Convex programming-based phase retrieval: Theory and applications

Thesis by
Kishore Jaganathan

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2016
Defended May 16, 2016

To my family and friends.

En vazhi thani vazhi (my way is a unique way).

- Rajnikanth

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Babak Hassibi. I could not have imagined having a better advisor and mentor for my Ph.D. studies. His immense knowledge, guidance, kindness and support over the years have played a crucial role in making this work possible. My understanding of many topics, including convex optimization, signal processing and entropy vectors, have significantly increased because of him. His exceptional problem solving abilities, teaching qualities and deep understanding of a wide variety of subjects have inspired me a lot. Furthermore, the intellectual freedom he offered throughout the course of my graduate studies helped me pursue my passion and grow as a research scientist.

Besides my advisor, I am also extremely indebted to Prof. Yonina C. Eldar. I have been privileged to have had the opportunity to collaborate with her. Her vision and ideas have played a very important role in shaping this work. Her vast knowledge, attention to detail, work ethic and energy have influenced me significantly. Additionally, I would like to thank her for providing me the opportunity to contribute to a book chapter on phase retrieval.

I would also like to thank Prof. Palghat P. Vaidyanathan, Prof. Changhuei Yang, Prof. Joel Tropp and Prof. Venkat Chandrasekaran for serving on my defense committee. I am very grateful to Prof. Vaidyanathan for his words of wisdom, both professional and personal. I am also grateful to Prof. Yang for his valuable feedback on the practical aspects of this work, and to Prof. Tropp and Prof. Chandrasekaran for their useful convex optimization related suggestions. I would also like to express my gratitude to my undergraduate advisor and mentor Prof. David Koilpillai for being a constant source of inspiration.

I am indebted to my labmates for their role in creating a wonderful research environment. I would like to thank Ravi Teja Sukhavasi, the First of His Name, for his help during the initial stages. To Teja, thank you for showing me that graduate students can own a convertible car. I am also indebted to Samet Oymak for his role in getting me started with academic research. He introduced me to the three fundamental laws of the lab, and one of them turned out to be incredibly useful. I have fond memories of the trip to Japan with Samet and Matthew Thill. Indeed, I do not have any photos of it despite doing multiple favors to both of them. I would like to thank Christos "Greek Machine" Thrampoulidis (CGMT) for his inputs on the use of colors apart

from red and black in a presentation. I would also like to thank Anatoly Khina for all of the advice and conversations. The lab has been a lively place due to the energy of Ehsan Abbasi, Ramya Korlakai Vinayak and Wael Halbawi. I acknowledge the technical contributions of James Saunderson, Philipp Walk and Roarke Horstmeyer. The lunch meetings with Roarke significantly improved my understanding of the practical aspects of this work. I would like to express my gratitude to Shirley Slattery for her kind and patient administrative assistance.

I would also like to use this opportunity to thank the people who made my stay at Caltech a truly memorable one. I am forever indebted to Anupama Lakshmanan and Samet Oymak for always being there for me. To Anupama, your "bhooli bhaaliness" has added a lot of color to my life. Several of my best memories at Caltech are due to you, and I thank you for that. To Samet, I formally accept the fact that you are better than me at racquetball. Your TAing and parking skills will always be an inspiration to me.

I will always cherish the time I spent with Neel Nadkarni and Pushkar Kopparla. To Neel, thank you in advance for making the students of IIT Gandhinagar read my thesis, and for letting me travel in your Neel 45 and Neel 65 yachts. To Pushkar, thank you for introducing me to quality games like hand cricket and DotA.

The Caltech Cricket Club has been an integral part of my stay at Caltech. I would like to sincerely thank Charles Steinhardt, Siddharth Jain (and his twin brother Siddhanth Jain) and David Hall for keeping the team active, motivated and organized. I would also like to thank Sumanth Dathathri, Anantha Ravi Kiran and Sisir Yalamanchili for their camaraderie.

I would like to express my appreciation to Wong Ming Fai, Asma Qureshi, Chandni Usha (and her mother) and Harish Ravishankar for their constant support. My thanks are due to Srikanth Tenneti, Tejaswi Venumadhav and Vikas Trivedi for being accommodating roommates. I am very grateful to Rakesh Misra, Chinmoy Venkatesh, Srinidhi Tirupattur, Karthikeyan Shanmugam and Ananda Theertha Suresh for their support and encouragement.

Finally, I would like to thank my parents Jaganathan Subramaniam and Alamelu Jaganathan, and my brother Sushil Jaganathan, for supporting me emotionally throughout my life.

ABSTRACT

Phase retrieval is the problem of recovering a signal from its Fourier magnitude. This inverse problem arises in many areas of engineering and applied physics, and has been studied for nearly a century. Due to the absence of Fourier phase, the available information is incomplete in general. Classic identifiability results state that phase retrieval of one-dimensional signals is impossible, and that phase retrieval of higher-dimensional signals is almost surely possible under mild conditions. However, there are no efficient recovery algorithms with theoretical guarantees. Classic algorithms are based on the method of alternating projections. These algorithms do not have theoretical guarantees, and have limited recovery abilities due to the issue of convergence to local optima.

Recently, there has been a renewed interest in phase retrieval due to technological advances in measurement systems and theoretical developments in structured signal recovery. In particular, it is now possible to obtain specific kinds of additional magnitude-only information about the signal, depending on the application. The premise is that, by carefully redesigning the measurement process, one could potentially overcome the issues of phase retrieval. To this end, another approach could be to impose certain kinds of prior on the signal, depending on the application. On the algorithmic side, convex programming based approaches have played a key role in modern phase retrieval, inspired by their success in provably solving several quadratic constrained problems.

In this work, we study several variants of phase retrieval using modern tools, with focus on applications like X-ray crystallography, diffraction imaging, optics, astronomy and radar. In the one-dimensional setup, we first develop conditions, which when satisfied, allow unique reconstruction. Then, we develop efficient recovery algorithms based on convex programming, and provide theoretical guarantees. The theory and algorithms we develop are independent of the dimension of the signal, and hence can be used in all the aforementioned applications. We also perform a comparative numerical study of the convex programming and the alternating projection based algorithms. Numerical simulations clearly demonstrate the superior ability of the convex programming based methods, both in terms of successful recovery in the noiseless setting and stable reconstruction in the noisy setting.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

MOTIVATION

Phase retrieval is the problem of recovering a signal from the *magnitude* of its Fourier transform. This inverse problem has a rich history [Pat34; Pat44], motivated by applications such as X-ray crystallography [Mil90], optics [Wal63] and astronomy [FD87], where the measurable quantity is the magnitude-square of the Fourier transform of the signal of interest. In applications such as radar [GZW88] and blind channel estimation [Bay04; Ton+95], measuring the Fourier magnitude-square of the signal of interest is significantly easier than measuring the Fourier phase. In such settings, phase retrieval leads to simple and cost-effective measurement systems. In the rest of this chapter, we briefly describe the origin of phase retrieval in various applications.

1.1 X-ray Crystallography/ Coherent Diffraction Imaging

X-ray crystallography is a technique used to identify molecular and atomic structures of crystals. This method has been used to identify the structure and function of many basic molecules, including table salt [Bra13], DNA [W+53] and proteins [Dre07]. A typical experimental setup, courtesy of [260], is detailed in Fig. 1.1. A focused monochromatic X-ray beam is incident on the crystal whose structure one wishes to determine. The crystal causes the incident beam to diffract in a specific manner. By rotating the crystal, multiple two-dimensional diffraction patterns are recorded using photosensitive films or CCD cameras. A three-dimensional picture of the density of the electrons is then reconstructed from these measurements by solving an inverse problem.

Let $\psi(x, y, z)$ denote the three-dimensional electron density of the object, centered at the origin. Also, let the direction of light be parallel to the z-axis, the plane of the two-dimensional detector be perpendicular to the z-axis such that $z = z'$, and $I(x', y')$ denote the diffraction pattern collected at (x', y', z') for various (x', y').

The Huygens-Fresnel principle states that every point which a luminous disturbance reaches becomes a source of a secondary spherical wave, and that the sum of these secondary waves determines the wave at any subsequent time [Huy85]. Let $\psi_{trans}(x, y)$ denote the secondary source at $(x, y, 0)$ produced by the electron

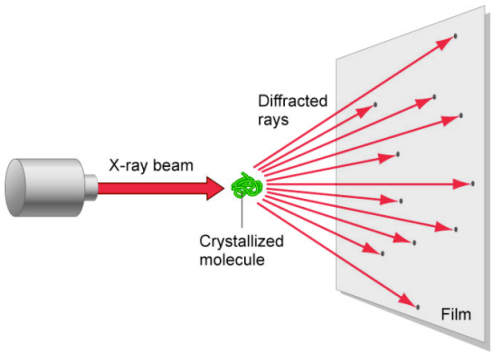


Figure 1.1: A typical X-ray Crystallography or Coherent Diffraction Imaging (CDI) setup (courtesy of [260]).

densities $\psi(x, y, z)$ for all z. The quantity $\psi_{trans}(x, y)$ is well-approximated by the line integral of $\psi(x, y, z)$ along the z direction [Goo05a], i.e.,

$$\psi_{trans}(x, y) = \int \psi(x, y, z)dz.$$

The wave at (x', y'), due to a unit point source at (x, y), is given by the scalar Green's function

$$\frac{e^{i\frac{2\pi}{\lambda}\sqrt{(x-x')^2+(y-y')^2+z'^2}}}{4\pi\sqrt{(x-x')^2+(y-y')^2+z'^2}}, \tag{1.1}$$

where λ is the wavelength [Goo05a]. Therefore, the wave at (x', y'), denoted by $\psi_{diff}(x', y')$, is such that

$$\psi_{diff}(x', y') \propto \iint \psi_{trans}(x, y)\frac{e^{i\frac{2\pi}{\lambda}\sqrt{(x-x')^2+(y-y')^2+z'^2}}}{\sqrt{(x-x')^2+(y-y')^2+z'^2}}dxdy.$$

The Fraunhofer approximation (also known as the far field approximation) involves the following steps: The $\sqrt{(x-x')^2+(y-y')^2+z'^2}$ term is approximated by $z' + \frac{x'^2+y'^2-(2xx'+2yy')}{2z'}$, which holds when $|z'| \gg |x-x'|$ and $|z'| \gg |y-y'|$ (detectors sufficiently far away). The $\sqrt{(x-x')^2+(y-y')^2+z'^2}$ term in the denominator is further approximated by z', which holds when $|z'| \gg |x-x'|$ and $|z'| \gg |y-y'|$ (object restricted to a small region). Consequently, the wave $\psi_{diff}(x', y')$ is Fraunhofer-approximated by

$$\psi_{diff}^{Fraunhofer}(x', y') \propto \frac{e^{i\frac{2\pi}{\lambda}z'}}{z'}e^{i\frac{\pi}{\lambda}\frac{x'^2+y'^2}{z'}}\iint \psi_{trans}(x, y)e^{i\frac{2\pi}{\lambda}\frac{-x'x-y'y}{z'}}dxdy.$$

Detection devices like CCD cameras and photosensitive films cannot measure the phase of the light wave, and instead measure the photon flux, which is proportional

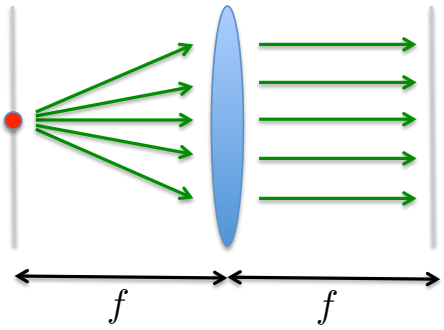


Figure 1.2: A picture depicting the Fourier transforming property of lenses.

to the intensity of the light wave. Therefore, the diffraction pattern measurements correspond to

$$\begin{aligned} I(x', y') &\propto \left| \iint \psi_{trans}(x, y) e^{i\frac{2\pi}{\lambda}\frac{-x'x-y'y}{z'}} dxdy \right|^2 \\ &\propto \left| \iiint \psi(x, y, z) e^{i\frac{2\pi}{\lambda}\frac{-x'x-y'y-0z'z}{z'}} dxdydz \right|^2 \\ &\propto \left| \hat{\psi}\left(\frac{x'}{\lambda z'}, \frac{y'}{\lambda z'}, 0\right) \right|^2, \end{aligned} \tag{1.2}$$

where $\hat{\psi}$ is the three-dimensional Fourier transform of ψ. Hence, the measurements correspond to the magnitude-square of the three-dimensional Fourier transform of the underlying signal, along a two-dimensional plane. By rotating the crystal, the magnitude-square along various two-dimensional planes of the three-dimensional Fourier transform are obtained. The electron density is then reconstructed by solving the phase retrieval problem.

1.2 Optics

The propagation of light through a lens is an essential part in many imaging systems. The phase retrieval problem arises in such setups due to the Fourier transforming property of lenses, i.e., if a transmissive object is placed one focal length in front of a lens, then its Fourier transform is formed one focal length behind the lens [Goo05b]. A pictorial representation of this property is provided in Fig. 1.2.

Let $\psi(x, y)$ denote the two dimensional object placed one focal length in front of the lens. As a consequence of Huygens principle, this is equivalent to placing a transmissive source $\psi(x, y)$ if the object is uniformly illuminated. The wave before

the lens, denoted by $\psi_-(x', y')$, is then given by

$$\psi_-(x', y') = \iint \psi(x, y) \frac{e^{i\frac{2\pi}{\lambda}\sqrt{(x-x')^2+(y-y')^2+f^2}}}{\sqrt{(x-x')^2+(y-y')^2+f^2}} dxdy,$$

using the steps described in the previous section (superposition principle, along with the scalar Green's function (1.1)).

The Fresnel approximation (also known as the near field approximation) involves the following steps: The $\sqrt{(x-x')^2+(y-y')^2+f^2}$ term is approximated by $f + \frac{x'^2+y'^2-(2xx'+2yy')+x^2+y^2}{2f}$, which holds when $|f| \gg |x-x'|$ and $|f| \gg |y-y'|$ (object restricted to a small region). The $\sqrt{(x-x')^2+(y-y')^2+f^2}$ term in the denominator is further approximated by f. Consequently, the Fresnel-approximated wave is given by

$$\psi_-^{Fresnel}(x', y') \propto e^{i\frac{\pi}{\lambda}\frac{x'^2+y'^2}{f}} \iint \psi(x, y) e^{i\frac{2\pi}{\lambda}\frac{-x'x-y'y}{f}} e^{i\frac{\pi}{\lambda}\frac{x^2+y^2}{f}} dxdy.$$

If the lens is thin, then the incoming wave at (x', y') leaves at (x', y'). Due to the fact that waves travel slower in a refractive medium when compared to free space, the wave at (x', y') undergoes a phase delay proportional to the thickness of the lens at (x', y'). The phase shift at (x', y') is calculated, using paraxial approximation [Goo05b], to be proportional to $e^{-i\frac{\pi}{\lambda}\frac{x'^2+y'^2}{f}}$. Therefore, the wave immediately after the lens is given by

$$\psi_+^{Fresnel}(x', y') \propto \iint \psi(x, y) e^{i\frac{2\pi}{\lambda}\frac{-x'x-y'y}{f}} e^{i\frac{\pi}{\lambda}\frac{x^2+y^2}{f}} dxdy.$$

The Fresnel-approximated wave at the detector is hence given by

$$\psi_{diff}^{Fresnel}(x'', y'') \propto e^{i\frac{\pi}{\lambda}\frac{x''^2+y''^2}{f}} \iint \psi_+^{Fresnel}(x', y') e^{i\frac{2\pi}{\lambda}\frac{-x''x'-y''y'}{f}} e^{i\frac{\pi}{\lambda}\frac{x'^2+y'^2}{f}} dx'dy',$$

which, upon substitution and integration with respect to x' and y', gives

$$\begin{aligned} \psi_{diff}^{Fresnel}(x'', y'') &\propto \iint \psi(x, y) e^{i\frac{2\pi}{\lambda}\frac{-x''x-y''y}{f}} dxdy \\ &\propto \hat{\psi}\left(\frac{x''}{\lambda f}, \frac{y''}{\lambda f}\right), \end{aligned} \tag{1.3}$$

where $\hat{\psi}$ is the two-dimensional Fourier transform of ψ. Hence, if photosensitive films or CCD cameras are used as detectors, then the reconstruction of the object involves solving the phase retrieval problem.

