathrm{e}^{\mathrm{i}\theta}\underline{x}$. This algorithm differs very little from Algorithm 5, except that in line 4, we use $\text{BlockMag}(X, \{J_i\}) \circ \tilde{x}$ instead of $\sqrt{|\text{vec}\,\text{diag}(X)|} \circ \tilde{x}$.

Algorithm 7 Phase Retrieval from Local Ptychographic Measurements

Input: A family of masks $\{m_j\}_{j=1}^D$ of support δ; $s, d, \overline{d} \in \mathbb{N}$ satisfying $d = \overline{d}s \geq 2\delta - 1$. A $(T_{\delta,s}, d)$-covering $\{J_i\}_{i\in[N]}$. Measurements $\mathbf{y} \approx \mathcal{A}(\underline{x}\underline{x}^*) \in \mathbb{R}^{\overline{d}D}$, as in (6.2).

Output: $x \in \mathbb{C}^d$ with $x \approx \mathrm{e}^{\mathrm{i}\theta}\underline{x}$ for some $\theta \in [0, 2\pi]$.

1: Compute the matrix $X = \mathcal{A}|^{-1}_{T_{\delta,s}(\mathbb{C}^{d\times d})} y \in T_{\delta,s}(\mathcal{H}^d)$ as an estimate of $T_{\delta,s}(\underline{x}\underline{x}^*)$.

2: Form the banded matrix of phases, $\widetilde{X} \in T_{\delta,s}(\mathcal{H}^d)$, by normalizing the non-zero entries of X.

3: Compute $\hat{Z}$, the solution to (5.6) with $L = (2\delta - 1)I - \widetilde{X}$, and take $\tilde{x} = \text{sgn}(u)$, where u is the top eigenvector of $\hat{Z}$.

4: Return $x = \text{BlockMag}(X, \{J_i\}) \circ \tilde{x}$.

Theorem 11 proves a bound on the accuracy of the output of Algorithm 7. Here, we use the Laplacian matrix of the graph whose adjacency matrix is $T_{\delta,s}(\mathbb{1}\mathbb{1}^*) - I$, namely

$$D - W, \text{ where } W = T_{\delta,s}(\mathbb{1}\mathbb{1}^*) - I \text{ and } D = \text{diag}(W\mathbb{1}). \tag{6.22}$$

Theorem 11. *Suppose we have a family of masks $\{m_j\}_{j\in[D]} \in \mathbb{C}^d$ of support δ; $s, \overline{d} \in \mathbb{N}$ such that $s\overline{d} = d$; and a $(T_{\delta,s}, d)$-covering $\{J_i\}_{i\in[N]}$. Further let $\underline{x} \in \mathbb{C}^d, n \in \mathbb{R}^{\overline{d}D}$ be arbitrary and set $\underline{X} = \underline{x}\underline{x}^*, X = \mathcal{A}^{-1}(\mathcal{A}(\underline{X}) + n)$. Define μ as in line 1 of Algorithm 6, $\tau_G = \lambda_2(D - W)$ for the matrices defined in* (6.22)*, and* $\mathrm{SNR} = \frac{\|\mathcal{A}(X)\|_2}{\|n\|_2}$*. Then if line 3 of Algorithm 7 solves* (5.3) *exactly, then the output x of Algorithm 7 satisfies*

$$\begin{aligned}
\min_{\theta\in[0,2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 &\leq \frac{4\sqrt{2}}{\sqrt{\tau_G}} \frac{\sigma_{\min}^{-1}\|n\|_2}{|(\underline{x})_{\min}|^2} + \frac{\max_j \mu_j}{\min_j \mu_j} \frac{1+2\sqrt{2}}{\min_i \|\underline{u}^{(J_i)}\|_2} \sigma_{\min}^{-1}\|n\|_2 \\
\min_{\theta\in[0,2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 &\leq \frac{4\sqrt{2}}{\sqrt{\tau_G}} \frac{\kappa\|\underline{X}\|_F}{|(\underline{x})_{\min}|^2 \,\mathrm{SNR}} + \frac{\max_j \mu_j}{\min_j \mu_j} \frac{1+2\sqrt{2}}{\min_i \|\underline{u}^{(J_i)}\|_2} \kappa \frac{\|\underline{X}\|_F}{\mathrm{SNR}}
\end{aligned} \tag{6.23}$$

Proof of Theorem 11. This proof is a straightforward synthesis of the results in Theorem 9 and Proposition 18. In detail, we set $\tilde{x} = \mathrm{sgn}(x), \underline{\tilde{x}} = \mathrm{sgn}(\underline{x})$ and apply these two results by expanding

$$\begin{aligned}
\min_{\theta\in[0,2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta}\underline{x}\|_2 &\leq \min_{\theta\in[0,2\pi]} \|\tilde{x} - \mathrm{e}^{\mathrm{i}\theta}\underline{\tilde{x}}\|_2 + \||x| - |\underline{x}|\| \\
&\leq \frac{2\sqrt{2}\|\widetilde{X} - \underline{\widetilde{X}}\|_F}{\sqrt{\tau_G}} + \frac{\max_j \mu_j}{\min_j \mu_j} \frac{1+2\sqrt{2}}{\min_i \|\underline{u}^{(J_i)}\|_2} \|X - \underline{X}\|_F.
\end{aligned}$$

At this point, we quote Lemma 24 to get $\|\widetilde{X} - \underline{\widetilde{X}}\|_F \leq 2\|X - \underline{X}\|_F / |(\underline{x}_{\min})|^2$ as usual, and finish by bounding $\|X - \underline{X}\|_F$ above by $\sigma_{\min}^{-1}\|n\|_2$ and $\kappa\frac{\|\underline{X}\|_F}{\mathrm{SNR}}$. □

We remark that a few elements of the bounds in (6.23) are a bit unclear, most notably $\frac{\max_j \mu_j}{\min_j \mu_j}$ and τ_G. Of these, $\frac{\max_j \mu_j}{\min_j \mu_j}$ is a design feature that is chosen when we select our $(T_{\delta,s}, d)$-covering for the magnitude estimation step. As is mentioned in Section 6.3.1, one strategy for keeping this term under control would be making $\{J_i\}_{i\in[N]}$ a partition of $[d]$. However, using the somewhat intuitive (if we follow the motivation behind Section 6.3.1), maximal (it uses the largest possible J_i, which is an advantage for the $\min_i \|u^{(J_i)}\|_2$ term) covering $J_i = [1 + (i-1)s, \delta + 1 + (i-1)s)$ for $i \in [\overline{d}]$, we notice that $\frac{\max_j \mu_j}{\min_j \mu_j} \in [1, 2]$.[11]

However, τ_G is still mysterious – we have yet to obtain a convenient expression for

[11] The proof of this is brief: $j \in [1 + (\ell-1)s, \delta + 1 + (\ell-1)s)$ iff $\ell - 1 \in [\lceil\frac{j-\delta}{s}\rceil, \lfloor\frac{j-1}{s}\rfloor + 1)$, so that $\mu_j = \lfloor\frac{j-1}{s}\rfloor - \lceil\frac{j-\delta}{s}\rceil + 1$. Clearly μ_j is periodic in $[s]$, so we take extrema over this interval. Using $\lfloor\frac{j-1}{s}\rfloor = 0$ for $j \in [s]$ and $-\lceil\frac{j-\delta}{s}\rceil = \lfloor\frac{\delta-j}{s}\rfloor$, this immediately gives $\lfloor\frac{\delta}{s}\rfloor \leq \mu_j \leq \lfloor\frac{\delta-1}{s}\rfloor + 1 \leq \lfloor\frac{\delta}{s}\rfloor + 1$.

τ_G, so we relegate its study to the numerical experiments of Section 6.4.3.

6.3.3 Blockwise Vector Synchronization Method

There is another method for retrieval from $X = \mathcal{A}^{-1}(y)$, based largely on the ideas behind the Blockwise Magnitude Estimation of Section 6.3.1, that permits a very straightforward and satisfyingly general recovery method. The idea here is to use something analogous to Algorithm 6, but which recovers the magnitudes *and phases* of each block, and then stitches them together with the vector synchronization technique discussed in Appendix C. The blockwise, eigenvector-based step, called Blockwise Vector Estimation, is stated in Algorithm 8, and the complete recovery process is stated in Algorithm 9.

Unfortunately, the theory behind this strategy is so far fairly weak: we are able to prove a recovery bound on it, but it suffers from tremendously poor scaling with problem dimension. The proof is also very technical, so we defer it to Appendix C.2. For the moment, we pose Algorithm 9 as a theoretically unvetted, but intuitive and empirically promising technique, and defer a more optimistic view of its performance to the numerical studies of Section 6.4.3.

Algorithm 8 Blockwise Vector Estimation

Input: $X \in T_{\delta,s}(\mathcal{H}^d)$, typically assumed to be an approximation $X \approx T_{\delta,s}(\underline{x}\underline{x}^*)$. A connected $(T_{\delta,s}, d)$-covering $(\{J_i\}_{i\in[N]}, G)$.
Output: Estimates $x^{(J_i)}$ of $\operatorname{diag}(\mathbb{1}_{J_i})\underline{x}$.
1: For $i \in [N]$, set $x^{(J_i)}$ to be the leading eigenvector of $X^{(J_i)} = \operatorname{diag}(\mathbb{1}_{J_i}) X \operatorname{diag}(\mathbb{1}_{J_i})$, normalized such that $\|x^{(J_i)}\|_2 = \sqrt{\|X^{(J_i)}\|_2}$.
2: Return $\operatorname{BlockVec}(X, (\{J_i\}, G)) = \begin{bmatrix} x^{(J_1)} & \cdots & x^{(J_N)} \end{bmatrix}$.

Despite the pessimistic result of Proposition 24, we expect this method to perform fairly well. Recalling the discussion of Section 5.3.4 concerning the advantages of using the magnitudes of X in the angular synchronization process (namely, that this will prioritize accurate relative phases between larger entries of x), it seems like a potentially sensible step forward to let the blockwise magnitude estimation process of Algorithm 6 simultaneously produce an estimate of the phases of x. This process operates by finding the eigenvectors of

dense, approximately rank-1 matrices, and the eigenvector perturbation bound used in the proof of Proposition 18 works just as well for complex matrices. Synthesizing these blockwise estimates by calculating appropriate relative phases between them (e.g., by appealing to the vector synchronization process defined in Appendix C) and averaging as in line 3 of Algorithm 6 seems satisfyingly symmetric, but our lack of theoretical understanding of the performance of the vector synchronization process prevents us from achieving a stronger theoretical promise on this method.

Algorithm 9 Phase Retrieval by Vector Synchronization

Input: Measurements $y \approx \mathcal{A}(\underline{x}\underline{x}^*) \in \mathbb{R}^{\bar{d}D}$, where $\mathcal{A}$ is invertible on $T_{\delta,s}(\mathbb{C}^{d\times d})$ as in (6.2). A connected $(T_{\delta,s}, d)$-covering $(\{J_i\}_{i\in[N]}, G')$.

Output: $x \in \mathbb{C}^d$ with $x \approx \mathrm{e}^{\mathrm{i}\theta}\underline{x}$ for some $\theta \in [0, 2\pi]$.

1: For $j \in [d]$, set $\mu_j = |\{i : j \in J_i\}|$ to be the number of appearences the index j makes in $\{J_i\}$.
2: Set $V = \mathrm{BlockVec}(X, (\{J_i\}, G))$.
3: Compute $\hat{Z}$, the solution to (5.6) with L_{vs} as defined in (6.24) using V and G' and set $\hat{z} = \mathrm{sgn}(u)$, where u is the top eigenvector of $\hat{Z}$.
4: Return $x = D_\mu^{-1} V \hat{z}$.

One remark that must be made before stating the vector synchronization-based recovery algorithm is that having a simple $(T_{\delta,s}, d)$-covering will not suffice for this method; we require a slightly stronger set of conditions on the sets $J_i \subseteq [d]$, which will allow vector synchronization to be performed on their blockwise estimates of $\underline{x}$. Namely, we define a *connected* $(T_{\delta,s}, d)$*-covering* to be an ordered pair $(\{J_i\}, G')$ satisfying that $\{J_i\}$ is a $(T_{\delta,s}, d)$-covering and $G' = (V = [N], E)$ is a connected graph with

$$E \subseteq \{(i, j) : J_i \cap J_j \neq \emptyset\},$$

which will permit the blocks $x^{(J_i)}$ associated with each J_i to have a meaningful sense of relative phase over a connected graph. The synthesis process will be to take, for each $i \in [N]$, $x^{(J_i)}$ to be the top eigenvector of $X^{(J_i)} = \mathrm{diag}(\mathbb{1}_{J_i}) X \,\mathrm{diag}(\mathbb{1}_{J_i})^*$ normalized such that $\|x^{(J_i)}\|_2 = \sqrt{\|X^{(J_i)}\|_2}$ and to synchronize them by taking the solution $\hat{z}$ of (5.3) for

$$L_{\mathrm{vs}} = D_{\mathrm{vs}} - W_{\mathrm{vs}} \circ X_{\mathrm{vs}}, \tag{6.24}$$

defining $D_{\text{vs}}, W_{\text{vs}}, X_{\text{vs}}$ as in (C.3). With μ_j defined as in line 1 of Algorithm 6, the final estimate of $\underline{x}$ is rendered as $x = D_{\mu}^{-1} \begin{bmatrix} x^{(J_1)} & \cdots & x^{(J_n)} \end{bmatrix} \hat{z}$, which simply averages together the phase-synced $x^{(J_i)}\hat{z}_i$ terms. With this, we state Algorithms 8 and 9.[12]

6.4 Numerical Evaluation

We now numerically evaluate the new structures and results enumerated and studied in this chapter. In Section 6.4.1, we compare the blockwise magnitude estimation technique stated in Section 6.3.1 to the previous magnitude estimation strategy of reading the magnitudes of x directly from the main diagonal of $X \approx T_{\delta,s}(xx^*)$. Section 6.4.2 considers the condition number $\kappa(\mathcal{A}|_{T_{\delta,s}(\mathbb{C}^{d\times d})})$ of the measurement operator in the ptychographic case, as well as the spectral gap τ_G associated with the unweighted graph whose adjacency matrix is $T_{\delta,s}(\mathbb{1}_d\mathbb{1}_d^*) - I_d$. Finally, in Section 6.4.3, we consider the robustness of the whole process – the linear solve, which is conditioned according to κ, composed with the angular synchronization process, conditioned according to τ_G – with ptychographically shifted masks. This section also includes a comparison with the vector synchronization technique described in Section 6.3.3.

We note that, in Sections 6.4.2 and 6.4.3, we exclusively use masks m_j that are randomly generated. Unfortunately, despite the structure uncovered in Section 6.2.4, we have not yet discovered a way to import the notion of a local Fourier measurement system, which produced such a satisfyingly simple and physically realizable family of masks for which the condition number was predictably controlled, to the case of general shifts $s > 1$. As of this dissertation, the author is unaware of any deterministic construction $\{m_j\}_{j\in[s(2\delta-s)]}$ that scales with s, δ, and d to consistently produce invertible measurement operators $\mathcal{A}$.

[12] In Algorithm 9, notice that $\text{BlockVec}(X, (\{J_i\}, G))$ is the output of Algorithm 8.

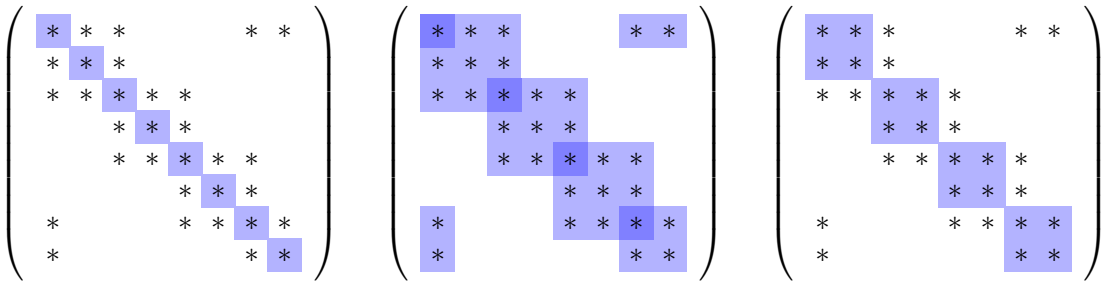
6.4.1 Numerical Study of Magnitude Estimation Techniques

The blockwise magnitude estimation technique of Section 6.3.1 was first introduced during the numerical analysis of Algorithm 1 presented in Section 3.6. This method had been invented as a mere intuition on how to slightly improve the numerical performance of Algorithm 1, and because it appeared to remove a few decibels of relative error from the reconstruction process as a whole, it was used in the evaluations. A theoretical analysis of this technique, however, was not developed until Section 6.3.1, when bounds on the relative error in the magnitudes were proven, and indeed the blockwise, eigenvector-based method was shown to have a sharper theoretical bound. Nonetheless, a direct numerical comparison between this strategy and the diagonal strategy was never made, so we present this comparison here.

Recall from the discussion of Section 6.3.1 that the blockwise magnitude estimation technique consists of applying Algorithm 6 to obtain $\mathrm{BlockMag}(X, \{J_i\}_{i\in[N]})$, where $J_i \subseteq [d]$ forms a $(T_{\delta,s}, d)$-covering, defined in Eq. (6.14). It was remarked that taking $|x|_i = \sqrt{|X_{ii}|}$, as in line 5 of Algorithm 1, was actually a special case of BlockMag, being equivalent to $\mathrm{BlockMag}(X, 1)$, using the notation of Eq. (6.15), while the "maximally dense" method described in Section 3.6.1 is equivalent to $\mathrm{BlockMag}(X, \delta)$. Therefore, a comparison of these two techniques is, in some sense, merely an evaluation of BlockMag over different parameters. We remark that, again according to (6.15), these techniques are trivially extensible to the ptychographic case of $s > 1$ by taking $\mathrm{BlockMag}(X, 1)$ and $\mathrm{BlockMag}(X, (\delta, s))$.