1.3 Astronomy

In optical astronomy, objects in space are imaged using a ground based telescope. Even at the best observation sites, the image resolution is typically limited by atmospheric turbulence. This is due to refractive index variations of the atmosphere [FD87].

Let $O(x, y)$ denote the object intensity one wishes to estimate. If $I(x, y)$ denotes the intensity measurements obtained using a telescope, then we have

$$I(x, y) = O(x, y) * |p(x, y)|^2,$$

where $|p(x, y)|^2$ is the point spread function introduced by the atmosphere [Har98]. In the spatial frequency domain, this relationship is equivalent to

$$\hat{I}(x', y') = \hat{O}(x', y')\hat{P}(x', y'),$$

where $\hat{I}(x', y')$, $\hat{O}(x', y')$ and $\hat{P}(x', y')$ are the spatial Fourier transforms of $I(x, y)$, $O(x, y)$ and $|p(x, y)|^2$ respectively.

It is well established that, when the measurements are taken at short exposures (in order to "freeze" the atmosphere), the atmospheric turbulence primarily affects the phase of $\hat{P}(x', y')$, and that $\left|\hat{P}(x', y')\right|^2$ obeys the same statistics across measurements with similar atmospheric conditions [Fri66]. A popular technique called speckle interferometry [Lab70] uses this fact to extract high spatial frequency information from such measurements. It involves collecting R measurements at short exposures under similar atmospheric conditions, so that we have

$$\hat{I}_r(x', y') = \hat{O}(x', y')\hat{P}_r(x', y') \quad \text{for} \quad r = 1, 2, \ldots, R.$$

Consequently, we have

$$\left(\frac{1}{R}\sum_{r=1}^{R}\left|\hat{I}_r(x', y')\right|^2\right) = \left|\hat{O}(x', y')\right|^2\left(\frac{1}{R}\sum_{r=1}^{R}\left|\hat{P}_r(x', y')\right|^2\right). \tag{1.4}$$

The term $\frac{1}{R}\sum_{r=1}^{R}\left|\hat{I}_r(x', y')\right|^2$ is calculated from the measurements and the term $\frac{1}{R}\sum_{r=1}^{R}\left|\hat{P}_r(x', y')\right|^2$ is reliably estimated by observing a point object under similar atmospheric conditions[1]. Therefore, the quantity $\left|\hat{O}(x', y')\right|^2$ is reliably obtained from these measurements. The object intensity $O(x, y)$ is then reconstructed by solving the phase retrieval problem.

[1]The average $\frac{1}{R}\sum_{r=1}^{R}\hat{I}_r(x', y')$ does not provide useful information due to the fact that $\hat{P}_r(x', y')$ has different statistics across measurements [Bat82].

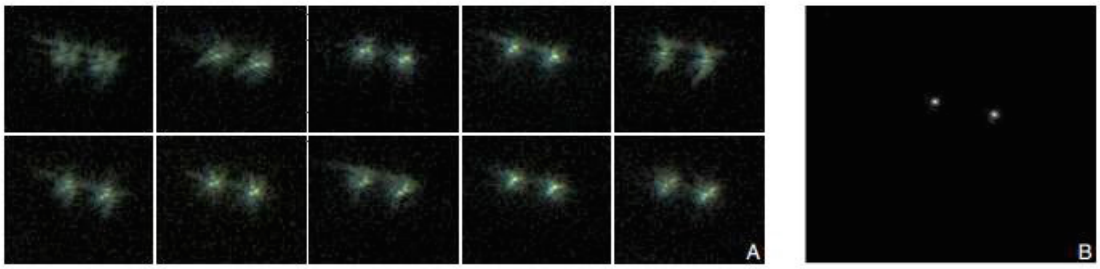


Figure 1.3: An example of the input and output data in speckle interferometry (courtesy of [Hir+11] and [Ran+13]). (A) A set of 10 low resolution speckle images. (B) The high resolution image of the stars is obtained through phase retrieval.

1.4 Direction of Arrival (DoA) Estimation

The need for estimating the direction of wave propagation arises in many applications, including radar [Zha+10], wireless communications [God97] and object tracking [RSZ94]. An active setup involves a transmitter which transmits narrow-band waves (with center frequency $\omega_c = \frac{2\pi c}{\lambda}$) and an array of, say M, receivers.

Consider the two-dimensional setup such that the transmitter and receivers are placed along the x-axis at the origin and $x = (\frac{\lambda}{2}, \frac{2\lambda}{2}, \ldots, \frac{M\lambda}{2})$ respectively, and the transmission is uniform in the positive y half of the two-dimensional space (see Fig. 1.4). Suppose there are K objects which reflect the transmitted wave, where the kth object is located at a distance r_k and an angle θ_k from the origin.

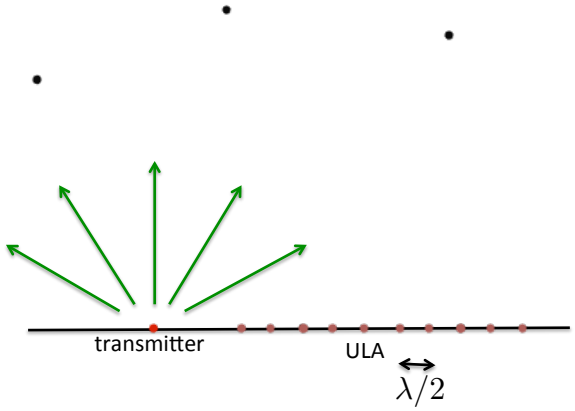


Figure 1.4: An active setup to estimate the position of objects in space (ULA = Uniform Linear Array).

If $s(t)$ denotes the base-band transmitted signal and $\mathbf{x}^{(t)} = (x^{(t)}[1], x^{(t)}[2], \ldots, x^{(t)}[M])^T$ denotes the $M \times 1$ narrow-band vector measured by the receivers at time t, then we

have

$$x^{(t)}[m] = \sum_{k=1}^{K} s\left(t - \frac{2r_k - \frac{m\lambda}{2}\sin\theta_k}{c}\right) e^{i\omega_c\left(t - \frac{2r_k - \frac{m\lambda}{2}\sin\theta_k}{c}\right)}\rho_k,$$

where ρ_k is a function of the reflectivity of the object and its distance from the transmitter [TF09]. Here, we use the fact that the total distance travelled by the wave reflected by the kth object onto the mth receiver is well-approximated by $2r_k - \frac{m\lambda}{2}\sin\theta_k$. Since $s(t)$ is slowly varying (base-band assumption), the quantity $s\left(t - \frac{2r_k - \frac{m\lambda}{2}\sin\theta_k}{c}\right)$ is approximated by $s(t)$. In the frequency domain, this leads to the following relationship:

$$\begin{aligned} y^{(\omega)}[m] &= \sum_{k=1}^{K} \hat{s}(\omega - \omega_c) e^{-i\omega_c \frac{2r_k - \frac{m\lambda}{2}\sin\theta_k}{c}} \rho_k \\ &\propto \sum_{k=1}^{K} e^{i\pi m \sin\theta_k} \left(\rho_k e^{-\frac{i2\omega_c r_k}{c}}\right), \end{aligned} \tag{1.5}$$

where $\hat{s}(\omega)$ is the Fourier transform of $s(t)$. Therefore, the vector $\mathbf{y}$ corresponds to the M low-frequency terms of the Fourier transform of a signal which has an amplitude $\rho_k e^{-\frac{i2\omega_c r_k}{c}}$ at location $\frac{\sin\theta_k}{2}$, for $1 \le k \le K$. The inverse problem of recovering the various θ_k from $\mathbf{y}$ is referred to as direction of arrival estimation (also commonly known as super-resolution). Classic algorithms to solve this problem include MUSIC [Sch86] and ESPRIT [RK89].

This setup requires coherent detection, i.e., the receivers must be perfectly synchronized and be able to measure the phase of the incoming wave accurately. In practice, this is very difficult to achieve, particularly when the number of receivers is large. The measurements, due to such errors, are of the form

$$y[m] \propto e^{i\phi_m} \sum_{k=1}^{K} e^{i\pi m \sin\theta_k} \left(\rho_k e^{-\frac{i2\omega_c r_k}{c}}\right),$$

for some unknown ϕ_m. A potential approach to overcome this issue is to discard the phase measurements and only consider the magnitude measurements, i.e.,

$$Z[m] \propto \left| \sum_{k=1}^{K} e^{i\pi m \sin\theta_k} \left(\rho_k e^{-\frac{i2\omega_c r_k}{c}}\right) \right|^2.$$

This inverse problem is a combination of phase retrieval and super-resolution. We refer to this problem as *phaseless super-resolution* [Jag+16].

Chapter 2

INTRODUCTION

In this chapter, we mathematically set up the phase retrieval problem, and provide an overview of the classic and the modern approaches. For the sake of exposition, we consider the discretized 1D setting[1]. Let $\mathbf{x} = (x[0], x[1], \ldots, x[N-1])^T$ be a signal of length N. Denote by $\mathbf{y} = (y[0], y[1], \ldots, y[N-1])^T$ its N point Discrete Fourier Transform (DFT) and let $\mathbf{Z} = (Z[0], Z[1], \ldots, Z[N-1])^T$ be the Fourier magnitude-square measurements (i.e., $Z[m] = |y[m]|^2$). Phase retrieval is the following recovery problem:

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && Z[m] = |\langle \mathbf{f}_m, \mathbf{x} \rangle|^2 \quad \text{for} \quad 0 \le m \le N-1, \end{aligned} \tag{2.1}$$

where $\mathbf{f}_m$ is the conjugate of the mth column of the N point DFT matrix, with elements $\{e^{i2\pi \frac{mn}{N}}\}_{n=0}^{N-1}$, and $\langle ., . \rangle$ is the standard inner product operator. Since Fourier magnitude-square (i.e., power spectral density) and circular autocorrelation are Fourier pairs, phase retrieval can also be equivalently stated as the problem of recovering a signal from its circular autocorrelation, denoted by $\mathbf{b} = (b[0], b[1], \ldots, b[N-1])^T$, i.e.,

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && b[m] = \sum_{n=0}^{N-1} x[n] x^\star[(n+m) \bmod N] \quad \text{for} \quad 0 \le m \le N-1. \end{aligned} \tag{2.2}$$

2.1 Uniqueness

Due to the absence of Fourier phase information, the available data is highly incomplete. For any given Fourier magnitude, the Fourier phase can be chosen from an N-dimensional set. Since distinct phases correspond to different signals in general, the feasible set of (2.1) is an N-dimensional manifold, rendering phase retrieval a very ill-posed problem.

In fact, it is well known that the Fourier phase quite often contains more information than the Fourier magnitude. To demonstrate this fact, a synthetic example is provided

[1] The theory and algorithms developed in this work generalize to higher dimensions. We provide more details in the appropriate sections.

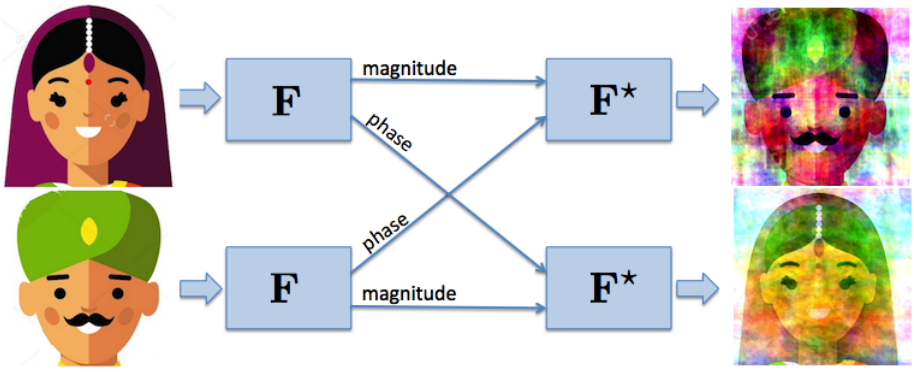


Figure 2.1: A synthetic example demonstrating the importance of Fourier phase in reconstructing a signal from its Fourier transform.

in Fig. 2.1. The figure shows the result of the following numerical simulation: Two images (of Alisha and Babu[2]) are Fourier transformed, their Fourier phases are swapped and then they are inverse Fourier transformed. The result clearly demonstrates the importance of Fourier phase.

A popular approach to mitigate the ill-posedness of phase retrieval is to use an $M > N$ point DFT. In practice, this is done by increasing the density of the detectors. A typical choice is $M = 2N$. This setting is mathematically equivalent to zero-padding the signal $\mathbf{x}$ with N zeros, and considering the $2N$ point DFT of $(x[0], x[1], \ldots, x[N-1], 0, 0, \ldots, 0)^T$. The term *oversampling* is used to refer to this setting.

Phase retrieval with oversampling can therefore be stated as the problem of recovering a signal from its autocorrelation, denoted by $\mathbf{a} = (a[0], a[1], \ldots, a[N-1])^T$, i.e.,

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && a[m] = \sum_{n=0}^{N-1-m} x[n]x^\star[n+m] \quad \text{for} \quad 0 \le m \le N-1. \end{aligned} \tag{2.3}$$

Remark: If $M \ge 2N$, then the inverse problem is equivalent to (2.3) irrespective of the value of M. Hence, increasing the density of the detectors does not help beyond a certain point.

[2]Alisha and Babu are the Indian counterparts of Alice and Bob respectively.

Trivial Ambiguities

Observe that the operations of time-shift, conjugate-flip and global phase-change on the signal do not affect the autocorrelation, or equivalently, the oversampled DFT magnitude. Indeed, if $\mathbf{y} = (y[0], y[1], \ldots, y[M-1])^T$ is the oversampled DFT of $\mathbf{x}$, then $\mathbf{y} = (y[0], e^{in_0} y[1], \ldots, e^{in_0(M-1)} y[M-1])^T$ is the oversampled DFT of $\mathbf{x}$ time-shifted by n_0 units, $\mathbf{y} = (y^\star[0], y^\star[1], \ldots, y^\star[M-1])^T$ is the oversampled DFT of the conjugate-flip of $\mathbf{x}$, and $e^{i\phi}\mathbf{y}$ is the oversampled DFT of $e^{i\phi}\mathbf{x}$. Each of these operations only affect the phase of the oversampled DFT.

Hence, a signal can only be reconstructed up to a time-shift, conjugate-flip and global phase without additional information. These ambiguities are referred to as *trivial ambiguities*, and signals obtained by these operations are considered to be *equivalent*. In most applications of phase retrieval, it is good enough if a signal is reconstructed up to these ambiguities. For example, in astronomy, where the underlying signal corresponds to stars in the sky, or in X-ray crystallography, where the underlying signal corresponds to atoms or molecules in a crystal, equivalent solutions are equally informative [Mil90].

In order to calculate the number of non-equivalent solutions to (2.3), we rewrite the equations in the z-transform domain. We have

$$A(z) = X(z)X^\star(z^{-\star}), \tag{2.4}$$

where $A(z)$ and $X(z)$ are the z-transforms of $\mathbf{a}$ and $\mathbf{x}$ respectively. Since $A(z) = A^\star(z^{-\star})$, if z_0 is a zero of $A(z)$, then $z_0^{-\star}$ is also a zero. Hence, the zeros of $A(z)$ appear in pairs of the form $(z_0, z_0^{-\star})$. The reconstruction of $\mathbf{x}$ from $\mathbf{a}$, or equivalently $X(z)$ from $A(z)$, is known as spectral factorization, and deals with the distribution of these pairs of zeros between $X(z)$ and $X^\star(z^{-\star})$.

The trivial ambiguities can be understood in this framework as follows: The z transform of $\mathbf{x}$ time-shifted by n_0 units is $X(z)z^{n_0}$. Consequently, the z transform of its autocorrelation is given by $X(z)z^{n_0} \times z^{-n_0}X^\star(z^{-\star}) = X(z)X^\star(z^{-\star})$. The z transform of the conjugate-flip of $\mathbf{x}$ is $X^\star(z^{-\star})$, due to which the z transform of its autocorrelation is given by $X^\star(z^{-\star})X(z)$. Indeed, this solution corresponds to "wrongly" assigning the zeros in every pair of zeros. The z transform of $e^{i\phi}\mathbf{x}$ is $e^{i\phi}X(z)$. Therefore, the z transform of its autocorrelation is $e^{i\phi}X(z) \times e^{-i\phi}X^\star(z^{-\star}) = X(z)X^\star(z^{-\star})$.

1D setting: Since $X(z)$ is a univariate polynomial of degree $N-1$, it has $N-1$ zeros (fundamental theorem of algebra [FR12]), denoted by r_n for $1 \leq n \leq N-1$.

Consequently, $A(z)$ has $N-1$ pairs of zeros $(r_n, r_n^{-\star})$. For every pair $(r_n, r_n^{-\star})$, we can either assign r_n to $X(z)$ and $r_n^{-\star}$ to $X^\star(z^{-\star})$, or assign $r_n^{-\star}$ to $X(z)$ and r_n to $X^\star(z^{-\star})$. Hence, the total number of non-equivalent solutions is at most 2^{N-1}. If the zeros of $X(z)$ are distinct, then the number of non-equivalent solutions is exactly 2^{N-1}. While this is a significant improvement when compared to the number of non-equivalent solutions of (2.2), 2^{N-1} is still a prohibitive number, due to which phase retrieval with oversampling in 1D remains ill-posed. Additional assumptions on the signal are required in order to be able to guarantee unique reconstruction.

≥2D setting: Here, $X(z_1, z_2, \ldots, z_d)$ is a multivariate polynomial. In [HM82], it is shown that *almost all* polynomials in two or more variables are irreducible. Hence, in theory, almost all signals can be uniquely reconstructed, up to trivial ambiguities, by factorizing the polynomial $A(z_1, z_2, \ldots, z_d)$. Consequently, with the exception of a set of signals of measure zero, phase retrieval in ≥2D with oversampling is a well-posed problem.

2.2 Classic Approaches

Earlier approaches to phase retrieval were based on the method of alternating projections, pioneered by the work of Gerchberg and Saxton [GS72]. The phase retrieval problem (with oversampling, i.e., $M = 2N$) is reformulated as the following least-squares problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \sum_{m=0}^{2N-1} \left(\sqrt{Z[m]} - |\langle \mathbf{f}_m, \mathbf{x} \rangle| \right)^2 \\ \text{subject to} \quad & x[n] = 0 \quad \text{for} \quad N \le n \le 2N-1. \end{aligned} \tag{2.5}$$

Here, $\mathbf{f}_m$ is the conjugate of the mth column of the $2N$ point DFT matrix, with elements $e^{i2\pi \frac{mn}{2N}}$. The underlying signal has nonzero values only within the interval $[0, N-1]$, and has a value 0 outside this interval, i.e., in the interval $[N, 2N-1]$.

The Gerchberg-Saxton (GS) algorithm attempts to minimize this non-convex objective by starting with a random initialization and iteratively imposing the time domain constraints (for example, nonzero values only within the interval $[0, N-1]$) and Fourier domain constraints (given Fourier magnitude measurements) using projections. The details of the various steps are provided in Algorithm 1.

The intuition behind the algorithm is the following: The underlying signal is known to be in $\mathcal{S}_1 \cap \mathcal{S}_2$, where $\mathcal{S}_1$ is the set of all signals which satisfy the time domain constraints, and $\mathcal{S}_2$ is the set of all signals which satisfy the Fourier magnitude

Algorithm 1 Gerchberg-Saxton (GS) Algorithm

Input: Fourier magnitude-square measurements $\mathbf{Z}$
Output: Estimate $\hat{\mathbf{x}}$ of the underlying signal
Initialize: Choose a random input signal $\mathbf{x}^{(0)}$, $\ell = 0$
while halting criterion false **do**
  $\ell \leftarrow \ell + 1$
  Compute the DFT of $\mathbf{x}^{(\ell-1)}$: $\mathbf{y}^{(\ell)} = \mathbf{F}\mathbf{x}^{(\ell-1)}$
  Impose Fourier magnitude constraints: $y'^{(\ell)}[m] = \frac{y^{(\ell)}[m]}{|y^{(\ell)}[m]|}\sqrt{Z[m]}$
  Compute the inverse DFT of $\mathbf{y}'^{(\ell)}$: $\mathbf{x}'^{(\ell)} = \mathbf{F}^{-1}\mathbf{y}'^{(\ell)}$
  Impose time domain constraints to obtain $\mathbf{x}^{(\ell)}$
end while
return $\hat{\mathbf{x}} \leftarrow \mathbf{x}^{(\ell)}$

measurements. From any signal, it is typically straightforward to calculate the projection onto $\mathcal{S}_1$ or $\mathcal{S}_2$. If $\mathcal{S}_1$ is the set of all signals which have nonzero values only within the interval $[0, N-1]$, then the projection onto this set is obtained by forcing the values outside this interval to 0. The projection onto $\mathcal{S}_2$ is the signal obtained by calculating the Fourier transform, replacing the magnitude with the measured magnitude, and taking the inverse Fourier transform.

If the sets $\mathcal{S}_1$ and $\mathcal{S}_2$ are both convex, then the method of alternating projections always converges to a signal which lies in $\mathcal{S}_1 \cap \mathcal{S}_2$ (assuming this set is not a null set). In the phase retrieval setup, since $\mathcal{S}_2$ is non-convex, this method has limited abilities. The objective function value is shown to be non-increasing with each iteration, due to which the algorithm always converges:

$$\begin{aligned}
\sum_{m=0}^{2N-1}\left(\sqrt{Z[m]}-\left|\left\langle \mathbf{f}_m, \mathbf{x}^{(\ell-1)}\right\rangle\right|\right)^2 &= \sum_{m=0}^{2N-1}\left(\left|\left\langle \mathbf{f}_m, \mathbf{x}'^{(\ell)}\right\rangle\right|-\left|\left\langle \mathbf{f}_m, \mathbf{x}^{(\ell-1)}\right\rangle\right|\right)^2 \\
&= \sum_{m=0}^{2N-1}\left|\left\langle \mathbf{f}_m, \mathbf{x}'^{(\ell)}\right\rangle - \left\langle \mathbf{f}_m, \mathbf{x}^{(\ell-1)}\right\rangle\right|^2 = \|\mathbf{x}'^{(\ell)} - \mathbf{x}^{(\ell-1)}\|_F^2 \\
&\geq \|\mathbf{x}'^{(\ell)} - \mathbf{x}^{(\ell)}\|_F^2 = \sum_{m=0}^{2N-1}\left|\left\langle \mathbf{f}_m, \mathbf{x}'^{(\ell)}\right\rangle - \left\langle \mathbf{f}_m, \mathbf{x}^{(\ell)}\right\rangle\right|^2 \\
&\geq \sum_{m=0}^{2N-1}\left(\left|\left\langle \mathbf{f}_m, \mathbf{x}'^{(\ell)}\right\rangle\right|-\left|\left\langle \mathbf{f}_m, \mathbf{x}^{(\ell)}\right\rangle\right|\right)^2 \\
&= \sum_{m=0}^{2N-1}\left(\sqrt{Z[m]}-\left|\left\langle \mathbf{f}_m, \mathbf{x}^{(\ell)}\right\rangle\right|\right)^2,
\end{aligned}$$

due to the fact that $\mathbf{x}'^{(l)}$ has the same Fourier magnitude as the measurements, $\mathbf{x}'^{(l)}$

and $\mathbf{x}^{(l-1)}$ have the same Fourier phase, Parseval's theorem, and $\mathbf{x}'^{(l)}$ is closer to $\mathbf{x}^{(l)}$ when compared to $\mathbf{x}^{(l-1)}$.

The converged signal is often a local minimizer of the objective function, due to the fact that $\mathcal{S}_2$ is non-convex. In order to mitigate this issue, Fienup, in his seminal work [Fie82], extended this method by introducing additional correction terms to the time domain step (see Hybrid Input-Output (HIO) algorithm [Fie82] for details). The HIO algorithm is not guaranteed to converge, and when it does converge, it may be to a local minimum. We refer the readers to [BCL02] and [Mar07] for a theoretical and numerical investigation of such methods, and to [Fie82] for a survey of classic approaches.

2.3 Modern Approaches

The classic algorithms have limited recovery abilities, and do not have theoretical recovery guarantees. Due to these reasons, phase retrieval is still an active research problem. Recent developments in measurement technologies and advances in optimization methods have inspired a host of new approaches to phase retrieval. The modern approaches can be broadly classified into three categories:

(i) *Additional prior information*: Inspired by results in the area of compressed sensing [CT05; EK12; Cha+12; Tro15], various researchers have explored the idea of sparsity as a prior information on the signal. A signal of length N is said to be k-sparse if it has k locations with nonzero values and $k \ll N$. The exact locations and values of the nonzero elements are not known a priori. The approach has been to develop conditions under which only one sparse signal satisfies the autocorrelation measurements, and to develop algorithms which exploit the sparsity prior.

(ii) *Additional magnitude-only measurements*: Technological advances have enabled the possibility of obtaining additional information about the signal. In particular, magnitude-square measurements of the form

$$Z[m] = \left|\langle \mathbf{f}_m, \mathbf{D}\mathbf{x}\rangle\right|^2 \tag{2.6}$$

can be obtained in many phase retrieval applications, where $\mathbf{D}$ is an $N \times N$ diagonal matrix. This can be done in practice in various ways, depending on the application. Common approaches include the use of masks [Joh+08], optical gratings [LP97] and structured illuminations [Far+10]. The idea is to overcome the uniqueness and algorithmic issues of phase retrieval by obtaining measurements from multiple carefully designed diagonal matrices.

(iii) *Random phaseless measurements*: A popular trend for analysis purposes is to replace the Fourier vectors with random vectors. The measurements considered are of the form

$$Z[m] = \left|\langle \mathbf{a}_m, \mathbf{x}\rangle\right|^2, \tag{2.7}$$

where $\mathbf{a}_m$ is a generic measurement vector. A natural question to ask is how many and which measurement vectors can uniquely identify the underlying signal. Another interesting problem is to identify a set of measurement vectors for which there is an efficient and stable reconstruction algorithm. Since our work focuses on Fourier vectors which naturally come up in many applications, we do not pursue this line of work. We refer the interested readers to [BCE06; Bal+09; CSV13; LV13; NJS13; EM14; Ale+14; BR15; CLS15b; Oym+15; PLR14; SR15; Tro15] for details.

Semidefinite Programming (SDP)

On the algorithmic front, one of the recent popular approaches to treat phase retrieval problems is to use semidefinite programming methods. SDP algorithms have been shown to yield robust solutions to various quadratic-constrained optimization problems (see [Lov79; GW95] and references therein). Since phase retrieval results in quadratic constraints, it is natural to use SDP techniques to try and solve such problems. An SDP formulation of phase retrieval (2.1) can be obtained by a procedure popularly known as *lifting*: We embed $\mathbf{x}$ in a higher dimensional space using the transformation $\mathbf{X} = \mathbf{x}\mathbf{x}^\star$. The Fourier magnitude measurements are then linear in the matrix $\mathbf{X}$:

$$Z[m] = \left|\langle \mathbf{f}_m, \mathbf{x}\rangle\right|^2 = \mathbf{x}^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{x} = \operatorname{trace}(\mathbf{f}_m \mathbf{f}_m^\star \mathbf{x}\mathbf{x}^\star) = \operatorname{trace}(\mathbf{f}_m \mathbf{f}_m^\star \mathbf{X}).$$

Consequently, phase retrieval reduces to finding a rank one positive semidefinite matrix $\mathbf{X}$ which satisfies these affine measurement constraints, leading to the following reformulation:

$$\begin{aligned} \text{minimize} \quad & \operatorname{rank}(\mathbf{X}) \\ \text{subject to} \quad & Z[m] = \operatorname{trace}(\mathbf{f}_m \mathbf{f}_m^\star \mathbf{X}) \quad \text{for} \quad 0 \le m \le N-1 \\ & \mathbf{X} \succcurlyeq 0. \end{aligned}$$

However, rank minimization is known to be NP-hard in general. To obtain an SDP algorithm, one possibility is to replace $\operatorname{rank}(\mathbf{X})$ by a convex surrogate $\operatorname{trace}(\mathbf{X})$

[RFP10], resulting in the following convex program:

$$\begin{aligned} &\text{minimize} && \operatorname{trace}(\mathbf{X}) \\ &\text{subject to} && Z[m] = \operatorname{trace}(\mathbf{f}_m \mathbf{f}_m^\star \mathbf{X}) \quad \text{for} \quad 0 \le m \le N-1 \\ & && \mathbf{X} \succcurlyeq 0. \end{aligned} \tag{2.8}$$

If the underlying signal is known to be sparse, then one could add an $\|\mathbf{X}\|_1$ cost to the objective function [CT05]. Measurements of the form (2.6) will appear as linear constraints of the form $Z[m] = \operatorname{trace}(\mathbf{D}^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{D}\mathbf{X})$. The approach has been to develop conditions under which $\mathbf{x}_0 \mathbf{x}_0^\star$, where $\mathbf{x}_0$ is the underlying signal, is the unique optimizer of (2.8). We refer the readers to [She+15; JEH15a] for an overview of contemporary methods.