A third strategy, worthy of consideration, is briefly mentioned at the end of Section 6.3.1. In this case, we let $\{J_i\}$ be a partition of $[d]$. Of course, taking $J_i = [1]_i$, as in $\mathrm{BlockMag}(X, 1)$, is one example of such a partition, but in Figs. 6.4a–6.4c, the "Part'n" data points refer to calculating $\mathrm{BlockMag}(X, (s\lfloor\frac{\delta-1}{s}\rfloor, s))$, which is in some sense the "largest" partition. To get a sense of how $\mathrm{BlockMag}(X, 1)$, $\mathrm{BlockMag}(X, (\delta, s))$, and $\mathrm{BlockMag}(X, (s\lfloor\frac{\delta-1}{s}\rfloor, s\lfloor\frac{\delta-1}{s}\rfloor))$ scale with s, examples are illustrated in Fig. 6.3.

With this, we present the results of this numerical experiment in Fig. 6.4, where we



(a) BlockMag$(X, 1)$ (b) BlockMag$(X, (3, 2))$ (c) BlockMag$(X, (2, 2))$

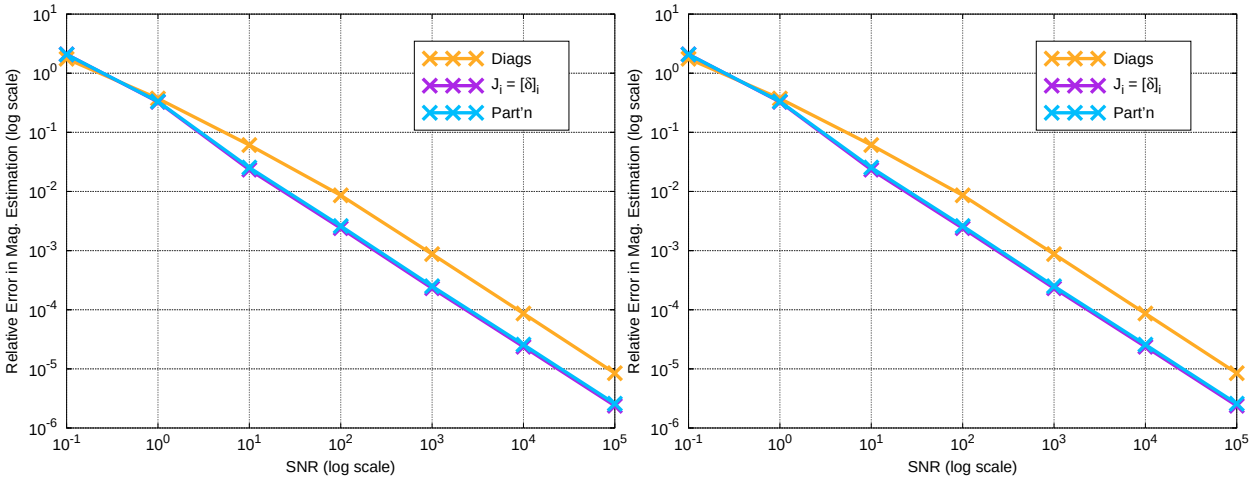
Figure 6.3: Blocks used for blockwise magnitude estimation in $T_{3,2}(\mathbb{C}^{8\times 8})$

have compared the relative error in magnitude estimation of these three techniques as a function of signal-to-noise ratio (SNR) for $s = 1, 3$, and 5, when $d = 60$ and $\delta = 6$. This gives us a comparison in the dense case (when $s = 1$), an overlap of 50% ($s = 3$), and a minimal overlap of 83.3% ($s = 5$).

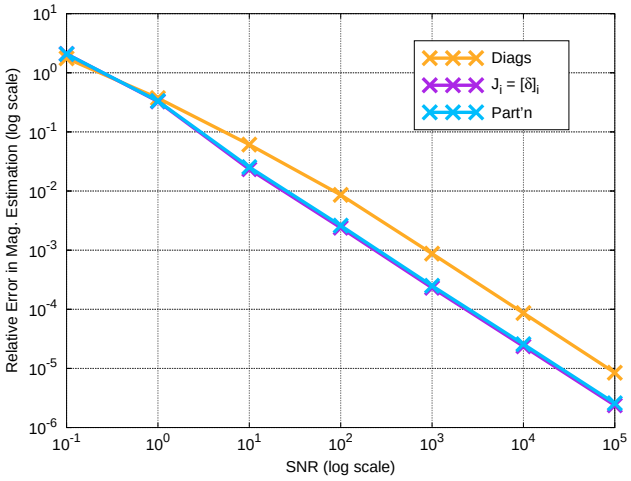
For each SNR, we randomly generated 128 objective vectors $\underline{x} \in \mathcal{CN}(0, I_d)$, and computed $\underline{X} = T_{\delta,s}(\underline{x}\underline{x}^*)$. We then randomly drew a Hermitian perturbation $N = T_{\delta,s}(N' + N'^*)$ where $N'_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, 1)$, and scaled it to ensure $\|\underline{X}\|_F / \|N\|_F = \text{SNR}$. We then wrote $X = \underline{X} + N$ and calculated $|x_{\text{diag}}| = \text{BlockMag}(X, 1), b\ |x_\delta| = \text{BlockMag}(X, (\delta, s))$, and $|x_{\text{part}}| = \text{BlockMag}(X, (s\lfloor\frac{\delta-1}{s}\rfloor, s\lfloor\frac{\delta-1}{s}\rfloor))$, along with the relative errors $\frac{\||x.|-|\underline{x}|\|_2}{\|\underline{x}\|_2}$ for each. These relative errors were averaged over the 128 trials and plotted against SNR for each technique in Fig. 6.4.

The results mostly affirm the intuition of Section 3.6.1 and the theory of Proposition 18: the larger block sizes show consistently stronger performance than the diagonal-only recovery method originally put forth. In fact, once the SNR becomes usable (at 10^1 or 10^2 – at SNR $= 10^{-1}$ or 10^0, the measurements are clearly useless), the relative error on the blockwise magnitude estimates is reduced by roughly a factor of five! Not too surprisingly, the partition method produces results comparable to those of the "full blocks" of BlockMag$(X, (\delta, s))$, since the block sizes differ by at most $\min\{s, \delta - s\}$, and the gain that x_{part} experiences from using off-diagonal entries to inform its magnitude estimates is on the same order as the gain experienced by x_δ.

On the other hand, one observation that is *not* expected is that the performance of



(a) $s = 1$ (b) $s = 3$



(c) $s = 5$

Figure 6.4: Relative error in magnitude estimation vs. SNR. $d = 60, \delta = 6, s \in \{1, 3, 5\}$

none of these methods appears to deteriorate at all with increasing s. This may be explained by supposing that it is the actual *eigenvector strategy*, rather than averaging over several estimates, that gives this method its strength. indeed, this intuition would also somewhat explain the comparability between x_{part} and x_δ, since the block sizes for each are roughly equal, and x_δ's main advantage is that each magnitude arises from an average of several estimates, rather than a single estimate as in the case of a partition. However, none of these intuitions are formalized any further at the moment.

6.4.2 Conditioning and Spectral Gap

In this section, we consider the measures we have that gauge the degree to which the recovery methods of Section 6.3.3, especially Algorithm 7, magnify noise in the measurements $\mathcal{A}(\underline{x}\underline{x}^*) + n$. Specifically, we are going to numerically evaluate κ, the condition number of the linear operator $\mathcal{A}$ defined in (6.2), and τ_G, the spectral gap of the graph that will appear in the angular synchronization phase (specifically, line 3 of Algorithm 7).

As already mentioned, there are so far no known constructions of masks $\{m_j\}_{j\in[D]}$ that have theoretical conditioning bounds (or even a theoretical guarantee of invertability), so Fig. 6.5 describes the distribution of $\kappa(\mathcal{A})$ over masks given by $m_j \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0, I_\delta), j \in [s(2\delta - s)]$, much like Fig. 4.8 in Section 4.5.3. In this experiment, we have fixed $d = 60$ and $\delta = 13$, but we vary s so that we can see whether and how much the level of overlap affects the distribution of condition numbers. For each value of s, we have drawn 1024 sets of masks and calculated the condition number κ of each set. Note that, to compare the distributions of κ between values of s, in lieu of presenting a histogram with four overlaid sets of bins, we have opted for a line plot which traces the heights that the bins would otherwise take.

Seen side by side, the distributions appear roughly equal – they each have their peak between 10^3 and 10^4 and possess relatively little probability mass after 10^5. It is difficult to comment on whether these results are predictable, however: considering the expressions obtained for A in (6.12) and its condition number in Theorem 10, where g_m^j is a vector of products of Gaussian random variables, and the g_m^j are dependent across m, there is

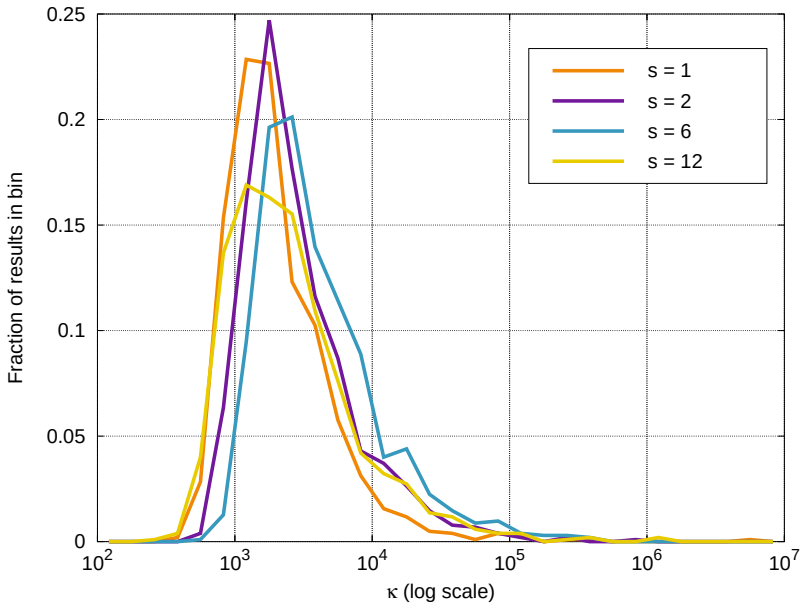


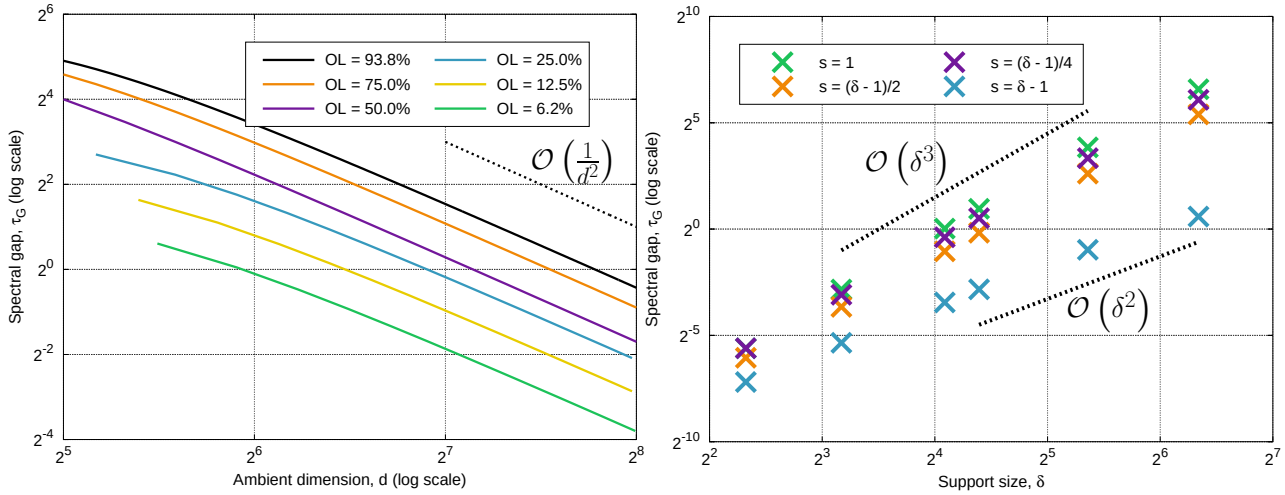
Figure 6.5: Distribution of condition numbers. $d = 60, \delta = 13$.

no obvious intuition for the singular values of A. Nonetheless, it is reassuring to see that the distribution of κ does not radically deteriorate when we introduce shifts $s > 1$, and somewhat interesting to see that, after slightly worsening when s goes through 2 and 6, at $s = 12$ (the maxiumum allowed shift for $\delta = 13$), the distribution comes back to a peak more commensurate with that of $s = 1$, although it does come with a thicker tail.

Figure 6.6 is a very simple display of the spectral gap τ_G of the graphs associated with the subspaces $T_{\delta,s}$ for different values of s. To be precise, we define $G^d_{\delta,s} = (V^d_{\delta,s} = [d], E^d_{\delta,s})$ to be the graph on d vertices such that the adjacency graph of $G^d_{\delta,s}$ is $A_G(d, \delta, s) := T_{\delta,s}(\mathbb{1}_d \mathbb{1}_d^*) - I_d$. For simplicity of notation, we will refer to $\tau_{G^d_{\delta,s}}$ as $\tau_G(d, \delta, s)$, or τ_G when the arguments are superfluous. Recall from Sections 3.4 and 5.3 that the spectral gap of this graph is an important component of the error bound that is proven for any of the angular synchronization techniques studied in this dissertation – the bound is inversely proportional to τ_G for the faster, eigenvector-based recovery method, and to $\sqrt{\tau_G}$ for the SDP recovery, so we want to have a spectral bound that is as large as possible.

However, τ_G is, in some sense, a measure of "how connected" the graph G is, and is in fact often called the *algebraic connectivity* of a graph in the literature [28, 38]. Indeed, both of

these works remark that τ_G is monotonically decreasing as edges are removed from a graph, so the introduction of larger shifts $s > 1$ can only serve to make the error bounds of Theorems 4 and 9 worse. Figure 6.6 numerically illustrates the dependence of $\tau_G(d, \delta, s)$ on each of its variables. For these charts, we constructed the Laplacian $L(d, \delta, s) = \text{diag}(A_G(d, \delta, s)\mathbb{1}_d) - A_G(d, \delta, s)$ for each (d, δ, s) and simply calculated the smallest eigenvalue of PLP^*, where P is an orthogonal projection onto $\mathbb{1}_d^{\perp} = \text{Col}(L(d, \delta, s))$. Notice, in each case, that the choices of parameters (d, δ, s) may seem somewhat erratic; this is simply because it is required that $\overline{d} = d/s$ be an integer. In any case, the data points we have produced are sufficient to illustrate the most pressing trends.



(a) Spectral gap of $G^d_{\delta,s}$ vs. d. $\delta = 16$. $s \in \{1, 4, 8, 12, 14, 15\}$.

(b) Spectral gap of $G^d_{\delta,s}$ vs. δ. $d = 240$. Overlap $\approx 100\%, 50\%, 25\%, \frac{1}{\delta}$.

Figure 6.6: Dependence of τ_G on d, δ, and s.

These results are qualitatively similar to those for the condition number: the behavior deteriorates in the direction we expect, but not catastrophically. In Fig. 6.6a, for example, we can see that, fixing δ and changing d, τ_G asymptotically behaves like $\mathcal{O}(1/d^2)$ for each level of overlap; recalling from Lemma 2 that, for $s = 1$, we had $\tau_G \approx \mathcal{O}(\delta^3/d^2)$, this is a reasonable observation. Nicely enough, as the shifts s get larger, the spectral gap doesn't appear to shrink too horrendously: between $s = 1$ and $s = \delta - 1 = 15$, the minumum and maximum possible shifts, τ_G only decreases by roughly a factor of 16.

Interestingly, on the other hand, s appears to play a larger role in shaping how τ_G

varies with δ. Figure 6.6b demonstrates this dependence, where we have now fixed $d = 240$, and we see that the relationship $\tau_G \approx \mathcal{O}(\delta^3)$ does *not* hold over different values of s. For $s = \delta - 1$, the maximal shift (and minimal overlap), τ_G doesn't even appear to level out at $\mathcal{O}(\delta^2)$. From a rough reading of Fig. 6.6b, we see that going from $s = 1$ to $s = \delta - 1$ at $\delta = 16$ costs about a factor of 2^4 in τ_G, whereas the same operation – full overlap to minimum overlap – at $\delta = 81$ costs a factor of 2^6. This increased "spectral cost" may be accounted for by considering that, with higher values of δ, the difference between the graphs $G^d_{\delta,1}$ and $G^d_{\delta,\delta-1}$ is a far greater number of edges, so the graph's "connectivity" may be expected to suffer more. For the present moment, we rest with these observations and defer a formalization of these results to future study.

6.4.3 Ptychography, Full System

Awaiting analysis of Section 5.4.

Chapter 7

Phase Retrieval from Local Measurements in Two Dimensions

In this chapter, we extend the model and methods of Chapter 3 to the case of a two dimensional signal $Q \in \mathbb{C}^{d \times d}$.

7.1 Introduction

In Chapter 3, we studied the phase retrieval problem (3.1) in the case of a measurement model that we referred to as *local measurement systems*. This meant we had a small number of vector $m_j \in \mathbb{C}^d$, whose magnitude-squared inner products were taken with the objective vector $x \in \mathbb{C}^d$, for several different circular translations of m_j, written explicitly as

$$(\mathbf{y}_\ell)_j = |\langle \mathbf{x}_0, S_\ell^* \mathbf{m}_j \rangle|^2, \quad (j, \ell) \in [K] \times P$$

in (3.2). Of course, considering this merely as a mathematical abstraction, it is entirely conceivable that $x \in \mathbb{C}^d$ represents a discretization and vectorization of a signal in two-dimensions, or for that matter, any arbitrary number of dimensions – the mathematical formulation does not force or exclude any particular interpretation of the model. Nonetheless,

as is mentioned in Section 3.1.4, one major strength of the model proposed and studied in this dissertation is that it somewhat more closely represents the reality of ptychography, the crucial application of phase retrieval that has motivated this work since the beginning. Therefore, if we wish to extend this work to two dimensions, we ought to analyze an analogous model that adequately reflects the structure of two-dimensional ptychography.