2.4 Organization

The rest of this work is organized as follows: In Chapter 3, we motivate sparse phase retrieval, which is the problem of recovering a sparse signal from its Fourier magnitude. We first give conditions, which when satisfied, allow unique reconstruction. Then, we develop an SDP based reconstruction algorithm (TSPR), and provide theoretical guarantees. Chapter 4 considers phase retrieval using masks. We propose two simple mask designs, and show that the SDP method provably reconstructs most signals when measurements are obtained using these masks. In Chapter 5, we study Short-Time Fourier Transform (STFT) phase retrieval in which the measurements correspond to the STFT magnitude. We give conditions under which signals can be uniquely reconstructed, and also provide theoretical guarantees for reconstruction using the SDP method (STliFT). In Chapter 6, we consider phaseless super-resolution, which is the combination of phase retrieval and super-resolution. We propose a simple structured illumination design, and show that the SDP method provably reconstructs most signals using the information from such measurements. Chapter 7 concludes the work with a summary and discussion on future directions.

Chapter 3

SPARSE PHASE RETRIEVAL

In many phase retrieval applications, the signal of interest is naturally sparse. For example, electron microscopy deals with sparsely distributed atoms or molecules [Mil90], while astronomical imaging tends to consider sparsely distributed stars [FD87]. If it is known a priori that the signal of interest is sparse, then one could potentially solve for the sparsest solution satisfying the Fourier magnitude measurements, and be able to uniquely and efficiently identify the underlying signal up to trivial ambiguities (the trivial ambiguities cannot be resolved with a sparsity prior). Sparse phase retrieval can be mathematically written as

$$\begin{aligned} &\text{minimize} && \|\mathbf{x}\|_0 \\ &\text{subject to} && Z[m] = |\langle \mathbf{f}_m, \mathbf{x}\rangle|^2 \quad \text{for} \quad 0 \le m \le M-1, \end{aligned} \tag{3.1}$$

where $\|.\|_0$ is the ℓ_0 norm which counts the number of nonzero entries of its argument, and M is the size of the DFT. When $M = 2N$, sparse phase retrieval is equivalent to the problem of recovering a sparse signal from its autocorrelation, i.e.,

$$\begin{aligned} &\text{minimize} && \|\mathbf{x}\|_0 \\ &\text{subject to} && a[m] = \sum_{n=0}^{N-1-m} x[n]x^\star[n+m] \quad \text{for} \quad 0 \le m \le N-1. \end{aligned} \tag{3.2}$$

3.1 Contributions

In this chapter, we first show that almost all signals with aperiodic support (defined in Section 3.2) can, in theory, be uniquely recovered by solving (3.2). In other words, if the signal of interest is known to have aperiodic support, then we show that the sparse phase retrieval problem is almost surely well-posed.

We then develop the TSPR algorithm to *efficiently* solve (3.2), and provide the following recovery guarantees: (i) Most $O(N^{\frac{1}{2}-\epsilon})$-sparse signals can be recovered uniquely by TSPR. (ii) Most $O(N^{\frac{1}{4}-\epsilon})$-sparse signals can be recovered robustly by TSPR when the measurements are corrupted by additive noise. Numerical simulations complement our theoretical analysis, and show that TSPR can perform better than alternating projection methods, and as good as the other popular sparse phase retrieval algorithms (which enjoy empirical success, but do not have theoretical guarantees).

Related Work

In [Ran+13], it is shown that the knowledge of the autocorrelation is sufficient to uniquely identify 1D sparse signals if the autocorrelation is "collision free", as long as the sparsity $k \neq 6$. A signal $\mathbf{x}$ is said to have a collision free autocorrelation if for all indices $\{i_1, i_2, i_3, i_4\}$ such that $\{x[i_1], x[i_2], x[i_3], x[i_4]\} \neq 0$, we have $|i_1 - i_2| \neq |i_3 - i_4|$. In words, a signal is said to have a collision free autocorrelation if no two pairs of locations with nonzero values in the signal are separated by the same distance. For higher dimensions, the authors show that the requirement $k \neq 6$ is not necessary. This result has been further refined in [OE14], where it is shown that $k^2 - k + 1$ Fourier magnitude measurements are sufficient to recover the autocorrelation.

We would like to note that the collision-free property generically holds only for $O(N^{\frac{1}{4}-\epsilon})$-sparse signals, whereas our uniqueness results apply for $(N-1)$-sparse signals. To the best of our knowledge, TSPR is the first efficient sparse phase retrieval algorithm with strong theoretical guarantees.

3.2 Uniqueness

In this section, we present our identifiability results for the sparse phase retrieval problem (3.2).

Definition: A signal is said to have periodic or aperiodic support if the locations of its nonzero components are uniformly spaced or not uniformly spaced respectively.

For example: Consider the signal $\mathbf{x} = (x[0], x[1], x[2], x[3], x[4])$ of length $N = 5$.

(i) Aperiodic support: $\{n | x[n] \neq 0\} = \{0, 1, 3\}, \{1, 2, 4\}$.

(ii) Periodic support: $\{n | x[n] \neq 0\} = \{0, 2, 4\}, \{0, 1, 2, 3, 4\}$.

We prove the following result:

Theorem 3.2.1. *Let $\mathcal{S}_k$ represent the set of all k-sparse signals of length n with aperiodic support, where $3 \leq k \leq n-1$. Almost all signals in $\mathcal{S}_k$ can be uniquely recovered by solving (3.2).*

Proof. The proof technique we use is popularly known in literature as *dimension counting*. Since $\mathcal{S}_k$ represents the set of all k-sparse signals with aperiodic support, it is a manifold with $2k$ degrees of freedom (each nonzero location has 2 degrees of freedom, as the value can be complex). We show that the set of signals in $\mathcal{S}_k$ which

cannot be uniquely recovered by solving (3.2) is a manifold with degrees of freedom less than or equal to $2k-1$ and hence, almost all signals in $\mathcal{S}_k$ can be uniquely recovered by solving (3.2). The details are provided in Appendix 8.1. □

Signals with sparsity $k \leq 2$ can always be recovered by solving (3.2) (the quadratic system of equations can be solved trivially).

Remark: Sparse signals with periodic support can be viewed as an oversampled version of a signal which is not sparse. The sparse phase retrieval problem (3.2) reduces to the phase retrieval problem (2.3), and hence these signals cannot be uniquely recovered from their autocorrelation without further assumptions. For a detailed discussion, we refer the readers to Section II in [LV11].

3.3 Two-stage Sparse Phase Retrieval (TSPR)

In this section, we discuss the drawbacks of the standard approaches to solve (3.2) and then develop TSPR [JOH13b].

The Fienup HIO algorithm has been extended to solve sparse phase retrieval by adapting the step involving time domain constraints to promote sparsity. This can be achieved in several ways. For example, the locations with absolute values less than a particular threshold may be set to zero. Alternatively, the k locations with the highest absolute values can be retained and the rest set to zero [MS12]. In the noiseless setting, the sparsity constraint partially alleviates the convergence issues if multiple random initializations are considered and the underlying signals are sufficiently sparse. However, in the noisy setting, convergence issues still remain.

In [SBE14; SBE12], a sparse optimization based greedy search method called GESPAR (GrEedy Sparse PhAse Retrieval) is proposed. Sparse phase retrieval is reformulated as the following sparsity-constrained least-squares problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \sum_{m=0}^{M-1} \left(Z[m] - \left|\langle \mathbf{f}_m, \mathbf{x} \rangle\right|^2 \right)^2 \\ \text{subject to} \quad & \|\mathbf{x}\|_0 \leq k. \end{aligned} \tag{3.3}$$

GESPAR is a local search method, based on iteratively updating the signal support, and seeking a vector that corresponds to the measurements under the current support. A location-search method is repeatedly invoked, beginning with an initial random support set. Then, at each iteration, a swap is performed between a support and a non-support index. Only two elements are changed in the swap (one in the

support and one in the non-support), following the so-called 2-opt method [PS82]. Given the support of the signal, phase retrieval is then treated as a non-convex optimization problem, and approximated using the damped Gauss-Newton method [Ber99]. While the algorithm enjoys empirical success, there are no theoretical guarantees.

Since the solution we desire is both sparse and low rank, a natural convex approach would be to solve:

$$\begin{aligned} &\text{minimize} && \operatorname{trace}(\mathbf{X}) + \lambda \|\mathbf{X}\|_1 \\ &\text{subject to} && a[m] = \operatorname{trace}(\mathbf{A}_m \mathbf{X}) \quad \text{for} \quad 0 \le m \le N-1 \\ & && \mathbf{X} \succcurlyeq 0, \end{aligned} \tag{3.4}$$

for some regularizer λ, where the matrices $\mathbf{A}_m$ are given by

$$\mathbf{A}_{mgh} = \begin{cases} 1 & \text{if} \quad |h-g| = m = 0 \\ \frac{1}{2} & \text{if} \quad |h-g| = m \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

However, this approach does not work, as the issue of trivial ambiguities (due to time-shift and conjugate-flip) is still unresolved. If $\mathbf{X}_0 = \mathbf{x}_0 \mathbf{x}_0^\star$ is the desired sparse solution, then $\tilde{\mathbf{X}}_0 = \tilde{\mathbf{x}}_0 \tilde{\mathbf{x}}_0^\star$, where $\tilde{\mathbf{x}}_0$ is the conjugate-flipped version of $\mathbf{x}_0$, $\mathbf{X}_i = \mathbf{x}_i \mathbf{x}_i^\star$, where $\mathbf{x}_i$ is the signal obtained by time-shifting $\mathbf{x}_0$ by i units, and $\tilde{\mathbf{X}}_i = \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\star$, where $\tilde{\mathbf{x}}_i$ is the signal obtained by time-shifting $\tilde{\mathbf{x}}_0$ by i units are also feasible with the same objective value as $\mathbf{X}_0$. Since (3.4) is a convex program, any convex combination of these solutions is also feasible and has an objective value less than or equal to that of $\mathbf{X}_0$, because of which the optimizer is neither sparse nor rank one. One approach to break this symmetry would be to solve a weighted ℓ_1 minimization problem, which can potentially introduce a bias towards a particular equivalent solution. Numerical simulations suggest that this approach does not help in the sparse phase retrieval setup.

Many iterative heuristics have been proposed to solve (3.4). In [Can+15], the log-det function is used as a surrogate for rank (see [FHB03]). In [Sza+12], the solution space is iteratively reduced by calculating bounds on the support of the signal. Reweighted minimization (see [CWB08]) is explored in [JOH12b; JOH13a], where the weights are chosen based on the solution of the previous iteration. While these methods enjoy empirical success, no theoretical guarantees are available.

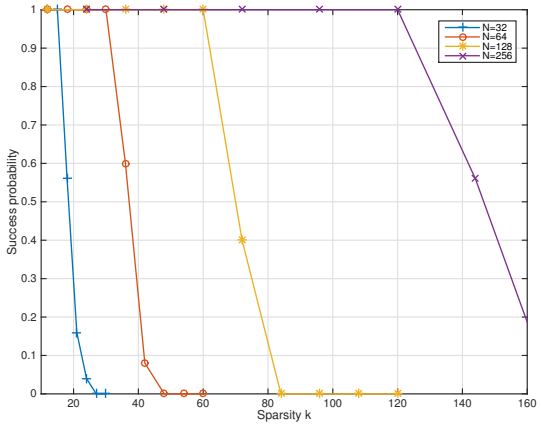


Figure 3.1: Probability of successful signal recovery of (3.6) (with $\lambda = 0$) for various sparsities for $N = 32, 64, 128, 256$.

The time-shift and time-reversal ambiguities stem from the fact that the support of the signal is not known. Therefore, let us momentarily assume that we somehow know the support of the signal (denoted from now on by V, which is the set of locations of the nonzero components of $\mathbf{x}$), (3.4) can be reformulated as

$$\begin{aligned} &\text{minimize} && \operatorname{trace}(\mathbf{X}) + \lambda \|\mathbf{X}\|_1 \\ &\text{subject to} && a[m] = \operatorname{trace}(\mathbf{A}_m \mathbf{X}) \quad \text{for} \quad 0 \le m \le N-1 \\ &&& X[n,m] = 0 \quad \text{if} \quad n, m \notin V \\ &&& \mathbf{X} \succcurlyeq 0. \end{aligned} \tag{3.6}$$

Fig. 3.1 plots the probability of successful recovery of (3.6) (with $\lambda = 0$) against various sparsities k for $N = 32, 64, 128, 256$. For a given signal length N and sparsity k, the k nonzero locations were chosen uniformly at random and the signal values in the support were chosen from an i.i.d. standard normal distribution. It can be observed that (3.6) recovers the signal with very high probability in the $k \le \frac{N}{2}$ regime[1]. This observation suggests a two-stage algorithm: one where we first recover the support of the signal and then use it to solve (3.6).

[1]This is an empirical observation. In this work, we provide recovery guarantees only for $O(N^{\frac{1}{2}-\epsilon})$-sparse signals.

Algorithm 2 Two-stage Sparse Phase Retrieval (TSPR)

Input: Autocorrelation **a** of the signal of interest
Output: Sparse signal **x** which has an autocorrelation **a**

(i) Recover V using Algorithm 3

(ii) Recover **x** by solving (3.6) with $\lambda = 0$.

It is difficult to characterize the set of signals that can be reconstructed using TSPR. In order to provide recovery guarantees, we consider a probabilistic approach. In particular, we assume that the sparse signal is drawn from the Bernoulli-Gaussian distribution $\mathcal{BN}(N, \theta)$, defined as follows:

(i) Support is chosen from an i.i.d. $Bern\left(\frac{N^\theta}{N}\right)$ distribution

(ii) Signal values in the support are chosen from an i.i.d. $\mathcal{CN}(0, 1)$ distribution.

We prove the following result:

Theorem 3.3.1. *If sparse signals are drawn from the $\mathcal{BN}(N, \theta)$ distribution, where the parameter θ satisfies $0 < \theta \leq \frac{1}{2} - \epsilon$ for some constant $\epsilon > 0$, then the failure probability of TSPR is $O(N^{-0.1\epsilon})$.*

Proof. This is a direct consequence of Theorem 3.3.2 and 3.3.3. □

For convenience of notation, we define the quantity $s = N^\theta$. Note that s controls the distribution of the sparsity of the signals. In particular, if k denotes the sparsity of the signal, then $E[k] = s$. Further, the probability that an integer belongs to the support is given by $\frac{s}{N}$.

Support Recovery

Consider the problem of recovery of the support of the signal V from the support of the autocorrelation (denoted from now on by W). We will assume that if $a[i] = 0$, then no two elements in **x** are separated by a distance i, i.e.,

$$a[i] = 0 \Rightarrow x[j]x^\star[i + j] = 0 \; \forall \; j.$$

This holds with probability one if the nonzero components of the signal are chosen from a continuous i.i.d. distribution. With this assumption, the support recovery

problem can be stated as

$$\text{find} \qquad V \qquad \text{subject to} \qquad \{|i-j| \mid i,j \in V\} = W, \tag{3.7}$$

which is the problem of recovering an integer set from its pairwise distance set (also known as the *Turnpike Problem*[2]).

For example, consider the set $V = \{2, 5, 13, 31, 44\}$. Its pairwise distance set is given by $W = \{0, 3, 8, 11, 13, 18, 26, 29, 31, 39, 42\}$. The Turnpike problem (and (3.7)) is the problem of reconstruction of the set V from the set W. We refer the interested readers to [JH13] for more details.

In [SSL90], a backtracking based algorithm is proposed to solve the turnpike problem. The algorithm needs multiplicity information of the pairwise distances which is not available in the phase retrieval setup, and is known to have a worst case exponential $O(2^k)$-complexity. In [LW88], a polynomial factorization based algorithm with complexity $O(k^d)$ is proposed, where d is the largest pairwise distance. [Dak00] provides a comprehensive summary of the existing algorithms for the turnpike problem. In the following part, we will develop a $O(k^4)$-complexity algorithm which can provably recover most $O(N^{\frac{1}{2}-\epsilon})$-sparse integer sets.

Suppose $V = \{v_0, v_1, \ldots, v_{k-1}\}$ is a set of k integers and $W = \{w_0, w_1, \ldots, w_{K-1}\}$ is its pairwise distance set[4].

If V has a pairwise distance set W, then sets $c \pm V$ also have a pairwise distance set W for any integer c, because of which there are trivial ambiguities. These solutions are considered equivalent, we attempt to recover the equivalent solution $U = \{u_0, u_1, \ldots, u_{k-1}\}$ defined as follows:

$$U = \begin{cases} V - v_0 & \text{if} \quad v_1 - v_0 \leq v_{k-1} - v_{k-2} \\ v_{k-1} - V & \text{otherwise,} \end{cases}$$

i.e., the equivalent solution set U we attempt to recover has the following properties:

(i) $u_0 = 0$

(ii) $u_1 - u_0 \leq u_{k-1} - u_{k-2}$.

[2]Many papers consider the problem of recovering a set of integers from the multiset of their pairwise distances, i.e., multiplicity of pairwise distances is known. We provide a solution without using multiplicity information.

[4]The elements of V and W are assumed to be in ascending order without loss of generality for convenience of notation, i.e., $v_0 < v_1 < \ldots < v_{k-1}$ and $w_0 < w_1 < \ldots < w_{K-1}$.

Let $u_{ij} = |u_j - u_i|$ for $0 \leq i \leq j \leq k-1$. With this definition, $W = \{u_{ij} : 0 \leq i \leq j \leq k-1\}$ and $U = \{u_{0j} : 0 \leq j \leq k-1\}$. The reason for choosing to recover the equivalent solution U is the following: We have the property $U \subseteq W$. Algorithm 3, in essence, crosses out all the integers in W that do not belong to U using two instances of *Intersection Step* and one instance of *Graph Step*.

Algorithm 3 Support Recovery: Combinatorial Algorithm

Input: Pairwise distance set W
Output: Integer set U which has W as its pairwise distance set

1. $u_{01} = w_{K-1} - w_{K-2}$
2. Intersection Step using u_{01}: get $Z = 0 \cup (W \cap (W + u_{01}))$
3. Graph Step using (Z, W): get $\{u_{0p} : 0 \leq p \leq t = \sqrt[3]{\log(s)}\}$ (smallest $t+1$ integers which have an edge with $u_{0,k-1}$)
4. Intersection Step using $\{u_{0p} : 1 \leq p \leq t\}$: get $U = \{u_{0p} : 0 \leq p \leq t-1\} \cup \left(W \cap \left(\bigcap_{p=1}^{t}(W + u_{0p})\right)\right)$

Inferring u_{01}

The largest integer in W (i.e., w_{K-1}) corresponds to the term $u_{0,k-1}$ and the second largest integer in W (i.e., w_{K-2}) corresponds to the term $u_{1,k-1}$ (due to $u_1 - u_0 \leq u_{k-1} - u_{k-2}$). Hence, $w_{K-1} - w_{K-2} = u_{0,k-1} - u_{1,k-1} = u_{01}$. Observe that $u_{01} = v_{01}$ if $v_1 - v_0 \leq v_{k-1} - v_{k-2}$ and $u_{01} = v_{k-2,k-1}$ otherwise.

Intersection Step

The key idea of this step can be summarized as follows: suppose we know the value of u_{0p} for some p, then

$$\{u_{0j} : p \leq j \leq k-1\} \subseteq W \cap (W + u_{0p}),$$

where the set $(W + u_{0p})$ is the set obtained by adding the integer u_{0p} to each integer in the set W. This can be seen as follows: $u_{0j} \in W$ by construction for $0 \leq j \leq k-1$. $u_{pj} \in W$ by construction for $p \leq j \leq k-1$, which when added by u_{0p}, gives u_{0j} and hence $u_{0j} \in (W + u_{0p})$ for $p \leq j \leq k-1$.

The idea can be generalized to multiple intersections. Suppose we know $\{u_{0p} : 1 \leq p \leq t\}$, we can construct $\{(W + u_{0p}) : 1 \leq p \leq t\}$ and see that

$$\{u_{0j} : t \leq j \leq k-1\} \subseteq W \cap \left(\cap_{p=1}^{t}(W + u_{0p})\right).$$

The idea can also be extended to the case when we know the value of $u_{q,k-1}$ for some q:

$$\{u_{j,k-1} : 0 \leq j \leq q\} \subseteq W \cap (W + u_{q,k-1}),$$

which can be seen as follows: $u_{j,k-1} \in W$ by construction for $0 \leq j \leq k-1$. $u_{jq} \in W$ by construction for $0 \leq j \leq q$, which when added by $u_{q,k-1}$, gives $u_{j,k-1}$ and hence $u_{j,k-1} \in (W + u_{q,k-1})$ for $0 \leq j \leq q$.

Consider the example $V = \{2, 5, 13, 31, 44\}$, $W = \{0, 3, 8, 11, 13, 18, 26, 29, 31, 39, 42\}$. We have $u_{01} = 3$, because of which $W_1 = \{3, 6, 11, 14, 16, 21, 29, 32, 34, 42, 45\}$ and hence $W \cap W_1 = \{3, 11, 29, 42\}$, which contains $\{u_{01}, u_{02}, u_{03}, u_{04}\} = \{3, 11, 29, 42\}$.

Graph Step

For an integer set U whose pairwise distance set is W, consider any set $Z = \{z_0, z_1, \ldots, z_{|Z|-1}\}$ which satisfies $U \subseteq Z \subseteq W$. Construct a graph $G(Z, W)$ with $|Z|$ vertices (each vertex corresponding to an integer in Z) such that there exists an edge between z_i and z_j iff the following two conditions are satisfied:

(i) $\forall z_g, z_h \in Z, z_g - z_h \neq z_i - z_j$ unless $(i, j) = (g, h)$

(ii) $|z_i - z_j| \in W$,

i.e., there exists an edge between two vertices iff their corresponding pairwise distance is unique and belongs to W.

The main idea of this step is as follows: suppose we draw a graph $G(Z, W)$ where $U \subseteq Z \subseteq W$. If there exists an edge between a pair of integers $z_i, z_j \in Z$, then $z_i, z_j \in U$. This holds because if $z_i, z_j \notin U$, then since $|z_i - z_j| \in W$ there has to be another pair of integers in U (and hence in Z) which have a pairwise distance $|z_i - z_j|$. This would contradict the fact that an edge exists between z_i and z_j in $G(Z, W)$.

Consider the example $V = \{2, 5, 13, 31, 44\}$, $W = \{0, 3, 8, 11, 13, 18, 26, 29, 31, 39, 42\}$. Suppose we have $Z = \{0, 3, 8, 11, 29, 42\}$. There will be an edge between 11 and 42

as they have a difference of 31, which belongs to W and there are no other integer pairs in Z which have a difference of 31. Hence, the only way a pairwise distance of 31 in W can be explained is if $11, 42 \in U$.

Theorem 3.3.2. *If sparse signals are drawn from the $\mathcal{BN}(n, \theta)$ distribution, where the parameter θ satisfies $0 < \theta \leq \frac{1}{2} - \epsilon$ for some constant $\epsilon > 0$, then the failure probability of Algorithm 3 is $O(n^{-0.1\epsilon})$.*

Proof. The proof of this theorem is constructive, i.e., we prove the correctness of the various steps involved in Algorithm 3 with the desired probability. The outline is as follows:

Due to $U \subseteq W$ property, we noted that Algorithm 3 aims to cross out integers in W that do not belong to U (undesired integers). The Intersection Step and Graph Step are designed such that they never cross out integers which belong to U, and cross out undesired integers with certain probabilities.

Lemma 8.2.2 provides a $O\left(\frac{s^4}{n^2}\right)$ bound on the probability that a particular undesired integer does not get crossed out in the first Intersection Step. If $0 < \theta \leq \frac{1}{5}$, then Lemma 8.2.3 shows that the support is recovered at the end of the first Intersection Step itself with the desired probability.

The Graph Step and the second instance of the Intersection Step cross out undesired integers, if any, when $\frac{1}{5} < \theta$. Lemma 8.2.6 shows that $\{v_{0p} : 1 \leq p \leq t = \sqrt[3]{\log(s)}\}$ can be recovered by Graph Step with the desired probability. Finally, Lemmas 8.2.4 and 8.2.5 show that the support is recovered at the end of the second Intersection Step with the desired probability. We refer the readers to Appendix 8.2 for details. □

Signal Recovery (with known support)

Once the support is recovered, the signal can be recovered by solving (3.6). We use $\lambda = 0$ as the support constraints promote sparsity by themselves.

Theorem 3.3.3. *If the sparse signal $\mathbf{x}_0$ is drawn from the $\mathcal{BN}(n, \theta)$ distribution, where the parameter θ satisfies $0 < \theta \leq \frac{1}{2} - \epsilon$ for some constant $\epsilon > 0$, then the probability that the optimizer of (3.6), with $\lambda = 0$, is not $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$ is $O(n^{-1})$.*

Proof. Analysis of semidefinite relaxation based programs with such deterministic measurements is a difficult task in general. We will instead analyze (3.8), which is a further relaxation of (3.6), and show that (3.8) has $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$ as its optimizer with

the desired probability, which is sufficient to prove the theorem as $\mathbf{x}_0\mathbf{x}_0^\star$ is a feasible point of (3.6).

We use the following notation: $H(U) = G(U, W)$ (see the description of Graph Step). In other words, $H(U)$ is a graph with k vertices, where each vertex corresponds to an integer in U and two vertices have an edge between them if their corresponding integers have a unique pairwise distance.

The key idea is the following: If there exists an edge between vertices corresponding to u_i and u_j in the graph $H(U)$, then $X[u_i, u_j]$ can be deduced from the autocorrelation. This is because if there is an edge between u_i and u_j, then $a[|u_i - u_j|] = x[u_i]x^\star[u_j]$, which by definition is $X[u_i, u_j]$. The convex program (3.6) can be relaxed by using only such autocorrelation constraints which fix certain entries of $\mathbf{X}$ (and discarding the rest), and by replacing the positive semidefinite constraint with the constraint that every 2×2 submatrix of $\mathbf{X}$ is positive semidefinite, i.e.,

$$\begin{aligned} \text{minimize} \quad & \text{trace}(\mathbf{X}) \\ \text{subject to} \quad & X[u_i, u_j] = a[|u_i - u_j|] \quad \text{if } u_i \leftrightarrow u_j \text{ in } H(U) \\ & X[i, j] = 0 \quad \text{if} \quad i, j \notin U \\ & X[i, i]X[j, j] \geq |X[i, j]|^2 \quad \forall \ i \neq j \ \ \& \ \ X[i, i] \geq 0 \quad \forall\, i, \end{aligned} \tag{3.8}$$

where $u_i \leftrightarrow u_j$ means that there exists an edge between vertices corresponding to u_i and u_j in $H(U)$.

Note that $\log^6(s) \leq k$ holds with the desired probability. The events are first conditioned with respect to a fixed k in this interval, a union bound over all values of k in this interval completes the bound.

Lemma 8.3.3 shows that the minimum degree of $H(U)$, denoted by $d_{min}(H(U))$, satisfies $d_{min}(H(U)) > k(1 - \frac{1}{t})$, where $t = \log^2(s)$, with the desired probability. Hajnal-Szemeredi theorem on disjoint cliques [HS70] states that such graphs contain $\frac{k}{t}$ vertex disjoint union of complete graphs of size t.

Lemma 8.3.1, along with a union bound, shows that the entries of the optimizer of (3.8) match with the entries of $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$ on each of these $\frac{k}{t}$ complete graphs with the desired probability. Consequently, the diagonal entries of the optimizer of (3.8) match with the diagonal entries of $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$ with the desired probability.

Also, since the graph $H(U)$ has a Hamiltonian cycle (Lemma 8.3.3), by rearranging the indices, we see that the first off-diagonal entries of the optimizer of (3.8) also

match with the first off-diagonal entries of $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$. Since the optimizer's diagonal and first off-diagonal entries are sampled from a rank one matrix, there is exactly one positive semidefinite completion, which is the rank one completion $\mathbf{x}_0\mathbf{x}_0^\star$. Since the optimizer also satisfies all the constraints of (3.6), $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$ is the unique minimizer of (3.6) with the desired probability.

We refer the readers to Appendix 8.3 for details. □

3.4 Stability

In practice, the measured autocorrelation is corrupted with additive noise, i.e., the measurements are of the form

$$a[m] = \sum_{n=0}^{N-1-m} x[n]x^\star[n+m] + z[m] \quad \text{for} \quad 0 \le m \le N-1,$$

where $\mathbf{z} = (z_0, z_1, \ldots, z_{N-1})$ is the additive noise. TSPR, in its pure form (support recovery using Algorithm 3), is not robust to noise as the u_{01} identification step and Graph Step are not robust. In this section, we present a modified version of TSPR, which in essence, considers the pairwise distance set of the pairwise distance set to identify $u_{i_0j_0}$, for some $0 \le i_0 < j_0 \le 2c+1$, robustly and then uses a sequence of *generalized* Intersection Steps to provably recover the true support of most $O(n^{\frac{1}{4}-\epsilon})$-sparse signals.

The support of the noisy autocorrelation, denoted by $W^\dagger = (w_0^\dagger, w_1^\dagger, \ldots, w_{K^\dagger-1}^\dagger)$, can be defined as the set of integers $\{\, n \mid |a[n]| \ge \tau\}$ where τ is a threshold parameter. Let $T^\dagger = \{(w_i^\dagger, w_j^\dagger) : 0 \le i < j \le K^\dagger - 1\}$ denote the set containing the $\binom{K^\dagger}{2}$ integer pairs formed using the $K^\dagger$ integers in $W^\dagger$. Let $T_{sub}^\dagger$ be a subset of $T^\dagger$ which contains all the integer pairs $(w_i^\dagger, w_j^\dagger)$ (where $j > i$), satisfying the following two conditions:

(i) $w_j^\dagger - w_i^\dagger \in W^\dagger$

(ii) $\exists \frac{\sqrt{K^\dagger}}{4}$ integers $\{g_1, g_2, \ldots, g_{\frac{\sqrt{K^\dagger}}{4}}\}$, such that $g_l, g_l + w_j^\dagger - w_i^\dagger \in T^\dagger$ for $1 \le l \le \frac{\sqrt{K^\dagger}}{4}$.