Conveniently enough, it is not difficult to state the model for what we have in mind: we shall consider measurements of the form

$$\left|\left\langle Q, S^{\ell} A S^{-\ell'} \right\rangle\right|^2 = \left|\sum_{j,k=1}^{d} A_{j-\ell,k-\ell'} Q_{jk}\right|^2, \tag{7.1}$$

where $Q \in \mathbb{C}^{d\times d}$ is the two-dimensional objective signal that we want to recover (analogous to x in (1.1)), $A \in \mathbb{C}^{d\times d}$ is a measurement matrix (analogous to one of the m_j), and $\ell, \ell' \in [d]_0$ are shifts (analogous to ℓ). For A to represent a "local measurement" (with a support size of δ), we require our measurement matrices to satisfy $\operatorname{supp}(A) \subseteq [\delta]^2$, and for simplicity of analysis, we are going to further require that these matrices are rank one; we will therefore index our measurement matrices by $A_{uv} = \mathbf{a}_u \mathbf{b}_v^*$, where $\operatorname{supp}(\mathbf{a}_u), \operatorname{supp}(\mathbf{b}_v) \subseteq [\delta]$. For convenience, we set

$$\mathbf{Q} := \operatorname{vec} Q \operatorname{vec} Q^*. \tag{7.2}$$

We remark that this setup may be used to model two-dimensional ptychography, in a way similar to the work in Section 3.1.4, by fixing $\mathbf{a}, \mathbf{b} \in \mathbb{C}^d$ (with $\operatorname{supp}(\mathbf{a}), \operatorname{supp}(\mathbf{b}) \subseteq [\delta]$) and setting

$$\mathbf{a}_u = \sqrt{K} f_u^K \circ \mathbf{a}, \ \ \mathbf{b}_v = \sqrt{K} f_v^K \circ \mathbf{b}.$$

7.2 An Efficient Method for Solving the Discrete 2D Phase Retrieval Problem

Our recovery method, outlined in Algorithm 10, aims to approximate an image $Q \in \mathbb{C}^{d\times d}$ from phaseless measurements of the form (7.1). Specifically, we consider the collection

Algorithm 10 Two Dimensional Phase Retrieval from Local Measurements

Input: Measurements $\mathbf{y} = \mathcal{M}(\mathbf{Q}) + n \in \mathbb{R}^{d^2D^2}$ and masks $\mathbf{a}_u, \mathbf{b}_v$ as per (7.3)
Output: $X \in \mathbb{C}^{d\times d}$ with $X \approx \mathrm{e}^{\mathrm{i}\theta}Q$ for some $\theta \in [0, 2\pi]$

1: Compute the Hermitian matrix $P = \left(\left(\mathcal{M}\big|_{\mathcal{P}}\right)^{-1}\mathbf{y}\right)/2 + \left(\left(\mathcal{M}\big|_{\mathcal{P}}\right)^{-1}\mathbf{y}\right)^*/2 \in \mathcal{P}\left(\mathbb{C}^{d^2\times d^2}\right)$ as an estimate of $\mathcal{P}(\mathbf{Q})$. $\mathcal{M}$ and $\mathcal{P}$ are as defined in (7.4) and §7.2.1. $\mathbf{Q}$ is as defined in (7.2).
2: Form the matrix of phases, $\widetilde{P} \in \mathcal{P}\left(\mathbb{C}^{d^2\times d^2}\right)$, by normalizing the non-zero entries of P.
3: Compute the principal eigenvector of $\widetilde{P}$ and use it to compute $U_{j,k} \approx \mathrm{sgn}(Q_{j,k})\ \forall j,k \in [d]$ as per §7.2.2.
4: Use the diagonal entries of P to compute $M_{j,k} \approx |Q_{j,k}|^2$ for all $j,k \in [d]$ as per §7.2.3.
5: Set $X_{j,k} = \sqrt{M_{j,k}} \cdot U_{j,k}$ for all $j,k \in [d]$ to form X

of measurements given by

$$y_{(\ell,\ell',u,v)} := \left|\langle Q, S_\ell \mathbf{a}_u \left(S_{\ell'}\mathbf{b}_v\right)^*\rangle\right|^2 \tag{7.3}$$

for all $(\ell,\ell',u,v) \in [d]^2 \times \Omega^2$ where $\Omega \subseteq [d]$ has $|\Omega| = 2\delta - 1$. Thus, we collect a total of $(2\delta-1)^2 \cdot d^2$ measurements where each measurement is due to a vertical and horizontal shift of a rank one illumination pattern $\mathbf{a}_u\mathbf{b}_v^* \in \mathbb{C}^{d\times d}$.

Algorithm 10 consists of first rephrasing the system (7.3) as a linear system on the space of $d^2 \times d^2$ matrices (following Candes, et al. [31]), and then estimating a projection $\mathcal{P}(\mathbf{Q})$ of the rank one matrix $\mathbf{Q}$ from this system. This process is described in Section 7.2.1. In Sections 7.2.2 and 7.2.3 we show how the magnitudes of the entries of Q are estimated directly from $\mathcal{P}(\mathbf{Q})$ and their phases are found from solving an eigenvector problem. Together, the magnitude and phase estimates provide an approximation of Q.

7.2.1 The Linear Measurement Operator $\mathcal{M}$ and Its Inverse

To produce the linear system of step 1, we observe that

$$\begin{aligned} y_{(\ell,\ell',u,v)} &= \left|\langle Q, S_\ell \mathbf{a}_u \left(S_{\ell'} \mathbf{b}_v\right)^* \rangle\right|^2 = \left|\langle \operatorname{vec} Q, S_{\ell'} \bar{\mathbf{b}_u} \otimes S_\ell \mathbf{a}_v \rangle\right|^2 \\ &= \left\langle \mathbf{Q},\ S_{\ell'} \bar{\mathbf{b}_u} \otimes S_\ell \mathbf{a}_v \left(S_{\ell'} \bar{\mathbf{b}_u} \otimes S_\ell \mathbf{a}_v\right)^* \right\rangle, \end{aligned}$$

which allows us to naturally define $\mathcal{M} : \mathbb{C}^{d^2 \times d^2} \mapsto \mathbb{R}^{[d]_0^2 \times [D]^2}$ as the linear measurement operator given by

$$\begin{aligned} \left(\mathcal{M}(Z)\right)_{(\ell,\ell',u,v)} &:= \left\langle Z,\ S_{\ell'} \bar{\mathbf{b}_u} \otimes S_\ell \mathbf{a}_v \left(S_{\ell'} \bar{\mathbf{b}_u} \otimes S_\ell \mathbf{a}_v\right)^* \right\rangle \\ &= \left\langle Z,\ S_{\ell'} \bar{\mathbf{b}_u} \bar{\mathbf{b}_u}^* S_{\ell'}^* \otimes S_\ell \mathbf{a}_v \mathbf{a}_v^* S_\ell^* \right\rangle, \end{aligned} \tag{7.4}$$

so that $\mathbf{y} = \mathcal{M}(\mathbf{Q})$. This allows us to solve for $\mathcal{P}(\mathbf{Q})$, the projection of $\mathbf{Q}$ onto the rowspace $\mathcal{P}(\mathbb{C}^{d^2 \times d^2})$ of $\mathcal{M}$. For clarity, we will abbreviate $\mathcal{P}(\mathbb{C}^{d^2 \times d^2})$ as $\mathcal{P}$, identifying this subspace with its orthogonal projection operator.

We observe that the local supports of $\mathbf{a}_u$ and $\mathbf{b}_v$ ensure that $\mathcal{M}(\operatorname{vec} E_{j,k} \operatorname{vec} E_{j',k'}{}^*) = \mathbf{0}$ whenever either $|j - j'| \geq \delta$ or $|k - k'| \geq \delta$ holds (this is clear from (7.4) and (4.15)). As a result we can see that $\mathcal{P} \subseteq \mathcal{B}$ where

$$\mathcal{B} := \operatorname{span}\{\operatorname{vec} E_{j,k} \operatorname{vec} E_{j',k'}{}^* \ \big|\ |j - j'| < \delta, |k - k'| < \delta\}. \tag{7.5}$$

In steps 2-4 of algorithm 10, recovery of Q from $\mathcal{P}(\mathbf{Q})$ relies on having $\mathcal{P} = \mathcal{B}$ exactly; we say in such a case that $\mathcal{M}|_{\mathcal{B}}$ is invertible. Clearly, the invertibility of $\mathcal{M}$ over $\mathcal{B}$ will depend on our choice of $\mathbf{a}$ and $\mathbf{b}$. We prove the following proposition, a corollary of which identifies pairs $\mathbf{a}, \mathbf{b}$ which produce an invertible linear system:

Proposition 19. *Let $T_\delta : \mathbb{C}^{d \times d} \to \mathbb{C}^{d \times d}$ be the operator given by*

$$T_\delta(X)_{ij} = \begin{cases} X_{ij}, & |i - j| < \delta \bmod d \\ 0, & \textit{otherwise} \end{cases}.$$

If the space $T_\delta(\mathbb{C}^{d\times d})$ is spanned by the collection $\{a_j a_j^*\}_{j=1}^K$, then $\mathcal{B}$ is spanned by

$$\{(a_j \otimes a_{j'})(a_j \otimes a_{j'})^*\}_{(j,j')\in[K]^2} = \{(a_j a_j^*) \otimes (a_{j'} a_{j'}^*)\}_{(j,j')\in[K]^2}.$$

Proof. By (4.15) and (7.5), it suffices to show that

$$(e_k e_{k'}^*) \otimes (e_j e_{j'}^*) \in \operatorname{span}\{(a_n a_n^*) \otimes (a_{n'} a_{n'}^*)\}_{(n,n')\in[K]^2}$$

for any $|j-j'|, |k-k'| < \delta$. Indeed, we have that $\{E_{jj'} : |j-j'| < \delta \bmod d\}$ forms a basis for $T_\delta(\mathbb{C}^{d\times d})$, so $E_{jj'}, E_{kk'} \in \operatorname{span}\{a_n a_n^*\}_{n\in[K]}$ and

$$(e_k e_{k'}^*) \otimes (e_j e_{j'}^*) \in \operatorname{span}\{(a_n a_n^*) \otimes (a_{n'} a_{n'}^*)\}_{(n,n')\in[K]^2}.$$

□

We remark that Proposition 19 permits us to import all of the theory developed in Chapter 4, as well as the mask constructions offered in Section 3.2. A trivial combination of Proposition 19 with Corollary 4 gives us Corollary 10, and with Example 2 of Section 3.2 gives us Corollary 11.

Corollary 10. *Fix $d, \delta \in \mathbb{N}$ such that $D := 2\delta - 1 \le d$. Suppose that $\gamma \in \mathbb{R}^d$ has $\operatorname{supp}(\gamma) \subseteq [\delta]$ and satisfies $f_j^d(\gamma \circ S^{-m}\gamma) \neq 0$ for all $j \in [d], m \in [\delta]_0$. Then fixing $K \ge D$ and setting $\mathbf{a}_u = f_u^K \circ \gamma$, we have*

$$\operatorname{span}\{S^\ell \mathbf{a}_u \mathbf{a}_u^* S^{-\ell} \otimes S^{\ell'} \mathbf{a}_v \mathbf{a}_v^* S^{-\ell'}\}_{(\ell,\ell,u,v)\in[d]^2\times[D]^2} = \mathcal{B}.$$

Corollary 11. *Fix $d, \delta \in \mathbb{N}$ such that $D := 2\delta - 1 \le d$. Set $\mathbf{a}_1 = e_1, \mathbf{a}_{2j} = e_1 + e_{j+1}$, and $\mathbf{a}_{2j+1} = e_1 + ie_{j+1}$ for $j \in [\delta - 1]$. Then*

$$\operatorname{span}\{S^\ell \mathbf{a}_u \mathbf{a}_u^* S^{-\ell} \otimes S^{\ell'} \mathbf{a}_v \mathbf{a}_v^* S^{-\ell'}\}_{(\ell,\ell,u,v)\in[d]^2\times[D]^2} = \mathcal{B}.$$

7.2.2 Computing the Phases of the Entries of Q after Inverting $\mathcal{M}\big|_{\mathcal{B}}$

Assuming that $\mathcal{P} = \mathcal{B}$ so that we can recover $\mathcal{P}(\mathbf{Q}) = \mathcal{B}(\mathbf{Q})$ from our measurements $\mathbf{y}$, we are still left with the problem of how to recover $\operatorname{vec} Q$ from $\mathcal{B}(\mathbf{Q})$. Our first step in solving for $\operatorname{vec} Q$ will be to compute all the phases of the entries of $\operatorname{vec} Q$ from $\mathcal{B}(\mathbf{Q})$. Thankfully, this can be solved as an angular synchronization problem as in Section 3.4 by using the method described in Theorem 4. Let $\mathbb{1} \in \mathbb{C}^{d^2 \times d^2}$ be the matrix of all ones, and $\operatorname{sgn} : \mathbb{C} \mapsto \mathbb{C}$ be

$$\operatorname{sgn}(z) = \begin{cases} \dfrac{z}{|z|}, & z \neq 0 \\ 1, & \text{otherwise} \end{cases}.$$

We now define $\widetilde{Q} \in \mathbb{C}^{d^2 \times d^2}$ by $\widetilde{Q} = \mathcal{B}(\operatorname{sgn}(\mathbf{Q}))$ (i.e. $\widetilde{Q}$ is $\mathcal{B}(\mathbf{Q})$ with its non-zero entries normalized). As we shall see, the principal eigenvector of $\widetilde{Q}$ will provide us with all of the phases of the entries of $\operatorname{vec} Q$.

Working toward that goal, we may note that

$$\widetilde{Q} = \operatorname{diag}\left(\operatorname{sgn}\left(\operatorname{vec} Q\right)\right) \mathcal{B}\left(\mathbb{1}\mathbb{1}^*\right) \operatorname{diag}\left(\overline{\operatorname{sgn}\left(\operatorname{vec} Q\right)}\right) \tag{7.6}$$

where sgn is applied component-wise to vectors, and where $\operatorname{diag}(\mathbf{x}) \in \mathbb{C}^{d^2 \times d^2}$ is diagonal with $\left(\operatorname{diag}(\mathbf{x})\right)_{j,j} := x_j$ for all $\mathbf{x} \in \mathbb{C}^{d^2}$ and $j \in [d^2]$. After noting that $\operatorname{diag}(\operatorname{sgn}(\cdot))$ always produces a unitary diagonal matrix, we can further see that the spectral structure of $\widetilde{Q}$ is determined by $\mathcal{B}\left(\mathbb{1}\mathbb{1}^*\right)$. In particular, the following theorem completely characterizes the eigenvalues and eigenvectors of $\mathcal{B}\left(\mathbb{1}\mathbb{1}^*\right)$.

Theorem 12. *Let $F \in \mathbb{C}^{d \times d}$ be the unitary discrete Fourier transform matrix with $F_{j,k} := \frac{1}{\sqrt{d}} \mathrm{e}^{2\pi \mathrm{i} \frac{(j-1)(k-1)}{d}}$ $\forall j, k \in [d]$, and let $\Lambda \in \mathbb{C}^{d \times d}$ be the diagonal matrix with $\Lambda_{j,j} = 1 + 2\sum_{k=1}^{\delta-1} \cos\left(\frac{2\pi(j-1)k}{d}\right)$ $\forall j \in [d]$. Then,*

$$\mathcal{B}\left(\mathbb{1}\mathbb{1}^*\right) = \left(F \otimes F\right)\left(\Lambda \otimes \Lambda\right)\left(F \otimes F\right)^*.$$

In particular, the principal eigenvector of $\mathcal{B}(\mathbb{1}\mathbb{1}^*)$ *is* $\mathbb{1}$ *and its associated eigenvector is* $(2\delta - 1)^2$.

Proof of Theorem 12. From the definition of $\mathcal{B}$ we have that

$$\begin{aligned}
\mathcal{B}(\mathbb{1}\mathbb{1}^*) &= \sum_{j=1}^{d} \sum_{|j-j'|<\delta} \sum_{k=1}^{d} \sum_{|k-k'|<\delta} \operatorname{vec} E_{j,k} \left(\operatorname{vec} E_{j',k'}\right)^* \\
&= \sum_{j=1}^{d} \sum_{|j-j'|<\delta} \sum_{k=1}^{d} \sum_{|k-k'|<\delta} E_{kk'} \otimes E_{jj'} \\
&= \left(\sum_{k=1}^{d} \sum_{|k-k'|<\delta} E_{kk'} \right) \otimes \left(\sum_{j=1}^{d} \sum_{|j-j'|<\delta} E_{jj'} \right) \\
&= T_\delta(\mathbb{1}\mathbb{1}^*) \otimes T_\delta(\mathbb{1}\mathbb{1}^*)
\end{aligned}$$

Of course, the eigenvectors and eigenvalues of $T_\delta(\mathbb{1}\mathbb{1}^*)$ have already been studied in Lemma 1, so $T_\delta(\mathbb{1}\mathbb{1}^*) = F\Lambda F^*$ (with Λ and F as defined in the statement of this theorem) which then yields the desired result by Theorem 4.2.12 of Horn and Johnson [54]. □

Theorem 12 in combination with (7.6) makes it clear that $\operatorname{sgn}(\operatorname{vec} Q)$ will be the principal eigenvector of $\widetilde{Q}$. As a result, we can rapidly compute the phases of all the entries of $\operatorname{vec} Q$ by using, e.g., a shifted inverse power method [94] in order to compute the eigenvector of $\widetilde{Q}$ corresponding to the eigenvalue $(2\delta - 1)^2$.