The first condition requires that the difference between the integers in the pair should be in $W^\dagger$ and the second condition requires that at least $\frac{\sqrt{K^\dagger}}{4}$ integer pairs in $W^\dagger$ should be separated by the same difference.

As earlier, let W denote the support of the autocorrelation (in the absence of noise). Let W_{ins} denote the set of integers which belong to $W^\dagger$ but do not belong to W: these

Algorithm 4 Two-stage Sparse Phase Retrieval: Noisy Setup

Input: Noisy autocorrelation $\mathbf{a}$ of the signal of interest, threshold τ, η such that $\|\mathbf{z}\|_2 \leq \eta$, constant c

Output: Sparse signal $\hat{\mathbf{x}}$ satisfying the noisy autocorrelation measurements

(i) $W^\dagger = \{\, n \mid |a[n]| \geq \tau\}$

(ii) $u_{i_0 j_0} = w^\dagger_{max} - w^\dagger_{min}$, where $0 \leq i_0 < j_0 \leq 2c+1$: $w^\dagger_{min}$ is the largest integer for which there exists an integer $w^\dagger_{max} > w^\dagger_{min}$ such that $(w^\dagger_{min}, w^\dagger_{max}) \in T^\dagger_{sub}$

(iii) Intersection Step using $u_{i_0 j_0}$: get $\{u_{i_0 q_0}, u_{i_0 q_1}, \ldots, u_{i_0 q_{c+1}}\}$, where $\{q_0, q_1, \ldots, q_{c+1}\} \geq (k-1)-(3c+1)$ (largest $c+2$ integers in $W^\dagger \cap (W^\dagger + u_{i_0 j_0})$)

(iv) Intersection Step using each of the $\binom{c+2}{2}$ terms $\{u_{q_i q_j} : 0 \leq i < j \leq c+1\}$: obtain $\{u_0, u_1, \ldots, u_{\frac{\sqrt{K^\dagger}}{4}-1}\}$ (largest $\frac{\sqrt{K^\dagger}}{4}$ integers in $\bigcup_{0\leq i<j\leq c+1} \left((W^\dagger \cap (W^\dagger + u_{q_i q_j})) + u_{q_j q_{c+1}}\right)$ correspond to $\{u_{i q_{c+1}} : 0 \leq i \leq \frac{\sqrt{K^\dagger}}{4} - 1\}$)

(v) Intersection Step using each of the $\binom{c+2}{2}$ terms $\{u_{ij} : 0 \leq i < j \leq c+1\}$: obtain $\{u_{\frac{\sqrt{K^\dagger}}{4}}, u_{\frac{\sqrt{K^\dagger}}{4}+1}, \ldots, u_{k-1}\}$ (all the integers greater than $u_{\frac{\sqrt{K^\dagger}}{4}-1}$ in $\bigcup_{0\leq i<j\leq c+1} \left((W^\dagger \cap (W^\dagger + u_{ij})) + u_{0i}\right)$)

(vi) Obtain $\mathbf{X}^\dagger$ by solving

$$\begin{aligned} &\text{minimize} && \operatorname{trace}(\mathbf{X}) \\ &\text{subject to} && |a[m] - \operatorname{trace}(\mathbf{A}_m \mathbf{X})| \leq \eta \quad \text{for} \quad 0 \leq m \leq N-1 \\ &&& X[n,m] = 0 \ \text{ if } \ n,m \notin U \ \ \& \ \ \mathbf{X} \succcurlyeq 0 \end{aligned} \tag{3.9}$$

(vii) Return $\mathbf{x}^\dagger$, where $\mathbf{x}^\dagger \mathbf{x}^{\dagger\star}$ is the best rank one approximation of $\mathbf{X}^\dagger$

---

are the integers which got inserted due to a noise value higher than the threshold. Also, let W_{del} denote the set of integers which belong to W but do not belong to $W^\dagger$: these are the integers which got deleted due to the autocorrelation value being below the threshold or due to noise reducing the autocorrelation value below the threshold. We have:

$$W^\dagger = (W \cup W_{ins}) \backslash W_{del}. \tag{3.10}$$

Theorem 3.4.1. *If the sparse signal $\mathbf{x}_0$ is drawn from the $\mathcal{BN}(n,\theta)$ distribution, where the parameter θ satisfies $0 < \theta \leq \frac{1}{4} - \epsilon$ for some constant $\epsilon > 0$, then TSPR*

(noisy setup) can recover it from its noisy autocorrelation measurements ($\|\mathbf{z}\|_2 \leq \eta$) with an estimation error

$$\|\mathbf{X}^{\dagger} - \mathbf{x}_0\mathbf{x}_0^{\star}\|_2 \leq 4k\eta$$

with probability at least $1 - c_0 n^{-4\epsilon}$, for some numerical constant c_0, if the noise vector $\mathbf{z}$ and threshold τ are such that for some constant c, we have

(i) W_{ins} *has i.i.d.* $Bern(p)$ *distribution, where* $p = o\left(\frac{s^2}{n}\right)$

(ii) For each $0 \leq i \leq k-1$, W_{del} *contains at most* c *terms of the form* $\{v_{ij} : 0 \leq j \leq k-1\}$, *and* $v_{0,k-1} \notin W_{del}$.

Proof. The proof of this theorem is constructive, i.e., we prove the correctness of the various steps involved with the desired probability.

We refer the readers to Appendix 8.4 for details. The outline is as follows: Lemma 8.4.1 bounds the probability of the first step failing by $O(n^{-4\epsilon})$. Then, a detailed discussion of the Generalized Intersection Step is provided. Finally, Lemma 8.4.2, combined with Lemma 8.2.3, shows that TSPR (noisy setup) can precisely recover the support of the signal with the desired probability. We then show that the signal values can be robustly recovered by the convex relaxation based program. □

3.5 Extension to $2D$

The theory and algorithms developed in this chapter can be generalized to $2D$ using the following trick: Let $\mathbf{x}$ be a two-dimensional signal with N_1 rows and N_2 columns, and $\mathbf{a}$ be its two-dimensional autocorrelation with $2N_1-1$ rows and $2N_2-1$ columns. Let $\mathbf{x}_{1D} = \text{vec}(\mathbf{x})$ denote the one-dimensional vector constructed by stacking the columns of $\mathbf{x}$ on top of each other. The one-dimensional autocorrelation of $\mathbf{x}_{1D}$, denoted by $\mathbf{a}_{1D}$, can be inferred from $\mathbf{a}$. This can be seen as follows:

$$\begin{aligned}
\mathbf{a}_{1D}[m] &= \sum_{n=0}^{N_1N_2-1-m} x_{1D}[n]x_{1D}^{\star}[n+m] \\
&= \sum_{l=0}^{N_2-1-\lfloor \frac{m}{N_1} \rfloor} \sum_{n=0}^{N_1-1-(m) \bmod N_1} x_{1D}[n+lN_1]x_{1D}^{\star}[n+lN_1+m] \\
&+ \sum_{l=0}^{N_2-2-\lfloor \frac{m}{N_1} \rfloor} \sum_{n=N_1-(m) \bmod N_1}^{N_1-1} x_{1D}[n+lN_1]x_{1D}^{\star}[n+lN_1+m]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l=0}^{N_2-1-\lfloor \frac{m}{N_1} \rfloor} \sum_{n=0}^{N_1-1-(m) \bmod N_1} x[n,l]x^{\star}[n+(m) \bmod N_1, l+\Big\lfloor \frac{m}{N_1} \Big\rfloor] \\
&+ \sum_{l=0}^{N_2-2-\lfloor \frac{m}{N_1} \rfloor} \sum_{n=N_1-(m) \bmod N_1}^{N_1-1} x[n,l]x^{\star}[n-N_1+(m) \bmod N_1, l+\Big\lfloor \frac{m}{N_1} \Big\rfloor + 1] \\
&= a[(m) \bmod N_1, \Big\lfloor \frac{m}{N_1} \Big\rfloor] + a[-N_1+(m) \bmod N_1, \Big\lfloor \frac{m}{N_1} \Big\rfloor + 1].
\end{aligned}
$$

Since we know the autocorrelation $\mathbf{a}_{1D}$ of the sparse one-dimensional signal $\mathbf{x}_{1D} = \text{vec}(\mathbf{x})$, the results derived in this chapter apply to $\mathbf{x}_{1D}$. Consequently, $\mathbf{x}_{1D}$ can be uniquely reconstructed up to trivial ambiguities. However, note that the time-shift ambiguity of $\mathbf{x}_{1D}$ and the time-shift ambiguity of $\mathbf{x}$ are slightly different. In order to overcome this issue, one needs to make use of the sparsity structure of $\mathbf{a}$ to reduce the number of possible time-shifts of $\mathbf{x}_{1D}$. We refer the interested readers to [KEO16] for a detailed discussion on this technique.

Remark: This trick also works when the one-dimensional signal is obtained by stacking the rows next to each other.

Alternately, one could also generalize TSPR to $2D$ directly as the principles involved in the Intersection Step and the Graph Step are dimension independent. However, the theoretical analysis needs to be redone if this approach is used.

3.6 Numerical Simulations

In this section, we demonstrate the performance of TSPR using numerical simulations. The procedure is as follows: for a given N and k, the k locations of the nonzero components were chosen uniformly at random. The signal values in the chosen support were drawn from an i.i.d. standard normal distribution.

Success probability

In the first set of simulations, we demonstrate the performance of TSPR for $N = 12500$, $N = 25000$ and $N = 50000$ for various sparsities. The results of the simulations are shown in Fig. 3.2, the $O(N^{\frac{1}{2}-\epsilon})$ theoretical prediction can be clearly seen. For instance, $N = 12500$, $k = 80$ and $N = 50000$, $k = 160$ have a success probability of 0.5 and so on.

Failure exponent

In the second set of simulations, we numerically study the failure probability of TSPR, denoted by δ. For $\theta = 0.42, 0.44, 0.46$, we plot $\log_2(\delta)$ versus $\log_2(N)$ for

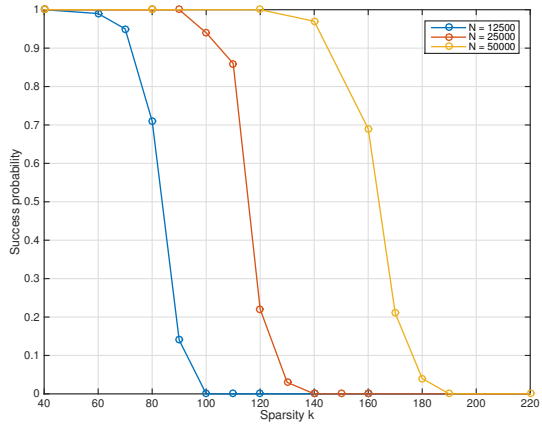


Figure 3.2: Probability of successful signal recovery of TSPR for various sparsities and $N = 12500, 25000, 50000$.

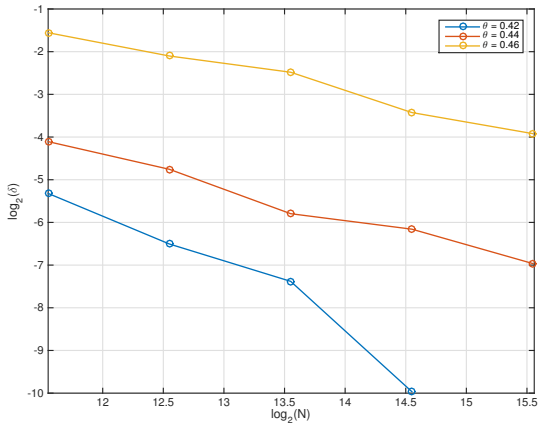


Figure 3.3: Failure probability of TSPR for various N and $\theta = 0.42, 0.44, 0.46$.

various choices of N. Theorem 3.3.1 upper bounds the slope by $-0.1 \times (\frac{1}{2} - \theta)$. The results of the simulations are shown in Fig. 3.3. It can be seen that the results are in accordance with the bounds provided by Theorem 3.3.1. It is also clear that $-0.1 \times (\frac{1}{2} - \theta)$ is not a tight bound, which is not surprising as the analysis in Theorem 3.3.1 involved many union bounds, which are typically not tight.

Comparison with fast algorithms

In this set of simulations, we compare the recovery ability of TSPR with other popular sparse phase retrieval algorithms. We choose $N = 6400$ and plot the success probabilities of the algorithms TSPR, GESPAR [SBE14] and Sparse-Fienup (100

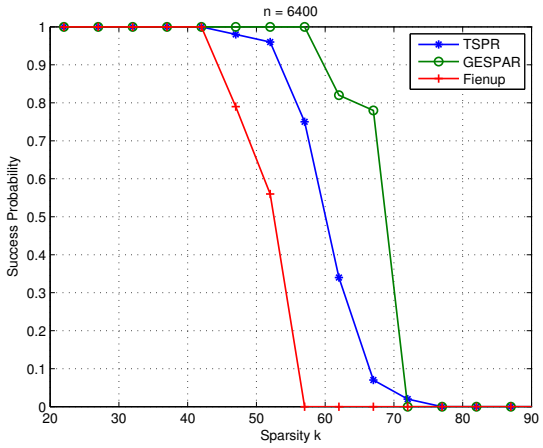


Figure 3.4: Probability of successful signal recovery of various efficient sparse phase retrieval algorithms for various sparsities and $N = 6400$.

random initializations) [Fie82] for sparsities $20 \leq k \leq 90$. The results of the simulations are shown in Fig. 3.4.

Fig. 3.4 shows that TSPR outperforms Sparse-Fienup algorithm and is almost on par with GESPAR. We expect TSPR to outperform GESPAR for higher values of N due to the fact that it can recover $O(N^{\frac{1}{2}-\epsilon})$-sparse signals (GESPAR empirically recovers $O(N^{\frac{1}{3}})$-sparse signals). We suspect that the recovery ability of the two algorithms for $N = 6400$ is similar due to the effect of the constants multiplying these terms. We were unable to compare the performances for higher values of N due to scalability limitations of GESPAR. For instance, TSPR took an average run time of $80ms$ to recover a signal with $N = 25000$ and $k = 100$ whereas GESPAR needed an average run time of $33s$ to recover a signal with $N = 512$ and $k = 35$.

Comparison with SDP algorithms

In this set of simulations, we compare the recovery ability of TSPR with the SDP heuristic (based on log-det minimization) proposed in [Can+15]. We choose $N = 64$ and plot the success probabilities for sparsities $0 \leq k \leq 20$. The results are shown in Fig. 3.5, we observe that the performances are similar.

Image reconstruction

Finally, we test the performance of TSPR on real images. To this end, we use a 54×64 image of the M73 asterism in the constellation of Aquarius [NCK]. The original image (Fig. 3.6a) is converted into a sparse binary image by thresholding.

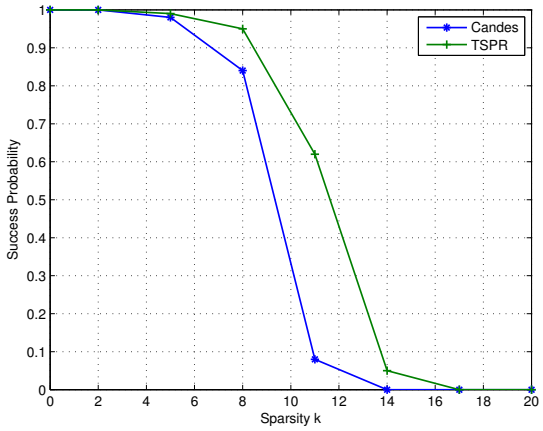


Figure 3.5: Probability of successful signal recovery of various SDP based sparse phase retrieval algorithms for various sparsities and $N = 64$.

(a) (b) (c) (d)

Figure 3.6: Reconstruction of sparse images using TSPR. (a) A 54×64 image of the M73 asterism in the constellation of Aquarius (courtesy of [NCK]). (b) A 44-sparse binary image obtained using hard-thresholding. (c) Output of TSPR. (d) Reconstruction error, after accounting for trivial ambiguities.

In particular, by using a threshold value equal to 25% of the maximum value, a binary image with sparsity 44 is obtained (Fig. 3.6b). The reconstructed image and the error are shown in Fig. 3.6c and 3.6d respectively. The output of TSPR has sparsity 47: the original 44 support locations are accurately reconstructed, and only 3 undesired support locations were not crossed out.

3.7 Conclusions and Future Work

We have identified the following problems as potential directions for future research:

- We showed that almost all signals with aperiodic support can be recovered by solving (3.2). Note that most signals with sparsity up to $N - 1$ have aperiodic support. TSPR can efficiently solve (3.2) with high probability if the signals are $O(N^{\frac{1}{2}-\epsilon})$-sparse. It is unclear whether signals with sparsity

greater than $O(N^{\frac{1}{2}})$ can be reconstructed efficiently by any algorithm. In several related sparse quadratic constrained problems like sparse PCA [B+13; BR13] and sparse recovery from random phaseless measurements [LV13; Oym+15], there is a fundamental gap between the set of signals that can be identified and the set of signals that can be efficiently identified (the bottleneck happens at $O(N^{\frac{1}{2}})$ sparsity [JOH13a]). Hence, a precise characterization of the set of signals that can be efficiently reconstructed from their autocorrelation by any algorithm would provide valuable insights into our understanding of general sparse quadratic constrained problems.

- We showed that $O(N^{\frac{1}{4}-\epsilon})$-sparse signals can be recovered robustly by TSPR in the presence of additive noise. A precise characterization of the set of signals that can be efficiently and robustly reconstructed by any algorithm is another interesting open question.

Chapter 4

PHASE RETRIEVAL WITH MASKS

In this chapter, we explore the idea of obtaining additional magnitude-only measurements, in order to be able to uniquely and efficiently identify the signal of interest. In particular, we consider measurements of the form

$$Z[m] = \left|\langle \mathbf{f}_m, \mathbf{D}\mathbf{x}\rangle\right|^2, \tag{4.1}$$

where $\mathbf{D}$ is an $N \times N$ diagonal matrix. Note that

$$\mathbf{D}\mathbf{x} = \begin{bmatrix} d[0] & 0 & \dots & 0 \\ 0 & d[1] & \dots & 0 \\ & & & \\ 0 & 0 & \dots & d[N-1] \end{bmatrix} \begin{bmatrix} x[0] \\ x[1] \\ \\ x[N-1] \end{bmatrix} = \begin{bmatrix} d[0]x[0] \\ d[1]x[1] \\ \\ d[N-1]x[N-1] \end{bmatrix}.$$

Effectively, the underlying signal is Hadamard-multiplied with a modulating vector $(d[0], d[1], \dots, d[N-1])^T$, and the Fourier magnitude-square of the modulated signal is assumed to be available as additional information.

There are many ways in which this can be done in practice, depending on the application. Several such methods are summarized in [CLS15a]. In this chapter, we focus on "masking", a technique where the signal is modified by the use of a mask or a phase plate [Joh+08]. A schematic representation, courtesy of [CLS15a], is provided in Fig. 4.1.

Suppose Fourier magnitude-square measurements are collected using R masks. For $0 \le r \le R-1$, let $\mathbf{D}_r$ be an $N \times N$ diagonal matrix, corresponding to the rth mask, with diagonal entries $(d_r[0], d_r[1], \dots, d_r[N-1])$. Let $\mathbf{Z}$ denote the $N \times R$ magnitude-square measurements, such that the rth column of $\mathbf{Z}$ corresponds to the magnitude-square of the N point DFT of the masked signal $\mathbf{D}_r\mathbf{x}$. Phase retrieval using masks then reduces to the following recovery problem:

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && Z[m,r] = \left|\langle \mathbf{f}_m, \mathbf{D}_r\mathbf{x}\rangle\right|^2 \quad \text{for} \quad 0 \le m \le N-1 \quad \text{and} \quad 0 \le r \le R-1, \end{aligned} \tag{4.2}$$

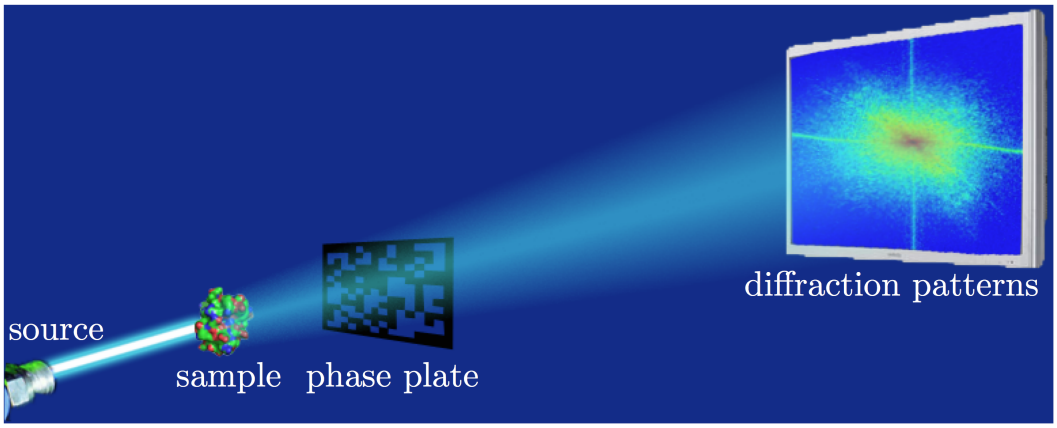


Figure 4.1: A typical setup for phase retrieval using masks (courtesy of [CLS15a]).

or equivalently,

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && b[m,r] = \sum_{n=0}^{N-1} d_r[n]x[n]d_r^\star[(n+m) \bmod N]x^\star[(n+m) \bmod N] \\ &&& \text{for} \quad 0 \le m \le N-1 \quad \text{and} \quad 0 \le r \le R-1. \end{aligned} \tag{4.3}$$

In the oversampled setting, this recovery problem can be rewritten as

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && a[m,r] = \sum_{n=0}^{N-1-m} d_r[n]x[n]d_r^\star[n+m]x^\star[n+m] \\ &&& \text{for} \quad 0 \le m \le N-1 \quad \text{and} \quad 0 \le r \le R-1. \end{aligned} \tag{4.4}$$

A natural question to ask is which masks guarantee uniqueness, and allow efficient recovery.

4.1 Literature Survey

Phase retrieval algorithms based on SDP and stochastic gradient descent (Wirtinger Flow algorithm [CLS15b]) have been adapted to solve phase retrieval for some choice of masks. Combinatorial algorithms have also been developed for specific mask designs which allow unique and efficient reconstruction in the noiseless setting. In what follows, we first review the existing literature, and then present our results.

A combinatorial algorithm is proposed in [Can+15] for the three masks $\{\mathbf{I}, \mathbf{I}+\mathbf{D}^s, \mathbf{I}-i\mathbf{D}^s\}$, where s is any integer coprime with N and $\mathbf{D}$ is a diagonal matrix with diagonal

entries

$$d[n] = e^{i2\pi \frac{n}{N}} \quad \text{for} \quad 0 \leq n \leq N-1.$$

It is shown that signals with non-vanishing N point DFT can be uniquely recovered using these masks up to a global phase. Indeed, the measurements obtained in this case provide the knowledge of $|y[n]|^2, |y[n] + y[n-s]|^2$ and $|y[n] - iy[n-s]|^2$ for $0 \leq n \leq N-1$ ($n-s$ is understood modulo N). Writing $y[n] = |y[n]|\, e^{i\phi[n]}$ for $0 \leq n \leq N-1$, we have

$$|y[n] + y[n-s]|^2 = |y[n]|^2 + |y[n-s]|^2 + 2|y[n]||y[n-s]| \operatorname{Re}(e^{i(\phi[n-s]-\phi[n])}),$$

$$|y[n] - iy[n-s]|^2 = |y[n]|^2 + |y[n-s]|^2 + 2|y[n]||y[n-s]| \operatorname{Im}(e^{i(\phi[n-s]-\phi[n])}).$$

Consequently, if $y[n] \neq 0$ for $0 \leq n \leq N-1$, then the measurements provide the relative phases $\phi[n-s] - \phi[n]$ for $0 \leq n \leq N-1$. Since s is coprime with N, by setting $\phi[0] = 0$ without loss of generality, $\phi[n]$ can be inferred for $1 \leq n \leq N-1$. Since most signals have a non-vanishing N point DFT, these three masks may be used to recover most signals efficiently.

In order to be able to recover all signals (as opposed to most signals), a polarization based technique is proposed in [Ale+14; BCM14]. It is shown that $O(\log N)$ masks (see [Ale+14] for design details) are sufficient for this technique.

In [PLR14], the authors consider a combinatorial algorithm, based on coding theoretic tools, for the 3 masks $\{\mathbf{I}, \mathbf{I} + \mathbf{e}_0\mathbf{e}_0^\star, \mathbf{I} + i\mathbf{e}_0\mathbf{e}_0^\star\}$, where $\mathbf{e}_0$ is the $N \times 1$ column vector $(1, 0, \ldots, 0)^T$. For signals with $x[0] \neq 0$, it is shown that the value of $|x[0]|$ can be uniquely found with high probability. The phase of $x[0]$ is set to 0 without loss of generality, and the phase of $x[n]$ relative to $x[0]$ for $0 \leq n \leq N-1$ is inferred by solving a set of algebraic equations.

These methods are typically unstable in the presence of noise, due to the issue of error propagation. SDP based phase retrieval has been adapted to account for random masks in [CLS15a] by solving

$$\begin{aligned} \text{minimize} \quad & \operatorname{trace}(\mathbf{X}) \\ \text{subject to} \quad & Z[m,r] = \operatorname{trace}(\, \mathbf{D}_r^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{D}_r \mathbf{X}\,) \quad \text{for} \quad 0 \leq m \leq N-1 \quad \text{and} \quad 0 \leq r \leq R-1 \\ & \mathbf{X} \succcurlyeq 0. \end{aligned} \tag{4.5}$$

In order to provide recovery guarantees, the masks in [CLS15a] are chosen from a random model. In particular, the diagonal matrices $\mathbf{D}_r$ are assumed to be i.i.d.

copies of a matrix $\mathbf{D}$, whose entries consist of i.i.d. copies of a random variable d satisfying the following properties:

$$\mathbb{E}[d] = 0 \qquad \mathbb{E}[d^2] = 0 \qquad \mathbb{E}|d|^4 = 2\mathbb{E}|d|^2.$$

An example of an admissible random variable is given by $d = b_1 b_2$, where b_1 and b_2 are independent and distributed as

$$b_1 = \begin{cases} 1 & \text{with prob.} \quad \frac{1}{4} \\ -1 & \text{with prob.} \quad \frac{1}{4} \\ i & \text{with prob.} \quad \frac{1}{4} \\ -i & \text{with prob.} \quad \frac{1}{4} \end{cases} \qquad b_2 = \begin{cases} 1 & \text{with prob.} \quad \frac{4}{5} \\ \sqrt{6} & \text{with prob.} \quad \frac{1}{5} \end{cases}. \tag{4.6}$$

Under this model, it is shown that $R \geq c \log^4 N$ masks, for some numerical constant c, are sufficient for the convex program (4.5) to uniquely recover the underlying signal up to a global phase with high probability in the noiseless setting. This result has been further refined to $R \geq c \log^2 N$ in [GKK15].

An alternative recovery approach for masked signals is based on the Wirtinger flow method [CLS15b], which applies gradient descent to the least squares problem:

$$\min_{\mathbf{x}} \quad \sum_{r=0}^{R-1} \sum_{m=0}^{N-1} \left(Z[m, r] - |\langle \mathbf{f}_m, \mathbf{D}_r \mathbf{x} \rangle|^2 \right)^2. \tag{4.7}$$

Minimizing such non-convex objectives is known to be NP-hard in general. Gradient descent-type methods have shown promise in solving such problems, however, their performance is very dependent on the initialization and update rules due to the fact that different initialization and update strategies lead to convergence to different (possibly local) minima.

Wirtinger flow (WF) is a gradient descent-type algorithm which starts with a careful initialization obtained by means of a spectral method. We refer the readers to [CLS15b] for a discussion on various spectral method based initialization strategies. The initial estimate is then iteratively refined using particular update rules. It is argued that the average WF update is the same as the average stochastic gradient scheme update. Consequently, WF can be viewed as a stochastic gradient descent algorithm, in which only an unbiased estimate of the true gradient is observed. The authors recommend the use of smaller step-sizes in the early iterations and larger step-sizes in later iterations. When the masks are chosen from a random

Robust methods	$O(\log^4 N)$ random masks, whose diagonal entries are i.i.d. copies of a random variable satisfying some properties, are sufficient for the SDP/ WF algorithm with high probability [CLS15a; CLS15b]
	2 specific masks with oversampling or 3 easy-to-implement masks with oversampling are sufficient for the SDP algorithm almost surely [This work]
Combinatorial methods	$O(\log N)$ masks are sufficient for a polarization based algorithm [BCM14]
	For signals with non-vanishing DFT, 3 specific masks are sufficient [Can+15]
	For signals satisfying $x[0] \neq 0$, 3 specific masks are sufficient [PLR14]

Table 4.1: Various results for phase retrieval using masks.

model with a distribution satisfying properties similar to (4.6), it is shown that $R \geq c \log^4 N$ masks, for some numerical constant c, are sufficient for the WF algorithm to uniquely recover the underlying signal up to a global phase with high probability in the noiseless setting. The aforementioned results are summarized in Table 4.1.