7.2.3 Computing the Magnitudes of the Entries of Q after Inverting $\mathcal{M}\big|_{\mathcal{B}}$

Having found the phases of each entry of $\operatorname{vec} Q$ using $\mathcal{B}(\mathbf{Q})$, it only remains to find each entry's magnitude as well. This is comparably easy to achieve. Note that $\mathcal{B}$ trivially contains $\operatorname{vec} E_{j,k} \operatorname{vec} E_{j,k}{}^* = e_k e_k^* \otimes e_j e_j^*$ for all $j, k \in [d]$, so $\mathcal{B}(\mathbf{Q})$ is guaranteed to always provide the diagonal entries of $\mathbf{Q}$,which are exactly the squared magnitudes of the entries of $\operatorname{vec} Q$. Combined with the phase information recovered in step 2 of 10, we are finally able

to reconstruct every entry of $\operatorname{vec} Q$ up to a global phase as in step 5.

7.3 Numerical Evaluation

We will now demonstrate the efficiency and robustness of Algorithm 10. Figs. 7.1b and 7.1d plot the reconstruction of two 256×256 test images (shown in Figs. 7.1a and 7.1c respectively) from squared magnitude local measurements of the type described in (7.3). Here $\mathbf{a}$ and $\mathbf{b}$ are chosen to be the *deterministic* measurement vectors of Corollary 11 with δ chosen to be 2. The reconstructions were computed using an implementation of Algorithm 10 in Matlab® running on a Laptop computer (Ubuntu Linux 16.04 `x86_64`, Intel® Core™M-5Y10c processor, 8GB RAM, Matlab® R2016b). More specifically, the linear measurement operator $\mathcal{M}|_{\mathcal{P}}$ was constructed by passing the standard basis elements $E_{i,j}, i, j \in [d], |i-j| < \delta \bmod d$ through (7.4). An LU decomposition of this sparse and structured matrix was pre-computed and stored for different values of d, δ for use in the numerical simulations below. We note that an FFT-based implementation of Step 1 of Algorithm 10 is likely to yield improved efficiency; we defer such an implementation to future work. The relative errors, defined by the expression $\dfrac{\min_\theta \|\mathrm{e}^{\mathrm{i}\theta} X - Q\|_F}{\|Q\|_F}$ (where X denotes the reconstruction (up to a global phase factor) of Q), were 4.288×10^{-16} and 2.857×10^{-16} for the reconstructions in Figures 7.1b and 7.1d respectively. The reconstructions were computed in 16.318 and 16.529 seconds respectively.

We next plot the execution time (in seconds, averaged over 50 trials) required to implement Algorithm 10 for different values of d in Fig. 7.2a. In each case, δ was chosen to be $\lceil \log_2(d) \rceil$, with the same choice of measurement vectors as for the reconstruction in Fig. 7.1b. The plot confirms that the proposed method is extremely efficient; indeed, the plot reveals an FFT-time empirical computational complexity of $\mathcal{O}(d^2 \log_2(d^2))$.

Finally, Fig. 7.2b illustrates the robustness of the proposed method to measurement errors. The figure plots the reconstruction error (averaged over 50 trials) in reconstructing a 64×64 random matrix with i.i.d zero-mean complex Gaussian entries from phaseless

(a) Test Image 1
(256×256 pixels)

(b) Reconstructed Image (Rel. error 4.288×10^{-16})

(c) Test Image 2
(256×256 pixels)

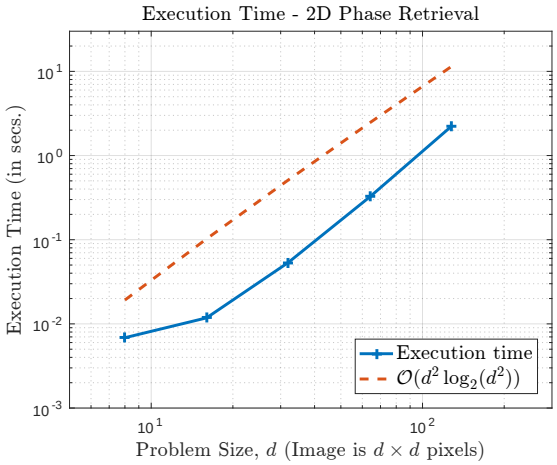
(d) Reconstructed Image (Rel. error 2.857×10^{-16})

Figure 7.1: Two Dimensional Image Reconstruction from Phaseless Local Measurements.

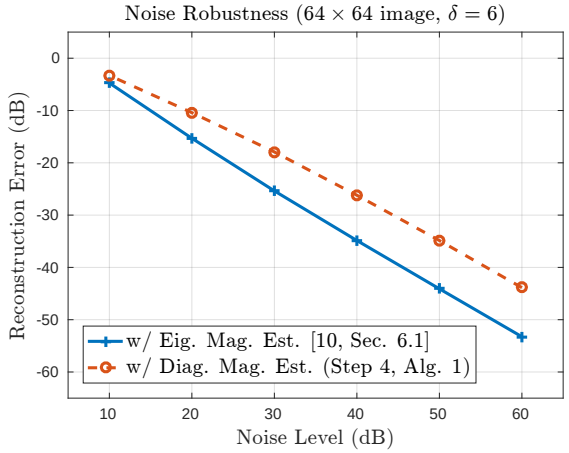
measurements (with $\delta = 6$, and with the same measurement construction as with Fig. 7.1b). An additive noise model with i.i.d. zero-mean Gaussian noise was used to corrupt the measurements. The added noise as well as reconstruction error are reported in decibels,[1]

$$\mathrm{SNR}_{\mathrm{db}}(dB) = 10\log_{10}\left(\frac{\|\mathbf{y}\|_2^2}{d^2D^2\sigma^2}\right), \qquad \text{Error (dB)} = 10\log_{10}\left(\frac{\min_\theta \|\mathrm{e}^{\mathrm{i}\theta}X - Q\|_F^2}{\|Q\|_F^2}\right).$$

We observe that the proposed algorithm (indicated by the dashed line) demonstrates robustness across a wide variety of SNRs. Additionally, the results from utilizing an improved magnitude estimation method (detailed in Section 3.6.1) in place of Step 4 of Algorithm 10 is plotted using the solid line. In both cases, we observe that the test signals are reconstructed to (almost) the level of added noise.



(a) Execution Time vs Problem Size



(b) Noise Robustness of the Proposed Method

Figure 7.2: Evaluating the Efficiency and Robustness of the Proposed Two Dimensional Phase Retrieval Algorithm.

Acknowledgments

This work was supported in part by NSF DMS-1416752.

[1]Note that we distinguish $\mathrm{SNR}_{\mathrm{db}}$ from $\mathrm{SNR} = \frac{\|\mathcal{A}(xx^*)\|_2}{\|n\|_2}$, though in the discussions of this section, we use "SNR" to refer to $\mathrm{SNR}_{\mathrm{db}}$.

7.4 Recovery Guarantees for 2D Phase Retrieval

In this section, we synthesize the results of Chapter 3 and Section 7.2 to produce guarantees of the robustness of Algorithm 10 of the type presented in Section 3.5. We begin with a few lemmas that help us reduce the 2D phase retrieval problem in such a way that permits us to import some of the results we have proven in the 1D case. First of all, we establish the relationship between the 2D measurement operator $\mathcal{M}$ defined in (7.4) and its 1D analog, defined in (4.1).

Proposition 20. *Fix $d, \delta \in \mathbb{N}$ such that $D := 2\delta - 1 \leq d$, and let $\{m_j\}_{j\in[D]}$ be a spanning family of masks. Then setting $\mathbf{a}_u = \mathbf{b}_u = m_u, u \in [D]$, and letting $\mathcal{M} : T_\delta(\mathbb{C}^{d^2\times d^2}) \to \mathbb{R}^{[d]_0^2\times[D]^2}$ and $\mathcal{A} : T_\delta(\mathbb{C}^{d\times d}) \to \mathbb{R}^{[d]_0\times[D]}$ be as defined in* (7.4) *and* (4.1)*. Then $\sigma_{\max}(\mathcal{M}) = \sigma_{\max}^2(\mathcal{A}), \sigma_{\min}(\mathcal{M}) = \sigma_{\min}^2(\mathcal{A})$, and $\kappa(\mathcal{M}) = \kappa(\mathcal{A})^2$.*

Proof of Proposition 20. The definition of the tensor power of a linear operator stipulates that, for linear operators $A_1 : V_1 \to W_1, A_1 : V_1 \to W_1, B : V_1 \otimes V_2 \to W_1 \otimes W_2$, we have that $B = A_1 \otimes A_2$ iff $B(u \otimes v) = A_1(u) \otimes A_2(v)$ for all $u, v \in V$. If we identify $\mathcal{B}(\mathbb{C}^{d^2\times d^2})$ with $(T_\delta(\mathbb{C}^{d\times d}))^{\otimes 2}$ and $\mathbb{R}^{[d]_0^2\times[D]^2}$ with $(\mathbb{R}^{[d]_0\times[D]})^{\otimes 2}$, then we claim that $\mathcal{M} = \overline{\mathcal{A}} \otimes \mathcal{A}$. Indeed, from (7.4) we can see that, for $U, V \in T_\delta(\mathbb{C}^{d\times d})$, we have

$$\begin{aligned}\mathcal{M}(U \otimes V)_{(\ell,\ell',u,v)} &= \langle U \otimes V, S_{\ell'}\overline{m_u m_u}^* S_{\ell'}^* \otimes S_\ell m_v m_v^* S_\ell^* \rangle \\ &= \langle U, S_{\ell'}\overline{m_u m_u}^* S_{\ell'}^* \rangle \langle V, S_\ell m_v m_v^* S_\ell^* \rangle,\end{aligned}$$

while

$$\begin{aligned}(\overline{\mathcal{A}(U)} \otimes \mathcal{A}(V))_{(\ell,\ell',u,v)} &= \overline{\mathcal{A}(U)}_{(\ell,u)} \otimes \mathcal{A}(V)_{(\ell',v)} \\ &= \langle U, S_{\ell'}\overline{m_u m_u}^* S_{\ell'}^* \rangle \langle V, S_\ell m_v m_v^* S_\ell^* \rangle,\end{aligned}$$

which proves the claim. To finish the proposition, we simply observe that the singular values of $\mathcal{M}$ are therefore the products of the singular values of $\mathcal{A}$ and $\overline{\mathcal{A}}$, which include $\sigma_{\min}(\mathcal{A})^2$ and $\sigma_{\max}(\mathcal{A})^2$ as extrema. □

Lemma 23 states a similar result regarding the process by which the main diagonal is extracted from the measurements.

Lemma 23. *With notation as in Propositions 11 and 20, we have that*

$$\kappa(P_{\mathrm{diag}(\mathbb{C}^{d^2\times d^2},0)} \circ \mathcal{M}^{-1}) = \kappa(P_{\mathrm{diag}(\mathbb{C}^{d\times d},0)} \circ \mathcal{A})^2.$$

Proof of Lemma 23. This comes immediately from observing that $P_{\mathrm{diag}(\mathbb{C}^{d^2\times d^2},0)} = P^{\otimes 2}_{\mathrm{diag}(\mathbb{C}^{d\times d},0)}$ and citing the fact that $\mathcal{M}^{-1} = \overline{\mathcal{A}}^{-1} \otimes \mathcal{A}^{-1}$ (as shown in the proof of Proposition 20). □

Proposition 21 establishes the spectral gap of the 2D analog of the graph G used in the angular synchronization step in line 3 of Algorithm 10, which will allow us to use results from Section 3.4 and Chapter 5.

Proposition 21. *Let $G = (V = [d]^2, E)$ be the unweighted graph whose adjacency matrix is given by $W = \mathcal{B}(\mathbb{1}_{d^2}\mathbb{1}_{d^2}^*) - I_{d^2}$, and let $D = \mathrm{diag}(W\mathbb{1}_{d^2})$ be the degree matrix of G. Then $\tau_G = \lambda_2(D - W)$, the spectral gap of G and the second smallest eigenvalue of $D - W$, is given by $(2\delta - 1)\tau_{1\mathrm{D}} = \mathcal{O}(\delta^4/d^2)$, where $\tau_{1\mathrm{D}} > \frac{\pi^2}{3}\frac{\delta^3}{d^2}$ is the spectral gap of the matrix described in Lemma 2.*

Proof of Proposition 21. Recognizing that $\mathbb{1}_{d^2} = \mathbb{1}_d^{\otimes 2}$, we calculate that

$$\begin{aligned} W\mathbb{1}_{d^2} &= (T_\delta(\mathbb{1}_d\mathbb{1}_d^*))^{\otimes 2}\mathbb{1}_{d^2} - \mathbb{1}_{d^2} = (T_\delta(\mathbb{1}_d\mathbb{1}_d^*)\mathbb{1}_d)^{\otimes 2} - \mathbb{1}_{d^2} \\ &= ((2\delta-1)\mathbb{1}_d)^{\otimes 2} - \mathbb{1}_{d^2} = ((2\delta-1)^2 - 1)\mathbb{1}_{d^2}, \end{aligned}$$

so that $D = ((2\delta-1)^2 - 1)I_{d^2}$. Therefore, by Theorem 12,

$$D - W = (2\delta-1)^2 I_{d^2} - (T_\delta(\mathbb{1}_d\mathbb{1}_d^*))^{\otimes 2} = F_d^{\otimes 2}((2\delta-1)^2 I_{d^2} - \Lambda^{\otimes 2})(F_d^*)^{\otimes 2},$$

where Λ is the diagonal matrix of $T_\delta(\mathbb{1}_d\mathbb{1}_d^*)$'s eigenvalues, as in Lemma 1. The conclusion of the proposition is then immediate:

$$\tau_G = (2\delta-1)^2 - \lambda_{d^2-1}(\Lambda^{\otimes 2}) = (2\delta-1)^2 - (2\delta-1)\lambda_{d-1}(\Lambda) = (2\delta-1)\tau_{1\mathrm{D}}.$$

□

The following lemma is a minor variant of (3.32) of Lemma 6.

Lemma 24. *Suppose $X, X_0 \in \mathbb{C}^{m\times n}$ have the same support set $\mathcal{I} \subseteq [m]\times[n]$ and $\min_{(j,k)\in\mathcal{I}}|(X_0)_{jk}| \geq a > 0$. Then if we define*

$$\widetilde{X}_{ij} = \begin{cases} \operatorname{sgn}(X_{ij}), & (i,j) \in \mathcal{I} \\ 0, & \textit{otherwise} \end{cases}$$

and $\widetilde{X}_0$ similarly, we have

$$\|\widetilde{X} - \widetilde{X}_0\|_F \leq \frac{2\|X - X_0\|_F}{a}$$

Proof of Lemma 24. We set $N = X - X_0$ and follow the argument of (3.33) to see that, for $(j,k) \in \mathcal{I}$,

$$|(\widetilde{X}_0)_{jk} - \widetilde{X}_{jk}| \leq 2\frac{|N_{jk}|}{|(X_0)_{jk}|} \leq 2\frac{|N_{jk}|}{a}.$$

The proof is complete by squaring both sides and summing over $\mathcal{I}$. □

Algorithm 11 2D Phase Retrieval, Improved Angular Synchronization

Input: Measurements $\mathbf{y} = \mathcal{M}(\mathbf{Q}) + n \in \mathbb{R}^{d^2D^2}$ and masks $\mathbf{a}_u, \mathbf{b}_v$ as per (7.3)
Output: $X \in \mathbb{C}^{d\times d}$ with $X \approx \mathrm{e}^{\mathrm{i}\theta}Q$ for some $\theta \in [0, 2\pi]$

1: Compute the Hermitian matrix $P = \left(\left(\mathcal{M}\big|_{\mathcal{P}}\right)^{-1}\mathbf{y}\right)/2 + \left(\left(\mathcal{M}\big|_{\mathcal{P}}\right)^{-1}\mathbf{y}\right)^*/2 \in \mathcal{P}\left(\mathbb{C}^{d^2\times d^2}\right)$ as an estimate of $\mathcal{P}(\mathbf{Q})$. $\mathcal{M}$ and $\mathcal{P}$ are as defined in (7.4) and §7.2.1.
2: Form the matrix of phases, $\widetilde{P} \in \mathcal{P}\left(\mathbb{C}^{d^2\times d^2}\right)$, by normalizing the non-zero entries of P.
3: Compute the solution $\hat{Z}$ to (5.6) with $L = (2\delta - 1)^2 I_{d^2} - \widetilde{P}$. Set $U = \operatorname{sgn}(\operatorname{mat}_{(d,d)} Z)$ to be the estimate of the phases of Q, where Z is the top eigenvector of $\hat{Z}$.
4: Use the diagonal entries of P to compute $M_{j,k} \approx |Q_{j,k}|^2$ for all $j,k \in [d]$ as per §7.2.3.
5: Set $X_{j,k} = \sqrt{M_{j,k}} \cdot U_{j,k}$ for all $j,k \in [d]$ to form X

Synthesizing the results in this section, we may give a complete error bound for 2D phase retrieval. We note that this uses the improved angular synchronization method proposed in Chapter 5. Although the numerical analysis in Section 7.3 was completed using Algorithm 10 (since the eigenvector-based angular synchronization is significantly cheaper and roughly as accurate, empirically), we present Algorithm 11 here for reference in Theorem 13.

Theorem 13. *Suppose that $\{m_j\}_{j\in[D]} \subseteq \mathbb{C}^d$ is a local measurement system with support δ satisfying $D = 2\delta - 1 \leq d$, and define $\mathcal{M}$ as in (7.3) with $\mathbf{a}_u = \mathbf{b}_u = m_u, u \in [D]$. Q is the ground truth objective image, and we set $\mathbf{Q} = \operatorname{vec} Q \operatorname{vec} Q^*$. Supposing further that line 3 of Algorithm 11 solves (5.3) exactly, then the output X satisfies*

$$\begin{aligned}
\min_{\theta\in[0,2\pi]} \|X - \mathrm{e}^{\mathrm{i}\theta} Q\|_F &\leq \frac{16}{5} \frac{\|\operatorname{vec} Q\|_\infty}{|\operatorname{vec} Q|_{\min}^2} \left(\frac{d}{\delta^2}\right) \sigma_{\min}^{-1}(\mathcal{M}) \|n\|_2 + d^{1/2} \sqrt{\sigma_{\min}^{-1}(\mathcal{M}) \|n\|_2} \\
\min_{\theta\in[0,2\pi]} \|X - \mathrm{e}^{\mathrm{i}\theta} Q\|_F &\leq \frac{16}{5} \frac{\|\operatorname{vec} Q\|_\infty}{|\operatorname{vec} Q|_{\min}^2} \left(\frac{d}{\delta^2}\right) \kappa(\mathcal{M}) \frac{\|\mathcal{B}(\mathbf{Q})\|_F}{\mathrm{SNR}} \\
&\qquad + d^{1/2} \sqrt{\kappa(\mathcal{M}) \frac{\|\mathcal{B}(\mathbf{Q})\|_F}{\mathrm{SNR}}}
\end{aligned} \tag{7.7}$$

Using γ_{flat} with a_δ as in Proposition 22, the second inequality becomes

$$\min_{\theta\in[0,2\pi]} \|X - \mathrm{e}^{\mathrm{i}\theta} Q\|_F \leq \frac{5}{2} \frac{d \|\operatorname{vec} Q\|_\infty}{|\operatorname{vec} Q|_{\min}^2} \frac{\|\mathcal{B}(\mathbf{Q})\|_F}{\mathrm{SNR}} \quad + d^{1/2}(1 + \frac{18}{\delta - 1}) \sqrt{\frac{\|\mathcal{B}(\mathbf{Q})\|_F}{\mathrm{SNR}}}$$

Proof of Theorem 13. We begin by splitting up $\|X - \mathrm{e}^{\mathrm{i}\theta} Q\|_F$ into "phase error" and "magnitude error" terms, by writing

$$\min_{\theta\in[0,2\pi]} \|X - \mathrm{e}^{\mathrm{i}\theta} Q\|_F \leq \min_{\theta\in[0,2\pi]} \| |Q| \circ (\operatorname{sgn}(X) - \mathrm{e}^{\mathrm{i}\theta} \operatorname{sgn}(Q)) \|_F + \| |X| - |Q| \|_F,$$

and the second term is immediately bounded by $(d^2)^{1/4} \sqrt{\sigma_{\min}^{-1}(\mathcal{M}) \|n\|_2}$ by quoting Lemma 7. To get the first term, we bound

$$\| |Q| \circ (\operatorname{sgn}(X) - \mathrm{e}^{\mathrm{i}\theta} \operatorname{sgn}(Q)) \|_F \leq \|\operatorname{vec} Q\|_\infty \|\operatorname{sgn}(X) - \mathrm{e}^{\mathrm{i}\theta} \operatorname{sgn}(Q)\|_F$$

and quote Lemma 24 to get, for $\widetilde{P}$ as defined as in line 2 of Algorithm 11,

$$\|\widetilde{P} - \operatorname{sgn} \mathbf{Q}\|_F \leq \frac{2\sigma_{\min}^{-1}(\mathcal{M}) \|n\|_2}{|\operatorname{vec} Q|_{\min}^2}.$$

We combine this with Theorem 9, using $\tau_G \geq \frac{\pi^2}{3} \frac{\delta^4}{d^2}$ from Proposition 21, to get

$$\|\operatorname{sgn}(X) - \mathrm{e}^{\mathrm{i}\theta} \operatorname{sgn}(Q)\|_F \leq \frac{2\sqrt{2} \|\widetilde{P} - \operatorname{sgn} \mathbf{Q}\|_F}{\sqrt{\tau_G}},$$

and combining these gives

$$\left\| |Q| \circ (\operatorname{sgn}(X) - \mathrm{e}^{\mathrm{i}\theta} \operatorname{sgn}(Q)) \right\|_F \leq \frac{\|\operatorname{vec} Q\|_\infty}{|\operatorname{vec} Q|_{\min}} \frac{4\sqrt{2}\sigma_{\min}^{-1}(\mathcal{M})\|n\|_2}{\sqrt{\frac{\pi^2}{3}\frac{\delta^4}{d^2}}},$$

and we combine the constants to obtain (7.7). The reduction in the case of γ_{flat} comes from Propositions 11 and 22. □

Acknowledgement of joint authorship

Sections 7.1–7.3, in full, are a reprint of material published with Mark Iwen, Rayan Saab, and Aditya Viswanathan in the Proceedings of SPIE vol. 10394, 2017. *Phase retrieval from local measurements in two dimensions.* Available online as of 24 August 2017.

Appendix A

Spanning Families for $\mathcal{H}^d$

In this section, we prove Proposition 3, as stated in Section 4.2.3. We restate the proposition here:

Proposition (Proposition 3)**.** *Suppose $\{m_j\}_{j\in[D]} \subseteq \mathbb{C}^d$ is a local Fourier measurement system of support δ with mask $\gamma \in \mathbb{R}^d$ and modulation index $K \geq D = \min(2\delta - 1, d)$. Then $\{m_j\}_{j\in[D]}$ is a spanning family, that is,*

$$\operatorname{span}_{\mathbb{R}}\{S^{\ell} m_j m_j^* S^{-\ell}\}_{(\ell,j)\in[d]_0\times[D]} = T_\delta(\mathcal{H}^d)$$

if and only if each of the sets $J_k := \{m \in [\delta]_0 : (F_d^*(\gamma \circ S^{-m}\gamma))_k \neq 0\}$, for all $k \in [d]$, satisfy

$$\begin{cases} 2|J_k| - 1 \geq D, & 0 \in J_k \\ 2|J_k| \geq D, & \textit{otherwise} \end{cases}.$$

Equivalently, we may require

$$\#\mathrm{nnz}\left(f_k^{d*} D_\gamma \begin{bmatrix} S^{1-\delta}\gamma & \cdots & S^{\delta-1}\gamma \end{bmatrix}\right) \geq D, \textit{ for all } k \in [d].$$

The proof makes use of the following technical lemmas concerning the linear independence of the real and imaginary parts of "truncated" Fourier vectors.

Lemma 25. *Define* $w_j = \mathcal{R}_{N_1}(f_j^{N_2}), j \in [N_2]$ *and set*

$$\rho_j = \mathrm{Re}(w_j) \quad \textit{and} \quad \mu_j = \mathrm{Im}(w_j)$$

to be vectors containing the real and imaginary components of w_j*. Then for* $1 \leq \ell_1 < \cdots < \ell_k \leq \frac{N_2+1}{2}$ *with* $k \leq N_1$*, we have*

$$\begin{aligned}\dim \operatorname{span}\{w_{\ell_i}, w_{2-\ell_i}\}_{i=1}^k &= \dim \operatorname{span}\{\rho_{\ell_i}, \mu_{\ell_i}\}_{i=1}^k \\ &= \begin{cases} 2k-1, & \ell_1 = 1, 2k-1 \leq N_1 \\ 2k, & \ell_1 \neq 1, 2k \leq N_1 \\ N_1, & \textit{otherwise} \end{cases},\end{aligned}$$

where the indices are taken modulo N_2*.*

Proof of Lemma 25. The first equality is clear by considering that $w_{2-i} = \overline{w_i}$, so $\rho_k = \frac{1}{2}(w_i + w_{2-i})$ and $\mu_i = -\frac{i}{2}(w_i - w_{2-i})$. We set $M = \dim \operatorname{span}\{w_{\ell_i}, w_{2-\ell_i}\}_{i=1}^k$ to be the common dimension of the two spaces under consideration.

We now divide into two cases: if $N_1 < N_2$, then $\{w_j\}_{j\in[N_2]}$ is full spark, as any $N_1 \times N_1$ submatrix of $\begin{bmatrix} w_1 & \cdots & w_{N_2} \end{bmatrix}$ will be a Vandermonde matrix of the form

$$V = \frac{1}{\sqrt{N_2}} \begin{bmatrix} w_{\ell_1} & \cdots & w_{\ell_{N_1}} \end{bmatrix}$$

with determinant

$$N_2^{-N_1/2} \prod_{1 \leq i < j \leq N_1} (\omega_{N_2}^{\ell_i - 1} - \omega_{N_2}^{\ell_j - 1}),$$

which is immediately non-zero since $\omega_{N_2}^{\ell_i-1} - \omega_{N_2}^{\ell_j-1} = 0$ only when $\ell_i - \ell_j = 0 \bmod N_2$, which cannot happen when $N_1 < N_2$.

When $N_1 \geq N_2, \{w_j\}_{j\in[N_2]}$ is linearly independent, since its members form the matrix $\begin{bmatrix} F_{N_2} \\ 0_{N_1 - N_2 \times N_2} \end{bmatrix}$.

In either case, M is equal to the cardinality of $\{\ell_i, 2-\ell_i\}_{i=1}^k$, which has $2k-1$ elements if and only if $\ell_1 = 1$; otherwise it has $2k$. We remark that a collision where $\ell_i = (2 - \ell_i \bmod N_2) = N_2/2 + 1$ is precluded since we have asserted $\ell_i \leq \frac{N_2+1}{2}$.

□

Lemma 26. *For $v \in \mathbb{R}^d$, we have*

$$\operatorname{circ}(v)\rho_k^d = \frac{1}{2}\operatorname{Re}((Fv)_k f_k^d) \tag{A.1}$$

$$\operatorname{circ}(v)\mu_k^d = \frac{1}{2}\operatorname{Im}((Fv)_k f_k^d). \tag{A.2}$$

In particular, if $(Fv)_k \neq 0$ and $k \notin \{1, \frac{d}{2}+1\}$, then $\rho_k^d, \mu_k^d \notin \operatorname{Nul}(\operatorname{circ}(v))$; if $k \in \{1, \frac{d}{2}+1\}$, then $\rho_k^d \notin \operatorname{Nul}(\operatorname{circ}(v))$ and $\mu_k^d = 0$. On the other hand, if $(Fv)_k = 0$, then $\rho_k^d, \mu_k^d \in \operatorname{Nul}(\operatorname{circ}(v))$.

Proof of Lemma 26. We set $\lambda_k^d = (Fv)_k$, and recalling that $\operatorname{circ}(v) = F \operatorname{diag}(Fv) F^*$, we observe that

$$\begin{aligned}\operatorname{circ}(v)\mu_k^d &= \operatorname{circ}(v)\frac{1}{2}(f_k^d + f_{2-k}^d) = \frac{1}{2}(\operatorname{circ}(v) f_k^d + \operatorname{circ}(v) f_{2-k}^d) \\ &= \frac{1}{2}(\lambda_k^d f_k^d + \lambda_{2-k}^d f_{2-k}^d).\end{aligned}$$

(A.1) follows immediately since $\lambda_k^d = \overline{\lambda_{2-k}^d}$ when $v \in \mathbb{R}^D$. (A.2) follows from an analogous calculation.

If $\lambda_k^d \neq 0$ and $k \notin \{1, \frac{d}{2}+1\}$, then ω_d^{k-1} is a non-real root of unity and there exists some j such that $\operatorname{Re}(\omega_d^{(j-1)(k-1)} \lambda_k^d) \neq 0$, and similarly for $\operatorname{Im}(\omega_d^{(j-1)(k-1)} \lambda_k^d) \neq 0$. When $k \in \{1, \frac{d}{2}+1\}, \omega_d^{(k-1)} \in \mathbb{R}$ so $\mu_k^d = 0$, but $\lambda_k^d \in \mathbb{R}$ in this case (because $v \in \mathbb{R}^d$), so $\operatorname{circ}(v)\rho_k^d = \lambda_k^d \rho_k^d \neq 0$. The claim concerning the case of $\lambda_k^d = 0$ is immediate from (A.1) and (A.2). □

With these in hand, we prove Proposition 3.

Proof of Proposition 3. For this proof, we set

$$(\rho_k^d, \mu_k^d) = (\mathrm{Re}(f_k^d), \mathrm{Im}(f_k^d))$$
$$(\rho_k, \mu_k) = (\mathrm{Re}(v_k), \mathrm{Im}(v_k))$$
$$(\rho_k^D, \mu_k^D) = (\mathrm{Re}(v_k^D), \mathrm{Im}(v_k^D))$$

In this case, we identify $\mathcal{L}_\gamma := \mathcal{L}_{\{m_j\}} = \mathrm{span}_{\mathbb{R}}\{S^\ell m_j m_j^* S^{-\ell}\}_{(\ell,j)\in[d]_0\times[D]}$. By a basic dimension count, $\{m_j\}_{j\in[D]}$ is a spanning family if and only if $\mathcal{L}_\gamma$ is linearly independent, so we consider the conditions under which a linear combination of this lifted measurement system can be equal to zero. To this end, we define the operator $\mathcal{A}^* : \mathbb{R}^{d\times D} \to \mathbb{C}^{d\times d}$ by

$$\mathcal{A}^*(C) = \sum_{\ell\in[d], j\in[D]} C_{\ell,j} S^\ell m_j m_j^* S^{-\ell} \tag{A.3}$$

and begin with the observation that, for any $A \in \mathbb{C}^{d\times d}$ we have

$$\mathrm{diag}(S^\ell A S^{-\ell}, m) = S^\ell \,\mathrm{diag}(A, m).$$

We then have

$$\begin{aligned}
&\sum_{j\in[D],\ell\in[d]} C_{\ell,j} S^\ell m_j m_j^* S^{-\ell} = 0 \\
\iff &\mathrm{diag}\left(\sum_{j\in[D],\ell\in[d]} C_{\ell,j} S^\ell m_j m_j^* S^{-\ell}, m\right) = 0 \quad \text{for all } m \in [\delta]_0 \\
\iff &\sum_{j\in[D],\ell\in[d]} C_{\ell,j}\,\mathrm{diag}(S^\ell m_j m_j^* S^{-\ell}, m) = 0 \quad \text{for all } m \in [\delta]_0 \\
\iff &\sum_{j\in[D],\ell\in[d]} C_{\ell,j} S^\ell\,\mathrm{diag}(m_j m_j^*, m) = 0 \quad \text{for all } m \in [\delta]_0
\end{aligned}$$

At this point, we recall from (4.21) that $\mathrm{diag}(m_j m_j^*, m) = \omega_K^{-m(j-1)}\,\mathrm{diag}(\gamma\gamma^*, m)$, so

we set $g_m := \mathrm{diag}(\gamma\gamma^*, m) = \gamma \circ S^{-m}\gamma$ and proceed with the previous chain of implications:

$$\begin{aligned}
&\sum_{j\in[D],\ell\in[d]} C_{\ell,j} S^\ell \,\mathrm{diag}(m_j m_j^*, m) = 0 \quad \text{for all } m \in [\delta]_0 \\
\iff &\sum_{j\in[D],\ell\in[d]} C_{\ell,j} S^\ell(\omega_K^{-m(j-1)} g_m) = 0 \quad \text{for all } m \in [\delta]_0 \\
\iff &\sum_{j\in[D],\ell\in[d]} C_{\ell,j} \omega_K^{-m(j-1)} S^\ell g_m = 0 \quad \text{for all } m \in [\delta]_0 \\
\iff &\qquad \mathrm{circ}(g_m) C\, v_{1-m}^D = 0 \quad \text{for all } m \in [\delta]_0
\end{aligned}$$

We now recall from Lemma 14 that any circulant matrix $\mathrm{circ}(v)$ is diagonalized by the Discrete Fourier Matrix, so that, by writing $\lambda_k^m = \sqrt{d}(F_d^* g_m)_k$, we get a natural decoupling of the previous equations: for a fixed m, we have that $\mathrm{circ}(g_m) C v_{1-m}^D = 0$ if and only if

$$\sum_{k=1}^{d} \lambda_k^m f_k^d f_k^{d*} C\, v_{1-m}^D = \sum_{k=1}^{d} (\lambda_k^m f_k^{d*} C\, v_{1-m}^D) f_k^d = 0.$$

Since this last expression is a linear combination of an orthonormal basis, it occurs only when $\lambda_k^m f_k^{d*} C\, v_{1-m}^D = 0$ for all $k \in [d]$. We collect these equations over $m \in [\delta]_0$, considering the definition of J_k and that $g_m \in \mathbb{R}^d$ implies $\lambda_k^m = 0 \iff \lambda_{2-k}^m = 0 \iff m \notin J_k$ to restate this condition as $\begin{bmatrix} f_k^d & f_{2-k}^d \end{bmatrix}^* C v_{1-m}^D = 0$ for all $k \in [d], m \in J_k$. Since $\mathrm{span}\{f_k^d, f_{2-k}^d\} = \mathrm{span}\{\rho_k^d, \mu_k^d\}$, we further restate this as $\begin{bmatrix} \rho_k^d & \mu_k^d \end{bmatrix}^* C v_{1-m}^D = 0$ for all $k \in [d], m \in J_k$; setting $W_k = C^* \begin{bmatrix} \rho_k^d & \mu_k^d \end{bmatrix} \in \mathbb{R}^{D\times 2}$, we now get that $\mathcal{A}^*(C) = 0 \iff \mathrm{Col}(W_k) \subseteq \{v_{1-m}^D\}_{m\in J_k}^\perp \cap \mathbb{R}^D$ for all $k \in [d]$.

We now claim that $\mathcal{A}^*$ is invertible if and only if the subspaces $\{v_{1-m}^D\}_{m\in J_k}^\perp \cap \mathbb{R}^D$ are all trivial. Indeed, if we fix a k and have some non-zero $u \in \{v_{1-m}^D\}_{m\in J_k}^\perp \cap \mathbb{R}^D$, then we may set $C = \rho_k^d u^*$, such that

$$\mathrm{circ}(g_m) C v_{1-m}^D = (\mathrm{circ}(g_m)\rho_k^d)(u^* v_{1-m}^D).$$

For $m \in J_k, u^* v_{1-m}^D = 0$ by hypothesis on u, and for $m \notin J_k, \mathrm{circ}(g_m)\rho_k^d = 0$ by definition of J_k and Lemma 26.

For the other direction, assume $\{v_{1-m}^D\}_{m\in J_k}^{\perp} \cap \mathbb{R}^D = \{0\}$ for each $k \in [d]$. Then $\mathcal{A}^*(C) = 0 \iff \mathrm{Col}(W_k) = \{0\} \iff W_k = 0$ for all k. However, $\{\rho_k^d\}_{k\in[d]} \cup \{\mu_k^d\}_{k\in[d]\setminus\{1,\frac{d}{2}+1\}}$ is an orthogonal basis for $\mathbb{R}^d$, so

$$\begin{aligned} & W_k = 0 \quad \text{for all } k \in [d] \\ \iff \quad & C^*\rho_k^d = C^*\mu_k^d = 0 \quad \text{for all } k \in [d] \\ \iff \quad & C = 0 \end{aligned}$$

We complete the proof by considering that, for $u \in \mathbb{R}^D$, $\langle v_j^D, u\rangle = 0$ if and only if $\langle \rho_j^D, v\rangle = \langle \mu_j^D, v\rangle = 0$, so

$$\{v_{1-m}^D\}_{m\in J_k}^{\perp} \cap \mathbb{R}^D = \{\rho_{1-m}^D, \mu_{1-m}^D\}_{m\in J_k}^{\perp}$$

which has dimension $\max\{D - (2|J_k| - \mathbb{1}_{0\in J_k}), 0\}$ by Lemma 25. Therefore, $\mathcal{A}^*$ is invertible if and only if $2|J_k| - \mathbb{1}_{0\in J_k} \leq D$ for all $k \in [d]$, as claimed. □

Appendix B

Conditioning Bound of Near-Flat Masks

In this section, we propose and prove an improvement to the condition number of the local Fourier measurement system studied in Section 4.3.2 by offering a different recommendation for a as a function of δ than that originally given. This result is stated in Proposition 22.

Proposition 22. *Let $\kappa(a, d, \delta)$ be the condition number of the measurement operator associated with the local Fourier measurement system $\{m_j\}_{j\in[2\delta-1]} \subseteq \mathbb{C}^d$ with support δ and mask $\gamma(a, d, \delta) \in \mathbb{R}^d$, where*

$$\gamma(a,d,\delta)_i = \begin{cases} a+1, & i = 1 \\ 1, & 2 \leq i \leq \delta \\ 0, & \textit{otherwise} \end{cases}$$

and we require $2\delta - 1 \leq d$. Then there exists a constant $C \in \mathbb{R}$ such that, setting $a_\delta = 2C\delta$, we have

$$\lim_{d,\delta\to\infty} \frac{\kappa(a_\delta, d, \delta)}{\delta} \leq \frac{7}{8}.$$

The proof makes use of Lemma 27.

Lemma 27. *Define the n^{th} Dirichlet kernel to be the function $D_n(\theta) : \mathbb{R} \to \mathbb{R}$ given by*

$$D_n(\theta) = \sum_{k=-n}^{n} \mathrm{e}^{\mathrm{i}k\theta}.$$

Then

$$\begin{aligned} D_n(\theta) &= 1 + 2\sum_{k=1}^{n} \cos(k\theta) \\ &= \frac{1}{2}\frac{\sin(\theta(n+1/2))}{\sin(\theta/2)} - \frac{1}{2} \end{aligned} \tag{B.1}$$

Proof of Lemma 27. The first line of (B.1) is obvious. The second comes from viewing D_n as a geometric sum, and taking

$$\begin{aligned} \sum_{k=-n}^{n} \mathrm{e}^{\mathrm{i}k\theta} &= \frac{\mathrm{e}^{\mathrm{i}-n\theta} - \mathrm{e}^{\mathrm{i}(n+1)\theta}}{1 - \mathrm{e}^{\mathrm{i}\theta}} \\ &= \frac{\mathrm{e}^{\mathrm{i}(-n-1/2)\theta} - \mathrm{e}^{\mathrm{i}(n+1/2)\theta}}{\mathrm{e}^{-\mathrm{i}\theta/2} - \mathrm{e}^{\mathrm{i}\theta/2}} \\ &= \frac{\sin((n+1/2)\theta)}{\sin(\theta/2)}. \end{aligned}$$

□

Proof of Proposition 22. Our strategy is as follows: recalling the form of

$$\kappa(a,d,\delta) = \frac{\|\gamma\|_2^2}{\min\limits_{(j,m)\in[d]\times[\delta]_0}\left|\sqrt{d}f_j^{d*}(\gamma\circ S^{-m}\gamma)\right|} = \frac{(a+1)^2 + (\delta-1)}{\min\limits_{(j,m)\in[d]\times[\delta]_0}\left|\sqrt{d}f_j^{d*}(\gamma\circ S^{-m}\gamma)\right|},$$

if we bound the denominator below by $\left|\sqrt{d}f_j^{d*}(\gamma\circ S^{-m}\gamma)\right| \geq a - C\delta + o(\delta)$, then we will have

$$\kappa(2C\delta, d, \delta) \leq \frac{(2C\delta+1)^2 + (\delta-1)}{C\delta + o(\delta)} = \frac{4C^2\delta^2 + (4C+1)\delta}{C\delta + o(\delta)},$$

which satisfies

$$\lim_{d,\delta\to\infty} \frac{\kappa(2C\delta, d, \delta)}{\delta} \leq 4C.$$

Therefore, it suffices to show that $\left|\sqrt{d}f_j^{d*}(\gamma\circ S^{-m}\gamma)\right| \geq a - C\delta + C' + o(1)$, for some $C \leq \frac{7}{32}$.

We define $h(a,j,m) = \sqrt{d} f_j^{d*}(\gamma \circ S^{-m}\gamma)$ and begin by recalling from (4.31) and (4.32) that

$$|h(a,j,\delta-k)| \geq \operatorname{Re} h(a,j,\delta-k) = a + \operatorname{Re}\left(\frac{1}{2} + \frac{\sin((k-\frac{1}{2})\theta)}{2\sin\frac{\theta}{2}}\right). \tag{B.2}$$

We then quote the sole proposition from [71], which tells us that

$$\lim_{n\to\infty} \frac{\min_\theta D_n(\theta)}{n} = C_0, \tag{B.3}$$

where D_n is the n^{th} Dirichlet kernel, as in Lemma 27, and $C_0 \approx -0.434467 > -7/16$ is the absolute minimum of $\frac{\sin(t)}{t}$. We then observe, citing (B.1) of Lemma 27, that (B.2) may be rewritten as $|h(a,j,\delta-k)| \geq a + 1/2 + \frac{1}{2}D_{k-1}(\theta)$, so that

$$\frac{|h(a,j,\delta-k)|}{\delta} \geq \frac{a}{\delta} + \frac{1}{2\delta} + \frac{1}{2}\min_\theta \frac{D_{k-1}(\theta)}{\delta}.$$

The last term may be bounded below by $C_0/2 > -7/32$, since $k-1 < \delta$. Using (B.3), this may be rephrased as

$$\left|\sqrt{d} f_j^{d*}(\gamma \circ S^{-m}\gamma)\right| \geq a - \frac{7}{32}\delta + o(\delta),$$

which completes the proof. □

Appendix C

Vector Synchronization

C.1 Vector Synchronization as Angular Synchronization

One subproblem that arises in some of the alternative reconstruction algorithms proposed in, e.g., Section 6.3.3, is that of *vector synchronization*, where, given a set of vectors $\{v_j\}_{j\in[n]} \subseteq \mathbb{C}^d$ and a graph $G = (V = [n], E)$, we want to find phases that minimize the pairwise distances between the v_j that are connected by the edges of the graph. More specifically, we want to solve

$$\min_{\theta_j\in[0,2\pi]} \sum_{(i,j)\in E} \|\mathrm{e}^{\mathrm{i}\theta_i} v_i - \mathrm{e}^{\mathrm{i}\theta_j} v_j\|_2^2. \tag{C.1}$$

In this section, we prove a short lemma that shows that this problem is equivalent to a certain instantiation of angular synchronization (see Chapter 5), and therefore may be solved using the technology and enjoying the recovery guarantees developed elsewhere in that chapter.

We derive this equivalence by showing that the minimizer (θ_i) for (C.1) is the same as the minimizer for

$$\min_{\theta\in[0,2\pi]} \sum_{(i,j)\in E} |v_i^* v_j| |\mathrm{e}^{\mathrm{i}\theta_i} - \mathrm{sgn}(v_i^* v_j)\mathrm{e}^{\mathrm{i}\theta_j}|^2, \tag{C.2}$$

which is, by following Eqs. (5.1)–(5.3), an angular synchronization problem with

$$\begin{aligned} W &= \left| \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^* \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \right| \circ A_G, \\ X &= \operatorname{sgn}\left(\begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^* \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \right) \circ A_G, \\ D &= \operatorname{diag}(W\mathbb{1}) \end{aligned} \tag{C.3}$$

where $|\cdot|$ and $\operatorname{sgn}(\cdot)$ are taken elementwise and A_G is the adjacency matrix of G. For convenience, we identify $\omega_j := \mathrm{e}^{\mathrm{i}\theta_j} \in \mathbb{S}^1$ and consider that the argument of (C.1) may be rewritten as

$$\sum_{(i,j)\in E} \|\omega_i v_i - \omega_j v_j\|_2^2 = \sum_{(i,j)\in E} \|v_i\|^2 + \|v_j\|^2 - 2\operatorname{Re}(\omega_i^* \omega_j v_i^* v_j).$$

Since the $\|v_i\|^2$ terms do not depend on ω_i, this is minimized by

$$\operatorname*{argmax}_{\omega_i \in \mathbb{S}^1} \sum_{(i,j)\in E} \operatorname{Re}(\omega_i^* \omega_j v_i^* v_j) = \operatorname*{argmax}_{\omega_i \in \mathbb{S}^1} |v_i^* v_j| \operatorname{Re}(\operatorname{sgn}(v_i^* v_j)\omega_i^* \omega_j).$$

Similarly, we rewrite the argument of (C.2) as

$$\sum_{(i,j)\in E} |v_i^* v_j| |\mathrm{e}^{\mathrm{i}\theta_i} - \operatorname{sgn}(v_i^* v_j)\mathrm{e}^{\mathrm{i}\theta_j}|^2 = \sum_{(i,j)\in E} |v_i^* v_j| \left(2 - 2\operatorname{Re}(\operatorname{sgn}(v_i^* v_j)\omega_i^* \omega_j)\right), \tag{C.4}$$

which is clearly minimized by

$$\operatorname*{argmax}_{\omega_i \in \mathbb{S}} \sum_{(i,j)\in E} |v_i^* v_j| \operatorname{Re}(\operatorname{sgn}(v_i^* v_j)\omega_i^* \omega_j).$$

The minimizers of these two problems coincide, so by solving (C.2), we obtain a solution to (C.1).

In order to state a recovery guarantee, we need to have a notion of a ground truth, and how we might consider the measurements v_i (or $v_i^* v_j, (i,j) \in E$, since (C.2) is completely determined by this data alone) to be noisy. To establish this, we observe, in light of (C.4), that the minimum value of (C.2) is zero iff there exist ω_i such that $\operatorname{sgn}(v_i^* v_j)\omega_i^* \omega_j = 1$ for all $(i,j) \in E$, which happens iff $D - W \circ X$ has a nullspace (further, as before, this solution

is unique iff this nullspace has dimension 1, which requires the graph described by W to be connected). Therefore, given a graph $G = (V = [n], E)$ and a set of vectors $\{v_j\}_{j\in[n]} \subseteq \mathbb{C}^d$, we say that v_j is *G-consistent* if $\dim \mathrm{Nul}(D - W \circ X) = 1$, with W, X, D defined as in (C.3). Then, with the previous discussion proving the coincident minimizers of Eqs. (C.1) and (C.2) and Theorem 8, we have Proposition 23. This proposition states that, if $\{v_j\}_{j\in[n]}$ is a perturbed version of some G-consistent $\{\underline{v}_j\}_{j\in[n]}$, then the synchronizing angles $\hat{\omega}_j$ of the v_j are close to the synchronizing angles $\underline{\omega}_j$ of the $\underline{v}_j$ and may be recovered according to the SDP relaxations discussed in Section 5.2 if the perturbation is small enough.

Proposition 23. *Suppose that $\{\underline{v}_j\}_{j\in[n]} \subseteq \mathbb{C}^d$ are G-consistent for some graph $G = (V = [n], E)$, and let $\{v_j\}_{j\in[n]}$ be arbitrary. Define $\underline{D}, \underline{W}, \underline{X}, D, W$, and X according to* (C.3)*, and $\underline{L} = \underline{D} - \underline{W} \circ \underline{X}, L - D - W \circ X$. Let $\tau = \lambda_2(\underline{D} - \underline{W})$. Suppose that $\hat{\omega} \in (\mathbb{S}^1)^n$, with $\hat{\omega}_j = \mathrm{e}^{\mathrm{i}\theta_j}$ minimizes* (C.1)*, and that $\underline{\omega} \in (\mathbb{S}^1)^n$ spans $\mathrm{Nul}(\underline{D} - \underline{W} \circ \underline{X})$. Then if $\|L - \underline{L}\|_2 < \frac{\tau}{1+\sqrt{n}}, \hat{\Omega} = \hat{\omega}\hat{\omega}^*$ and $\mathfrak{R}(\hat{\Omega})$ are the unique minimizers of* (5.6) *and* (5.7)*. In any case, we have*

$$\min_{\theta\in[0,2\pi]} \|\hat{\omega} - \mathrm{e}^{\mathrm{i}\theta}\underline{\omega}\|_2 \leq \frac{2\sqrt{2n}\|\underline{L} - L\|_2}{\tau}.$$

The end goal of this proposition is, setting $\underline{V} = \begin{bmatrix} \underline{v}_1 & \cdots & \underline{v}_n \end{bmatrix}$ and $V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}$, to find alignments $\underline{V}D_{\omega_1}$ and VD_{ω_2} of the $\underline{v}_j$'s and the v_j's for which we can control the distance between them in terms of $\|\underline{V} - V\|$.[1] Specifically, we want to bound

$$\begin{aligned}
\min_{\theta\in[0,2\pi]} \|\underline{V}D_{\underline{\omega}} - \mathrm{e}^{\mathrm{i}\theta}VD_{\hat{\omega}}\|_F &\leq \min_{\theta\in[0,2\pi]} \|\underline{V}(D_{\underline{\omega}} - \mathrm{e}^{\mathrm{i}\theta}D_{\hat{\omega}})\|_F + \|\underline{V} - V\|_F \\
&\leq \min_{\theta\in[0,2\pi]} \|\underline{V}\|_F \|\hat{\omega} - \mathrm{e}^{\mathrm{i}\theta}\underline{\omega}\|_\infty + \|\underline{V} - V\|_F \\
&\leq \frac{2\sqrt{2n}\|\underline{L} - L\|_2}{\tau}\|\underline{V}\|_F + \|\underline{V} - V\|_F
\end{aligned}$$

To bound the $\|\underline{L} - L\|_2$ term, we observe

$$\underline{L} - L = \mathrm{diag}(((\underline{V}^*\underline{V} - V^*V) \circ A_G)\mathbb{1}) - (\underline{V}^*\underline{V} - V^*V) \circ A_G.$$

[1] We remark that the bound we find here is markedly unsharp, as we use several costly inequalities to prove it. Its interest lies primarily in establishing that the vector synchronization problem is continuous, in the sense that the alignment $VD_{\hat{\omega}}$ satisfies $VD_{\hat{\omega}} \to \underline{V}D_{\underline{\omega}}$ as $V \to \underline{V}$.

If we then write $V = \underline{V} + N$, then $\underline{V}^*\underline{V} - V^*V = -N^*\underline{V} - \underline{V}^*N - N^*N$ which gives $\|\underline{V}^*\underline{V} - V^*V\|_F \leq \|N\|_F(2\|\underline{V}\|_2 + \|N\|_F)$. We use

$$\begin{aligned}\|\mathrm{diag}(X \circ A_G\mathbb{1})\|_2 = \|X \circ A_G\mathbb{1}\|_\infty &\leq \|X \circ A_G\mathbb{1}\|_2 \\ &\leq \sqrt{n}\|X \circ A_G\|_2 \leq \sqrt{n}\|X\|_F\end{aligned}$$

to obtain $\|\underline{L} - L\|_2 \leq (\sqrt{n}+1)\|N\|_F(2\|\underline{V}\|_2 + \|N\|_F)$, which gives the following corollary of Proposition 23.

Corollary 12. *Accept the notation of Proposition 23, and additionally set $N = \underline{V} - V$, where $\underline{V} = \begin{bmatrix}\underline{v}_1 & \cdots & \underline{v}_n\end{bmatrix}$ and $V = \begin{bmatrix}v_1 & \cdots & v_n\end{bmatrix}$. Then*

$$\min_{\theta\in[0,2\pi]} \|\underline{V}D_{\underline{\omega}} - \mathrm{e}^{\mathrm{i}\theta}VD_{\hat{\omega}}\|_F \leq \|N\|_F\left(\frac{4\sqrt{2}n}{\tau}\|\underline{V}\|_F(2\|\underline{V}\|_2 + \|N\|_F) + 1\right) \tag{C.5}$$

C.2 Phase Retrieval by Vector Synchronization

Here, we recall the setting of Section 6.3.3 and the notation of Algorithms 8 and 9. To prove that this process is continuous with respect to n, in the sense that

$$\lim_{\|n\|_2\to 0}\ \min_{\theta\in[0,2\pi]} \|x - \underline{x}\|_2 = 0,$$

we straightforwardly combine Lemma 9 and Corollary 12 with Lemma 28 to get Proposition 24.

Lemma 28. *Let $X, (\{J_i\}_{i\in[N]}, G)$ be a valid input to Algorithm 8, and suppose $\underline{X} = \underline{x}\underline{x}^*$ for some $\underline{x} \in \mathbb{C}^d$. Then*

$$\begin{aligned}\min_{\omega\in(\mathbb{S}^1)^N} \left\|\mathrm{BlockVec}(X, \{J_i\}) - \begin{bmatrix}\underline{x}^{(J_1)} & \cdots & \underline{x}^{(J_N)}\end{bmatrix} D_\omega\right\|_2 \leq \\ \frac{(1+2\sqrt{2})\sqrt{\max_j \mu_j}}{\min_i \|\underline{x}^{(J_i)}\|_2}\|X - \underline{X}\|_F,\end{aligned} \tag{C.6}$$

where $\mu_j = |\{i : i \in J_j\}|$.

Proof of Lemma 28. The proof proceeds very similarly to that of Proposition 18. Namely, we set $V = \mathrm{BlockVec}(X, \{J_i\}), \underline{V} = \mathrm{BlockVec}(\underline{X}, \{J_i\})$, and $\Delta_\omega = V - \underline{V} D_\omega$ and find

$$\begin{aligned}
\min_{\omega \in (\mathbb{S}^1)^N} \|\Delta_\omega\|_F^2 &= \sum_{i=1}^N \min_{\omega_i \in \mathbb{S}^1} \|x^{(J_i)} - \omega_i \underline{x}^{(J_i)}\|_2^2 \\
&\leq (1+2\sqrt{2})^2 \frac{\|X^{(J_i)} - \underline{X}^{(J_i)}\|_F^2}{\|\underline{x}^{(J_i)}\|_2^2} \\
&\leq \frac{(1+2\sqrt{2})^2 \max_j \mu_j}{\min_i \|\underline{x}^{(J_i)}\|_2^2} \|X - \underline{X}\|_F^2
\end{aligned}$$

□

Proposition 24. *Let $y, (\{J_i\}_{i\in[N]}, G')$ be a valid input to Algorithm 9 where $y = \mathcal{A}(\underline{x}\underline{x}^*) + n$ for some $\underline{x} \in \mathbb{C}^d, n \in \mathbb{R}^{\bar{d}D}$. Then setting*

$$\begin{aligned}
X = \mathcal{A}^{-1}(y), \underline{X} = \underline{x}\underline{x}^*, V = \begin{bmatrix} x^{(J_1)} & \cdots & x^{(J_N)} \end{bmatrix}, \\
\underline{V} = \begin{bmatrix} \underline{x}^{(J_1)} & \cdots & \underline{x}^{(J_N)} \end{bmatrix}, \Delta = V - \underline{V} D_{\hat{\omega}},
\end{aligned}$$

where

$$\hat{\omega} = \operatorname*{argmin}_{\omega \in (\mathbb{S}^1)^N} \|V - \underline{V} D_\omega\|_F$$

and μ_j as in line 1 of Algorithm 8, the output x satisfies

$$\begin{aligned}
\min_{\theta \in [0,2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta} \underline{x}\|_2 \leq \frac{d^{1/2}}{\min_j \mu_j} \|\Delta\|_F \left(\frac{4\sqrt{2}N}{\tau_{G'}} \|\underline{V}\|_F (2\|\underline{V}\|_2 + \|\Delta\|_F) + 1 \right), \quad \textit{where} \\
\|\Delta\|_F \leq (1+2\sqrt{2}) \frac{\sqrt{\max_j \mu_j}}{\min \|\underline{x}^{(J_i)}\|_2} \|X - \underline{X}\|_F \quad \textit{and} \\
\|X - \underline{X}\|_F \leq \sigma_{\min}^{-1}(\mathcal{A}) \|n\|_2,
\end{aligned}$$

so that

$$\lim_{\|n\|_2 \to 0} \min_{\theta \in [0,2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta} \underline{x}\|_2 = 0.$$

Proof of Proposition 24. We begin with an attempt to apply Corollary 12, so we accept its

notation and write

$$x = D_\mu^{-1} V \hat{z} = D_\mu^{-1} V D_{\hat{z}} \mathbb{1}_N \text{ and } \underline{x} = D_\mu^{-1} \underline{V} D_{\underline{\omega}} \mathbb{1}_N = D_\mu^{-1} \underline{V} D_{\hat{\omega}} D_{\underline{\omega} \circ \overline{\hat{\omega}}} \mathbb{1}_N,$$

so that we may immediately reach the bound

$$\min_{\theta \in [0,2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta} \underline{x}\|_2 \leq \frac{\sqrt{N}}{\min_j \mu_j} \min_{\theta \in [0,2\pi]} \|V D_{\hat{z}} - \mathrm{e}^{\mathrm{i}\theta} \underline{V} D_{\hat{\omega}} D_{\underline{\omega} \circ \overline{\hat{\omega}}}\|_2.$$

We remark that if $\underline{\omega}$ minimizes (C.1) for $\underline{V}$, then $\underline{\omega} \circ \overline{\hat{\omega}}$ minimizes (C.1) for $\underline{V} D_{\hat{\omega}}$. Hence, upper bounding $\|\cdot\|_2$ with $\|\cdot\|_F$, we may apply Corollary 12 to obtain

$$\min_{\theta \in [0,2\pi]} \|x - \mathrm{e}^{\mathrm{i}\theta} \underline{x}\|_2 \leq \frac{\sqrt{N}}{\min_j \mu_j} \|\Delta\|_F \left(\frac{4\sqrt{2}N}{\tau_{G'}} \|\underline{V}\|_F (2\|\underline{V}\|_2 + \|\Delta\|_F) + 1 \right),$$

which is the first of the three inequalities desired. The second comes immediately from Lemma 28, and the last comes from $\mathcal{A}(X - \underline{X}) = n$. □

Appendix D

Correction to Proof of Theorem 12 in [80]

The work of Chapter 5 was largely inspired by a paper authored by Rosen, Carlone, Bandeira, and Leonard in which the authors prove the tightness of an SDP relaxation analogous to ours in (5.6) which applies to the much more general problem of synchronization over $SE(d)$, the special Euclidean group on $\mathbb{R}^d$ [80]. In the course of preparing Theorem 8 of this dissertation, a flaw was discovered in the proof of Theorem 12 in [80], which is crucial to the proof of the main results of that paper.

Fortunately, however, the statement of Theorem 12 – and the results that depend on it – is still true. In this appendix, we describe the mistake in the original argument, a new argument that mends the proof, and a proof of an alternative bound that is slightly sharper in some cases.

D.1 Notation and Setting

We recall that $\underline{R}, R^*$ are elements of $O(d)^n$, meaning

$$\underline{R} = \begin{bmatrix} \underline{R}_1 & \cdots & \underline{R}_n \end{bmatrix} \text{ and } R^* = \begin{bmatrix} R_1^* & \cdots & R_n^* \end{bmatrix},$$

where $\underline{R}_i, R_i^* \in \mathbb{R}^{d\times d}$ are orthogonal matrices. P is the orthogonal projection of R^* onto the orthogonal complement of $\underline{R}$. In $O(d)^n$, this means

$$P = R^* - \frac{1}{n} R^* \underline{R}^T \underline{R},$$

and in (143) we discover that

$$\|P\|_F^2 = dn - \frac{1}{n} \left\| \underline{R} R^{*T} \right\|_F^2.$$

We define the $O(d)$ orbital distance between $\underline{R}$ and R^* by

$$d_{\mathcal{O}}(\underline{R}, R^*) = \min_{G\in O(d)} \|\underline{R} - GR^*\|_F,$$

and from Theorem 5 (p. 35) we have that

$$d_{\mathcal{O}}(\underline{R}, R^*)^2 = 2dn - 2\|\underline{R} R^{*T}\|_*.$$

D.2 The argument for (150) is incorrect

Theorem 12, stated on p. 43 of [80], is proven on pp. 41-43. The final line of the proof says "combining inequalities (134), (140), (151), and (152), we obtain the following [Theorem 12]," but unfortunately (151) and (152) are not true in general.

The goal is to prove

$$\|P\|_F^2 \geq \frac{1}{2} d_{\mathcal{O}}(\underline{R}, R^*)^2, \tag{D.1}$$

which is essential in the proof of Theorem 12 (and Theorem 12 is quoted in the proof of the main result, Proposition 2, on p. 44!). However, the "optimizaton strategy" of (147)-(150) does not work. Setting $\delta = d_{\mathcal{O}}(\underline{R}, R^*)$, we can rephrase the optimization problem of (147) as

$$\begin{aligned} \max \quad & \sum_{i=1}^{d} \sigma_i^2 \\ \text{s.t.} \quad & \sum_{i=1}^{d} \sigma_i = a \\ & \sigma_i \geq 0 \end{aligned},$$

where $a = dn - \delta^2/2$. This, in turn, is obviously equivalent to

$$\max_{\|x\|_1 = a} \|x\|_2^2.$$

From standard norm inequalities, we have that $\frac{1}{\sqrt{d}}\|x\|_1 \leq \|x\|_2 \leq \|x\|_1$, with equality on the left when $x_i = t\mathbb{1}$ and equality on the right when $x = e_i$ for some $i \in [d]$. In this instance, the maximal value will be $\|x\|_1^2$, or

$$\epsilon^2 = \left(\sum_{i=1}^{d} \sigma_i\right)^2 = \left(dn - \frac{\delta^2}{2}\right)^2 = d\left(d\left(n - \frac{\delta^2}{2d}\right)^2\right),$$

which is greater than the value given in (150) by a factor of d. This leads to (151) becoming

$$\|P\|_F^2 \geq dn - \frac{d^2}{n}\left(n - \frac{\delta^2}{2d}\right)^2 = d\delta^2 + dn(1-d) - \frac{\delta^4}{4n}.$$

Following the argument of (152), we get

$$\|P\|_F^2 \geq \frac{d}{2}\delta^2 - dn(d-1),$$

which, if combined with (134) and (140), gives

$$\delta^2 \leq \frac{4n\|\Delta Q\|_2}{\lambda_{d+1}(Q)} + 2n(d-1).$$

In this state, the bound proves nothing; in particular, we don't have that $\lim_{\|\Delta Q\|\to 0}\|\Delta R\| = 0$, not to mention that Theorem 5 (p. 35) trivially bounds $\delta^2 \leq 2dn$.

D.3 An alternative argument

Consider that, from (143) and (144), $\|P\|_F^2 \geq \frac{1}{2}d_{\mathcal{O}}(\underline{R}, R^*)^2$ holds for $\underline{R}, R^* \in O(d)^n$ iff

$$dn - \frac{1}{n}\left\|\underline{R}R^{*T}\right\|_F^2 \geq dn - \left\|\underline{R}R^{*T}\right\|_* \iff \left\|\underline{R}R^{*T}\right\|_* \geq \frac{1}{n}\left\|\underline{R}R^{*T}\right\|_F^2,$$

so it suffices to prove this last inequality for all $\underline{R}, R^* \in O(d)^n$. Fix $\underline{R}, R^* \in O(d)^n$; taking $\sigma_1 \geq \ldots \geq \sigma_d$ to be the singular values of $\underline{R}R^{*T}$ and setting $(\sigma_1, \ldots, \sigma_n)^T =: \sigma \in \mathbb{R}^n$, this happens iff

$$\|\sigma\|_1 \geq \frac{1}{n}\|\sigma\|_2^2. \tag{D.2}$$

By Hölder's inequality, we have

$$\|\sigma\|_2^2 \leq \|\sigma\|_1\|\sigma\|_\infty. \tag{D.3}$$

Now

$$\|\sigma\|_\infty = \left\|\underline{R}R^{*T}\right\| = \left\|\sum_{i=1}^n \underline{R}_i R_i^{*T}\right\| \leq \sum_{i=1}^n \left\|\underline{R}_i R_i^{*T}\right\| = n, \tag{D.4}$$

since $\underline{R}_i R_i^{*T}$ is orthogonal. Combining (D.3) and (D.4), we have

$$\frac{1}{n}\|\sigma\|_2^2 \leq \frac{1}{n}\|\sigma\|_1\|\sigma\|_\infty \leq \frac{1}{n}\|\sigma\|_1 n = \|\sigma\|_1,$$

which yields (D.2) as needed.

D.4 Improvement to the bound of Theorem 12

In sections D.2 and D.3, we ignored the origin of $\underline{R}$ and R^*, but in the present section D.4, we note that $\underline{R}$ and R^* arise as optima of certain constrained optimization

problems, though for now we ignore the specifics and assume only the relevant identities. Given symmetric matrices $\tilde{Q}, \underline{Q} \succeq 0$ in $\mathbb{R}^{dn \times dn}$, we assume $\mathrm{Tr}(\underline{Q}\underline{R}^T\underline{R}) = 0$ and that $\mathrm{Tr}(\tilde{Q}R^{*T}R^*) \leq \mathrm{Tr}(\tilde{Q}\underline{R}^T\underline{R})$. Setting $\Delta Q = \tilde{Q} - \underline{Q}$, these may be combined into

$$\begin{aligned}\mathrm{Tr}(\tilde{Q}\underline{R}^T\underline{R}) &= \mathrm{Tr}(\Delta Q\underline{R}^T\underline{R}) + \mathrm{Tr}(\underline{Q}\underline{R}^T\underline{R}) \\ &\geq \mathrm{Tr}(\Delta QR^{*T}R^*) + \mathrm{Tr}(\underline{Q}R^{*T}R^*) = \mathrm{Tr}(\tilde{Q}R^{*T}R^*),\end{aligned}$$

which we rearrange to get

$$\mathrm{Tr}(\underline{Q}R^{*T}R^*) \leq \mathrm{Tr}(\Delta Q\underline{R}^T\underline{R}) - \mathrm{Tr}(\Delta QR^{*T}R^*). \tag{D.5}$$

In the style of the proof of Lemma 4.1 in [9], we are able to achieve a notable improvement to the bound of Theorem 12 in [80]. From (D.5), and using $\mathrm{Tr}(\Delta Q\underline{R}^T\underline{R}) = \mathrm{vec}(\underline{R})^T(\Delta Q \otimes I_n)\,\mathrm{vec}(\underline{R})$, we get

$$\begin{aligned}\mathrm{Tr}(\underline{Q}R^{*T}R^*) &\leq \mathrm{vec}(\underline{R} - R^*)^T(\Delta Q \otimes I_n)\,\mathrm{vec}(\underline{R} + R^*) \\ &\leq \|\mathrm{vec}(\underline{R} - R^*)\|_2\|\Delta Q \otimes I_n\|_2\|\mathrm{vec}(\underline{R} + R^*)\|_2 \\ &= \|\underline{R} - R^*\|_F\|\Delta Q\|_2\|\underline{R} + R^*\|_F \\ &\leq 2\sqrt{dn}\|\underline{R} - R^*\|_F\|\Delta Q\|_2\end{aligned}$$

Assuming, without loss of generality, that $\underline{R}$ and R^* are representatives of their orbits such that $\|\underline{R} - R^*\|_F = d_{\mathcal{O}}(\underline{R}, R^*)$ and combining this with (D.1) and (140) of [80], which gives

$$\mathrm{Tr}(\underline{Q}R^{*T}R^*) \geq \lambda_{d+1}(Q)\|P\|_F^2,$$

we have

$$d_{\mathcal{O}}(\underline{R}, R^*) \leq \frac{4\sqrt{dn}\|\Delta Q\|_2}{\lambda_{d+1}(Q)}. \tag{D.6}$$

We compare this to the original result, which gives

$$d_{\mathcal{O}}(\underline{R}, R^*) \leq \sqrt{\frac{4dn\|\tilde{Q} - \underline{Q}\|_2}{\lambda_{d+1}(\underline{Q})}}. \tag{D.7}$$

We remark that, as $\|\tilde{Q} - \underline{Q}\|_2$ goes to zero, the higher exponent in the new bound guarantees a faster rate of convergence of $R^* \to \underline{R}$. Since both bounds are valid, we may always use whichever is stronger: indeed, the square root bound of (D.7) is stronger exactly when $\|\tilde{Q} - \underline{Q}\|_2/\lambda_{d+1}(\underline{Q}) > \frac{1}{4}$. We remark that (D.6) is trivial when $\|\tilde{Q} - \underline{Q}\|_2/\lambda_{d+1}(\underline{Q}) \geq \frac{1}{2\sqrt{2}}$, while (D.7) is trivial when $\|\tilde{Q} - \underline{Q}\|_2/\lambda_{d+1}(\underline{Q}) \geq \frac{1}{2}$, as $d_{\mathcal{O}}(\underline{R}, R^*) \leq \sqrt{2dn}$.

Appendix E

Software Statement

The software and hardware used for the numerical evaluations included in Sections 3.6 and 7.3 is described in those sections. The remainder of the numerical experiments (specifically, those contained in Sections 4.5, 5.4 and 6.4.3) were run on a laptop computer running Linux Mint 18.1 with an Intel© Core™ i5-6300U CPU @ (2.40 GHz × 2). For these sections, all computation was done using GNU Octave, version 4.0.0 [32], plots were produced using Gnuplot, version 4.6.7 [103], and the convex optimization problems were solved using SeDuMi version 1.32 [89]. The scripts used to test the methods described in the dissertation and produce the charts and graphs presented are available on GitHub at https://github.com/bpreskit/pr_code.

References

[1] S. Abrahamsson, D. Hodgkin, and E. Maslen. The crystal structure of phenoxymethylpenicillin. *Biochemical Journal*, 86(3):514–535, 1963. ISSN 0264-6021. doi: 10.1042/bj0860514. URL http://www.biochemj.org/content/86/3/514.

[2] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon. Phase retrieval with polarization. *SIAM Journal on Imaging Sciences*, 7(1):35–66, 2014.

[3] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton. Complementarity and nondegeneracy in semidefinite programming. *Mathematical Programming*, 77(1):111–128, Apr 1997. ISSN 1436-4646. doi: 10.1007/BF02614432.

[4] R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

[5] R. Balan, B. Bodmann, P. Casazza, and D. Edidin. Fast algorithms for signal reconstruction without phase. In *Optical Engineering+ Applications*, pages 67011L–67011L. International Society for Optics and Photonics, 2007.

[6] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin. Painless reconstruction from magnitudes of frame coefficients. *Journal of Fourier Analysis and Applications*, 15(4): 488–501, 2009.

[7] A. S. Bandeira and D. G. Mixon. Near-optimal phase retrieval of sparse vectors. In *Wavelets and Sparsity XV*, volume 8858, page 88581O. International Society for Optics and Photonics, 2013.

[8] A. S. Bandeira, A. Singer, and D. A. Spielman. A Cheeger inequality for the graph connection Laplacian. *SIAM Journal on Matrix Analysis and Applications*, 34(4): 1611–1630, 2013. doi: 10.1137/120875338.

[9] A. S. Bandeira, N. Boumal, and A. Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*, 163 (1):145–167, May 2017. ISSN 1436-4646. doi: 10.1007/s10107-016-1059-6.

[10] R. E. Bank and C. C. Douglas. Sparse matrix multiplication package (smmp). *Advances in Computational Mathematics*, 1(1):127–137, Feb 1993. ISSN 1572-9044. doi: 10.1007/BF02070824. URL https://doi.org/10.1007/BF02070824.

[11] H. Bauschke, P. Combettes, and D. Luke. Hybrid projection-reflection method for phase retrieval. *Journal of the Optical Society of America. A, Optics, Image science, and Vision*, 20(6):1025–1034, 2003.

[12] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Phase retrieval, error reduction algorithm, and Fienup variants: A view from convex optimization. *Journal of the Optical Society of America. A, Optics, Image science, and Vision*, 19(7):1334–1345, 2002.

[13] T. Bendory, Y. C. Eldar, and N. Boumal. Non-convex phase retrieval from STFT measurements. *IEEE Transactions on Information Theory*, 64(1):467–484, 2018.

[14] B. G. Bodmann and N. Hammen. Stable phase retrieval with low-redundancy frames. *Advances in computational mathematics*, 41(2):317–331, 2015.

[15] W. H. Bragg and L. Bragg. *X-rays and crystal structure*. G. Bell and Sons, Ltd London, 1915.

[16] J. Briales and J. Gonzalez-Jimenez. Cartan-sync: Fast and global $SE(d)$-synchronization. *IEEE Robotics and Automation Letters*, 2(4):2127–2134, Oct 2017. ISSN 2377-3766. doi: 10.1109/LRA.2017.2718661.

[17] D. Brodersen, W. Clemons, A. P. Carter, R. J. Morgan-Warren, B. Wimberly, and V. Ramakrishnan. The structural basis for the action of the antibiotics Tetracycline, Pactamycin, and Hygromycin B on the 30S ribosomal subunit. *Cell*, 103:1143–54, 01 2001.

[18] G. C. Calafiore, L. Carlone, and F. Dellaert. *Lagrangian Duality in Complex Pose Graph Optimization*, pages 139–184. Springer International Publishing, Cham, 2016. ISBN 978-3-319-42056-1. doi: 10.1007/978-3-319-42056-1_5.

[19] E. J. Candes and X. Li. Solving quadratic equations via Phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5): 1017–1026, 2014. ISSN 1615-3375.

[20] E. J. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

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

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

[23] E. J. Cands, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, April 2015. ISSN 0018-9448. doi: 10.1109/TIT.2015.2399924.

[24] L. Carlone, G. C. Calafiore, C. Tommolillo, and F. Dellaert. Planar pose graph optimization: Duality, optimal solutions, and verification. *IEEE Transactions on Robotics*, 32(3):545–565, June 2016. ISSN 1552-3098. doi: 10.1109/TRO.2016.2544304.

[25] H. Chang, P. Enfedaque, Y. Lou, and S. Marchesini. Partially coherent ptychography by gradient decomposition of the probe. *Acta Crystallographica Section A*, 74(3):157–169, May 2018. doi: 10.1107/S2053273318001924.

[26] F. R. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

[27] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[28] N. M. M. de Abreu. Old and new results on algebraic connectivity of graphs. *Linear Algebra and its Applications*, 423(1):53 – 73, 2007. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2006.08.017. URL http://www.sciencedirect.com/science/article/pii/S0024379506003971. Special Issue devoted to papers presented at the Aveiro Workshop on Graph Spectra.

[29] M. Dierolf, O. Bunk, S. Kynde, P. Thibault, I. Johnson, A. Menzel, K. Jefimovs, C. David, O. Marti, and F. Pfeiffer. Ptychography & lensless X-ray imaging. *Europhysics News*, 39(1):22–24, 2008.

[30] R. Diestel. *Graph Theory*. Graduate Texts in Mathematics. Springer Berlin Heidelberg, 2017. ISBN 9783662536223. URL https://books.google.com/books?id=Qjm0tAEACAAJ.

[31] T. S. E. J. Cands and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013. doi: 10.1002/cpa.21432.

[32] J. W. Eaton, D. Bateman, S. Haubergand, and R. Wehbring. *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*. 2015. URL http://www.gnu.org/software/octave/doc/interpreter.

[33] Y. Eldar, P. Sidorenko, D. Mixon, S. Barel, and O. Cohen. Sparse phase retrieval from short-time fourier measurements. *IEEE Signal Proc. Letters*, 22(5), 2015.

[34] Y. C. Eldar and S. Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[35] V. Elser. Phase retrieval by iterated projections. *Journal of the Optical Society of America. A, Optics, Image science, and Vision*, 20(1):40–55, 2003.

[36] O. Enqvist, F. Kahl, and C. Olsson. Non-sequential structure from motion. In *Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras*, 2011.

[37] A. P. Eriksson, C. Olsson, F. Kahl, O. Enqvist, and T. Chin. Why rotation averaging is easy. *CoRR*, abs/1705.01362, 2017. URL http://arxiv.org/abs/1705.01362.

[38] M. Fiedler. Algebraic connectivity of graphs. 23:298–305, 01 1973.

[39] J. R. Fienup. Reconstruction of an object from the modulus of its Fourier transform. *Optics letters*, 3(1):27–29, 1978.

[40] J. R. Fienup. Phase retrieval algorithms: A comparison. *Applied Optics*, 21(15): 2758–2769, 1982.

[41] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782. doi: 10.1145/358669.358692.

[42] R. E. Franklin and R. G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171:740–741, Apr 1953.

[43] R. G. Gallager. *Principles of Digital Communication*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511813498.

[44] S. Galli. X-ray crystallography: One century of Nobel prizes. *Journal of Chemical Education*, 91(12):2009–2012, 2014. doi: 10.1021/ed500343x.

[45] R. Gerchberg and W. Saxton. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*, 35:237246, 1972.

[46] J. W. Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[47] V. M. Govindu. Robustness in motion averaging. In P. J. Narayanan, S. K. Nayar, and H. Shum, editors, *Computer Vision - ACCV 2006, 7th Asian Conference on Computer Vision, Hyderabad, India, January 13-16, 2006, Proceedings, Part II*, volume 3852 of *Lecture Notes in Computer Science*, pages 457–466. Springer, 2006. ISBN 3-540-31244-7. doi: 10.1007/11612704_46.

[48] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag, 2008. http://stanford.edu/~boyd/graph_dcp.html.

[49] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, Mar. 2014.

[50] R. M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006. ISSN 1567-2190. doi: 10.1561/0100000006.

[51] D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 2015.

[52] H. A. Hauptman and J. Karle. *Solution of the phase problem,*. American Crystallographic Association, Ann Arbor, Mich., 1953.

[53] R. Hausbrand, G. Cherkashinin, H. Ehrenberg, M. Grting, K. Albe, C. Hess, and W. Jaegermann. Fundamental degradation mechanisms of layered oxide Li-ion battery cathode materials: Methodology, insights and novel approaches. *Materials Science and Engineering: B*, 192:3 – 25, 2015. ISSN 0921-5107. doi: https://doi.org/10.1016/j.mseb.2014.11.014. URL http://www.sciencedirect.com/science/article/pii/S0921510714002657. Electrical Fatigue.

[54] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[55] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

[56] M. Iwen, F. Krahmer, and A. Viswanathan. Technical note: A minor correction of Theorem 1.3 from [1]. *Unpublished note available at http://users.math.msu.edu/users/markiwen/Papers/PhaseLiftproof.pdf*, April 2015.

[57] M. Iwen, A. Viswanathan, and Y. Wang. Fast phase retrieval from local correlation measurements. *SIAM Journal on Imaging Sciences*, 9(4):1655–1688, 2016.

[58] M. Iwen, Y. Wang, and A. Viswanathan. BlockPR: Matlab software for phase retrieval using block circulant measurement constructions and angular synchronization, version 2.0. https://bitbucket.org/charms/blockpr, Apr. 2016.

[59] K. Jaganathan, Y. C. Eldar, and B. Hassibi. STFT phase retrieval: Uniqueness guarantees and recovery algorithms. *IEEE Journal of selected topics in signal processing*, 10(4):770–781, 2016.

[60] M. S. Kimber, F. Vallee, S. Houston, A. Neakov, T. Skarina, E. Evdokimova, S. Beasley, D. Christendat, A. Savchenko, C. H. Arrowsmith, M. Vedadi, M. Gerstein, and A. M. Edwards. Data mining crystallization databases: Knowledge-based approaches to optimize protein crystal screens. *Proteins: Structure, Function, and Bioinformatics*, 51 (4):562–568, 2003. doi: 10.1002/prot.10340.

[61] W. Krätschmer, L. D. Lamb, K. Fostiropoulos, and D. R. Huffman. Solid C_{60}: a new form of carbon. *Nature*, 347:354 EP –, Sep 1990.

[62] H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley. C_{60}: Buckminsterfullerene. *Nature*, 318:162 EP –, Nov 1985.

[63] R. Kueng, D. Gross, and F. Krahmer. Spherical designs as a tool for derandomization: The case of phaselift. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 192–196, May 2015. doi: 10.1109/SAMPTA.2015.7148878.

[64] R. Kueng, H. Zhu, and D. Gross. Low rank matrix recovery from clifford orbits. *CoRR*, abs/1610.08070, 2016. URL http://arxiv.org/abs/1610.08070.

[65] A. J. Laub. *Matrix Analysis For Scientists And Engineers*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2004. ISBN 0898715768.

[66] X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

[67] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and Computational Harmonic Analysis*, 2018. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2018.01.001. URL http://www.sciencedirect.com/science/article/pii/S1063520318300149.

[68] S. Ling and T. Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015. URL http://stacks.iop.org/0266-5611/31/i=11/a=115002.

[69] S. Ling and T. Strohmer. Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing. *Information and Inference: A Journal of the IMA*, page iax022, 2018. doi: 10.1093/imaiai/iax022.

[70] S. Marchesini, Y.-C. Tu, and H.-t. Wu. Alternating projection, ptychographic imaging and phase synchronization. *Applied and Computational Harmonic Analysis*, 41(3): 815–851, 2016.

[71] I. D. Mercer. The minimum value and the L^1 norm of the Dirichlet kernel. URL http://www.idmercer.com/dirichletkernel.pdf.

[72] S. Merhi, A. Viswanathan, and M. Iwen. Recovery of compactly supported functions from spectrogram measurements via lifting. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pages 538–542. IEEE, 2017.

[73] R. Millane. Phase retrieval in crystallography and optics. *Journal of the Optical Society of America. A, Optics, Image science, and Vision*, 7(3):394–411, 1990.

[74] B. Mohar. Eigenvalues, diameter, and mean distance in graphs. *Graphs and Combinatorics*, 7(1):53–64, Mar 1991. ISSN 1435-5914. doi: 10.1007/BF01789463.

[75] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

[76] A. L. Patterson. A fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.*, 46:372–376, Sep 1934. doi: 10.1103/PhysRev.46.372.

[77] G. E. Pfander and P. Salanevich. Robust phase retrieval algorithm for time-frequency structured measurements. *eprint arXiv:1611.02540*, 2016.

[78] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. ISBN 0-13-015157-2.

[79] S. G. F. Rasmussen, H.-J. Choi, D. M. Rosenbaum, T. S. Kobilka, F. S. Thian, P. C. Edwards, M. Burghammer, V. R. P. Ratnala, R. Sanishvili, R. F. Fischetti, G. F. X. Schertler, W. I. Weis, and B. K. Kobilka. Crystal structure of the human β_2 adrenergic g-protein-coupled receptor. *Nature*, 450:383 EP –, Oct 2007. Article.

[80] D. M. Rosen, L. Carlone, A. S. Bandeira, and J. J. Leonard. *SE*-sync: A certifiably correct algorithm for synchronization over the special euclidean group. *CoRR*, abs/1611.00128, 2016. URL http://arxiv.org/abs/1611.00128.

[81] P. Salanevich and G. E. Pfander. Polarization based phase retrieval for time-frequency structured measurements. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 187–191. IEEE, 2015.

[82] T. Schindler, W. Bornmann, P. Pellicena, W. T. Miller, B. Clarkson, and J. Kuriyan. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science*, 289 (5486):1938–1942, 2000. ISSN 0036-8075. doi: 10.1126/science.289.5486.1938. URL http://science.sciencemag.org/content/289/5486/1938.

[83] D. A. Shapiro, Y.-S. Yu, T. Tyliszczak, J. Cabana, R. Celestre, W. Chao, K. Kaznatcheev, A. L. D. Kilcoyne, F. Maia, S. Marchesini, Y. S. Meng, T. Warwick, L. L. Yang, and H. A. Padmore. Chemical composition mapping with nanometre resolution by soft x-ray microscopy. *Nature Photonics*, 8:765–769, Sep 2014.

[84] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20 – 36, 2011. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2010.02.001. URL http://www.sciencedirect.com/science/article/pii/S1063520310000205.

[85] A. Singer, M. Zhang, S. Hy, D. Cela, C. Fang, T. A. Wynn, B. Qiu, Y. Xia, Z. Liu, A. Ulvestad, N. Hua, J. Wingert, H. Liu, M. Sprung, A. V. Zozulya, E. Maxey, R. Harder, Y. S. Meng, and O. G. Shpyrko. Nucleation of dislocations and their dynamics in layered oxide cathode materials during battery charging. *Nature Energy*, 3(8):641–647, 2018. ISSN 2058-7546. doi: 10.1038/s41560-018-0184-2.

[86] S. S. Skiena. *Graph Traversal*, pages 145–190. Springer London, London, 2012. ISBN 978-1-84800-070-4. doi: 10.1007/978-1-84800-070-4_5.

[87] D. Starodub, P. Rez, G. Hembree, M. Howells, D. Shapiro, H. N. Chapman, P. Fromme, K. Schmidt, U. Weierstall, R. B. Doak, and J. C. H. Spence. Dose, exposure time and resolution in serial X-ray crystallography. *Journal of Synchrotron Radiation*, 15(1): 62–73, Jan 2008. doi: 10.1107/S0909049507048893.

[88] G. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[89] J. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999. Latest version available from https://github.com/sqlp/sedumi.

[90] H. Takajo, T. Takahashi, H. Kawanami, and R. Ueda. Numerical investigation of the iterative phase-retrieval stagnation problem: Territories of convergence objects and holes in their boundaries. *Journal of the Optical Society of America. A, Optics, Image science, and Vision*, 14(12):3175–3187, 1997.

[91] H. Takajo, T. Takahashi, R. Ueda, and M. Taninaka. Study on the convergence property of the hybrid input–output algorithm used for phase retrieval. *Journal of the Optical Society of America. A, Optics, Image science, and Vision*, 15(11):2849–2861, 1998.

[92] H. Takajo, T. Takahashi, and T. Shizuma. Further study on the convergence property of the hybrid input–output algorithm used for phase retrieval. *Journal of the Optical Society of America. A, Optics, Image science, and Vision*, 16(9):2163–2168, 1999.

[93] Y. L. Tong. *The multivariate normal distribution*. Springer-Verlag, New York, 1990.

[94] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*, volume 50. SIAM, 1997.

[95] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996. doi: 10.1137/1038003. URL https://doi.org/10.1137/1038003.

[96] L. Varsani, T. Cui, M. Rangarajan, B. S. Hartley, J. Goldberg, C. Collyer, and D. M. Blow. Arthrobacter d-xylose isomerase: protein-engineered subunit interfaces. *Biochemical Journal*, 291(2):575–583, 1993. ISSN 0264-6021. doi: 10.1042/bj2910575. URL http://www.biochemj.org/content/291/2/575.

[97] A. Viswanathan and M. Iwen. Fast angular synchronization for phase retrieval via incomplete information. In *Wavelets and Sparsity XVI*, volume 9597, page 959718. International Society for Optics and Photonics, 2015.

[98] A. Walther. The Question of Phase Retrieval in Optics. *Optica Acta*, 10:41–49, 1963. doi: 10.1080/713817747.

[99] J. D. Watson. The involvement of RNA in the synthesis of proteins. In *Nobel Lectures, Physiology or Medicine 1942-1962*. Elsevier Publishing Company, December 1962.

[100] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, Apr 1953.

[101] J. H. M. Wedderburn. *Lectures on matrices*. American Mathematical Society New York, 1934.

[102] M. H. F. Wilkins, A. R. Stokes, and H. R. Wilson. Molecular structure of nucleic acids: Molecular structure of deoxypentose nucleic acids. *Nature*, 171:738–740, Apr 1953.

[103] T. Williams, C. Kelley, and many others. Gnuplot 4.6: an interactive plotting program. http://gnuplot.sourceforge.net/, July 2015.

[104] D. B. Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the Twenty-eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pages 296–303, New York, NY, USA, 1996. ACM. ISBN 0-89791-785-5. doi: 10.1145/237814.237880.

[105] C. Yang, J. Qian, A. Schirotzek, F. Maia, and S. Marchesini. Iterative Algorithms for Ptychographic Phase Retrieval. *ArXiv e-prints*, May 2011.

[106] Y. Yu, T. Wang, and R. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.