4.2 Contributions

Note that the stable algorithms require at least $O(\log^2(N))$ i.i.d. masks. Such masks are difficult to implement in practice, and $O(\log^2(N))$ is a prohibitive number in general. In this chapter, inspired by practical applications, we focus our attention on simple masks which physically block the light from reaching parts of the sample (see Fig. 4.2 for a pictorial example). In particular, we propose two simple mask designs, one uses only 2 specific masks and the other uses only 3 easy-to-implement masks. We show that the SDP algorithm can provably reconstruct most signals when oversampled measurements are obtained using such masks. If oversampled measurements are unavailable, then the number of masks increases to 5 and 7 respectively. Numerical simulations show that the reconstruction is stable in the presence of measurement noise. These results are a significant improvement over the existing results, due to the simplicity and the number of masks considered, and the fact that there exists a stable reconstruction algorithm.

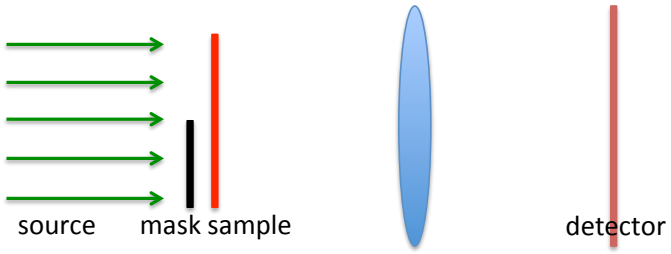


Figure 4.2: A pictorial example of the implementation of a simple mask in an optical setting.

4.3 Design #1

Let $\mathbf{D}_1$ and $\mathbf{D}_2$ be diagonal matrices with diagonal entries

$$d_1[n] = 1 \quad \text{for} \quad 0 \le n \le N-1 \tag{4.8}$$

$$d_2[n] = \begin{cases} 0 & \text{for} \quad n = 0 \\ 1 & \text{for} \quad 1 \le n \le N-1. \end{cases}$$

Essentially, $\mathbf{D}_1$ corresponds to measurements without using any mask, and $\mathbf{D}_2$ corresponds to measurements where the mask blocks only the first location in the sample (see Fig. 4.3a for an example). We consider the SDP formulation

$$\begin{aligned} \text{minimize} \quad & \text{trace}(\mathbf{X}) \\ \text{subject to} \quad & \text{trace}(\mathbf{D}_r^\star \mathbf{A}_m \mathbf{D}_r \mathbf{X}) = a[m,r] \quad \text{for} \quad 0 \le m \le N-1 \quad \text{and} \quad 0 \le r \le R-1 \\ & \mathbf{X} \succcurlyeq 0, \end{aligned} \tag{4.9}$$

where the matrices $\mathbf{A}_m$ are defined in (3.5), and prove the following result:

Theorem 4.3.1. *Consider any signal $\mathbf{x}_0$ such that $x_0[0] \neq 0$. Suppose oversampled measurements are taken using the masks defined by $\mathbf{D}_1$ and $\mathbf{D}_2$. The convex program (4.9) has a unique feasible point, namely, $\mathbf{x}_0\mathbf{x}_0^\star$, and hence $\mathbf{x}_0$ can be uniquely recovered up to a global phase.*

Proof. In the oversampled setting, there is a simple combinatorial recovery algorithm for this particular choice of masks. The measurements obtained using the masks defined by $\mathbf{D}_1$ and $\mathbf{D}_2$ are

$$\begin{aligned} a_1[m] &= \sum_{n=0}^{N-1-m} x[n]x^\star[n+m] \\ a_2[m] &= \sum_{n=1}^{N-1-m} x[n]x^\star[n+m], \end{aligned} \tag{4.10}$$

for $0 \leq m \leq N-1$. Since $a_1[0] - a_2[0] = x[0]x^\star[0]$, we can infer $x[0]$ up to a phase. Using $a_1[m] - a_2[m] = x[0]x^\star[m]$ for $1 \leq m \leq N-1$, we can infer the entire signal $\mathbf{x}$ up to a global phase. However, this method of recovery is unstable in the presence of measurement noise as it does not optimally make use of the available measurements.

From a matrix sensing perspective, the set of measurements

$$a_1[m] = \sum_{n=0}^{N-1-m} X[n, n+m] \quad \& \quad a_2[m] = \sum_{n=1}^{N-1-m} X[n, n+m], \tag{4.11}$$

denoted by $\mathcal{A}(\mathbf{X}) = \mathbf{c}$, fix (i) the entries of the first row and column of $\mathbf{X}$ (can be seen by subtracting $\mathbf{a}_2$ from $\mathbf{a}_1$) (ii) the sum along the mth off-diagonal of $\mathbf{X}$ excluding the first row and column for each m (can be seen as measurements due to $\mathbf{a}_2$). We will show the following: If $\mathbf{x}_0\mathbf{x}_0^\star$ satisfies (4.11), then it is the only positive semidefinite matrix which satisfies (4.11).

Let T be the set of symmetric matrices of the form

$$T = \{\mathbf{X} = \mathbf{x}_0\mathbf{w}^\star + \mathbf{w}\mathbf{x}_0^\star : \mathbf{w} \in \mathbb{C}^N\}$$

and $T^\perp$ be its orthogonal complement. T can be interpreted as the tangent space at $\mathbf{x}_0\mathbf{x}_0^\star$ to the manifold of symmetric matrices of rank one. Influenced by [CSV13], we use $\mathbf{X}_T$ and $\mathbf{X}_{T^\perp}$ to denote the projection of a matrix $\mathbf{X}$ onto the subspaces T and $T^\perp$ respectively.

Standard duality arguments in semidefinite programming show that the following set of conditions are sufficient for $\mathbf{x}_0\mathbf{x}_0^\star$ to be the unique optimizer to (4.9):

(i) *Condition 1*: $\mathbf{X} \in T \quad \& \quad \mathcal{A}(\mathbf{X}) = 0 \Rightarrow \mathbf{X} = 0$

(ii) *Condition 2*: There exists a *dual certificate* $\mathbf{W}$ in the range space of $\mathcal{A}^\star$ obeying:

- $\mathbf{W}\mathbf{x}_0 = 0$
- $\operatorname{rank}(\mathbf{W}) = N - 1$
- $\mathbf{W} \succcurlyeq 0$.

First, we will show that the measurement operator $\mathcal{A}$ obtained with the masks defined by $\mathbf{D}_1$ and $\mathbf{D}_2$ satisfies *Condition 1*.

The set of constraints $\mathcal{A}(\mathbf{X}) = 0$ fix the entries of the first row and column of $\mathbf{X}$ to 0, i.e., $X[0,m] = X[m,0] = 0 \quad \text{for} \quad 0 \leq m \leq N-1$. Since $\mathbf{X} \in T$, we can write $\mathbf{X} = \mathbf{x}_0\mathbf{w}^\star + \mathbf{w}\mathbf{x}_0^\star$ for some $\mathbf{w} = (w[0], w[1], \ldots, w[N-1])^T$, from which we infer $w[m] = icx_0[m]$ for some constant c. Hence, $\mathbf{X} = \mathbf{x}_0\mathbf{w}^\star + \mathbf{w}\mathbf{x}_0^\star = -ic\mathbf{x}_0\mathbf{x}_0^\star + ic\mathbf{x}_0\mathbf{x}_0^\star = 0$.

Next, we will show that *Condition 2* is satisfied. The range space of $\mathcal{A}^\star$ obtained with the masks defined by $\mathbf{D}_1$ and $\mathbf{D}_2$ is the set of all symmetric matrices whose principal submatrix obtained by removing the first row and column has Toeplitz structure (this can be easily seen by writing the dual of (4.9)). Suppose $\mathbf{z} = -(x_0[1], x_0[2], \ldots, x_0[N-1])^T / x_0[0]$ (well defined if $x_0[0] \neq 0$) and $\mathbf{I}_{N-1}$ is the identity matrix of size $N-1$. Consider the following dual certificate:

$$\mathbf{W} = \begin{bmatrix} \mathbf{z}^\star\mathbf{z} & \mathbf{z}^\star \\ \mathbf{z} & \mathbf{I}_{N-1} \end{bmatrix}. \tag{4.12}$$

$\mathbf{W}$ is in the range space of $\mathcal{A}^\star$ as $\mathbf{I}_{N-1}$ has Toeplitz structure. Also, $\mathbf{W}\mathbf{x}_0 = 0$ as $\mathbf{z}x_0[0] + (x_0[1], x_0[2], \ldots, x[N-1])^T = 0$. By writing out the characteristic equation, it is straightforward to see that the eigenvalues of $\mathbf{W}$ are $\{1 + \|\mathbf{z}\|^2, 1, 1, \ldots, 1, 0\}$. Hence, $\text{rank}(\mathbf{W}) = N-1$ and $\mathbf{W} \succcurlyeq 0$. This completes the proof. □

This result can also be extended to the setting with N point DFT measurements. Note that, if a signal is such that it has $\frac{N}{2}$ consecutive zeros in the beginning or the end, then its circular autocorrelation and autocorrelation are the same. Consider the following SDP formulation:

$$\begin{aligned} \text{minimize} \quad & \text{trace}(\mathbf{X}) \\ \text{subject to} \quad & \text{trace}(\mathbf{D}_r^\star \mathbf{B}_m \mathbf{D}_r \mathbf{X}) = b[m,r] \quad \text{for} \quad 0 \leq m \leq N-1 \quad \text{and} \quad 0 \leq r \leq R-1 \\ & \mathbf{X} \succcurlyeq 0, \end{aligned} \tag{4.13}$$

where the matrices $\mathbf{B}_m$ are given by

$$\mathbf{B}_{mgh} = \begin{cases} 1 & \text{if} \quad h - g = m = 0 \\ 1 & \text{if} \quad (h-g) \bmod N = m \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose the measurements are obtained using the masks defined by

$$d_3[n] = \begin{cases} 1 & 0 \le n \le \frac{N}{2} - 1 \\ 0 & \frac{N}{2} \le n \le N - 1 \end{cases} \tag{4.14}$$

$$d_4[n] = \begin{cases} 0 & n = 0 \\ 1 & 1 \le n \le \frac{N}{2} - 1 \\ 0 & \frac{N}{2} \le n \le N - 1 \end{cases}$$

$$d_5[n] = \begin{cases} 0 & 0 \le n \le \frac{N}{2} \\ 1 & \frac{N}{2} + 1 \le n \le N - 1 \end{cases}$$

$$d_6[n] = \begin{cases} 0 & 0 \le n \le \frac{N}{2} - 1 \\ 1 & \frac{N}{2} \le n \le N - 1 \end{cases}$$

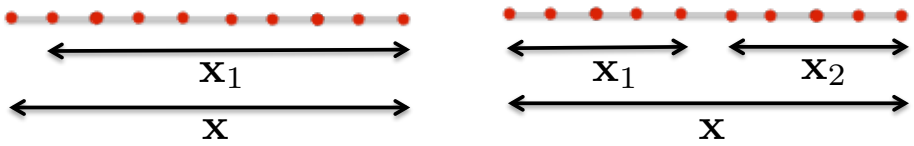
$$d_7[n] = \begin{cases} 0 & 0 \le n \le \frac{N}{4} - 1 \\ 1 & \frac{N}{4} \le n \le \frac{3N}{4} - 1 \\ 0 & \frac{3N}{4} \le n \le N - 1. \end{cases}$$

The matrices $\mathbf{D}_3$ and $\mathbf{D}_4$ are such that the nth diagonal element is 0 when $\frac{N}{2} \le n \le N - 1$. Consequently, the circular autocorrelation and the autocorrelation of the modulated signals $\mathbf{D}_3\mathbf{x}$ and $\mathbf{D}_4\mathbf{x}$ are the same. The values of $x_0[n]$ in the region $0 \le n \le \frac{N}{2} - 1$ can be inferred up to a global phase using calculations identical to the calculations following (4.10). Similarly, using the matrices $\mathbf{D}_5$ and $\mathbf{D}_6$, the values of $x_0[n]$ in the region $\frac{N}{2} \le n \le N - 1$ can be inferred up to a global phase. The matrix $\mathbf{D}_6$ resolves the relative phase between these two regions.

Theorem 4.3.2. *Consider any signal* $\mathbf{x}_0$ *such that* $x_0[0], x_0[\frac{N}{2} - 1], x_0[\frac{N}{2}] \neq 0$. *Suppose measurements are taken using the masks defined by* $\mathbf{D}_3, \mathbf{D}_4, \mathbf{D}_5$, $\mathbf{D}_6$ *and* $\mathbf{D}_7$. *The convex program (4.13) has a unique feasible point, namely,* $\mathbf{x}_0\mathbf{x}_0^\star$, *and hence* $\mathbf{x}_0$ *can be uniquely recovered up to a global phase.*

Proof. Consider the set of measurements obtained with the masks defined by $\mathbf{D}_3$ and $\mathbf{D}_4$. Since both these masks are zero throughout the region $\frac{N}{2} \le n \le N - 1$, Theorem 4.3.1 applies with N replaced by $\frac{N}{2}$. Hence, if $x[0] \neq 0$, then $X[n, m]$ in the region $0 \le n, m \le \frac{N}{2} - 1$ can be uniquely recovered.

Similarly, using the measurements obtained with the masks defined by $\mathbf{D}_5$ and $\mathbf{D}_6$, $X[n, m]$ in the region $\frac{N}{2} \le n, m \le N - 1$ can be uniquely recovered if $x[\frac{N}{2}] \neq 0$.



(a) (b)

Figure 4.3: An example of measurements using the proposed mask designs. (a) The autocorrelation of the signals $\mathbf{x}$ and $\mathbf{x}_1$ are obtained as measurements. (b) The autocorrelation of the signals $\mathbf{x}$, $\mathbf{x}_1$ and $\mathbf{x}_2$ are obtained as measurements.

The measurements obtained with the mask defined by $\mathbf{D}_7$ recover the value of $X[\frac{N}{2}-1, \frac{N}{2}]$. If $X[\frac{N}{2}-1, \frac{N}{2}] \neq 0$, then, given the aforementioned determined entries, it is straightforward to see that $\mathbf{x}_0\mathbf{x}_0^\star$ is the only feasible positive semidefinite completion. Hence, $\mathbf{x}_0$ can be uniquely identified up to a global phase. □

4.4 Design #2

The masks described in the previous section, albeit simple, have two practical drawbacks:

(a) The reconstruction is sensitive to the value of $x_0[0]$. In fact, if $x_0[0] = 0$, then the problem reduces to the standard phase retrieval problem with no additional measurements.

(b) In practice, it is not easy to accurately implement masks which have zero at only one location.

With this in mind, for the oversampled setup, we propose the following easy-to-implement design: Let $\mathbf{D}_8$, $\mathbf{D}_9$ and $\mathbf{D}_{10}$ be diagonal matrices with diagonal entries given by

$$d_8[n] = \begin{cases} 1 & 0 \leq n \leq \frac{N}{2}-1 \\ 0 & \frac{N}{2} \leq n \leq N-1 \end{cases} \tag{4.15}$$

$$d_9[n] = \begin{cases} 0 & 0 \leq n \leq \frac{N}{2}-1 \\ 1 & \frac{N}{2} \leq n \leq N-1 \end{cases}$$

$$d_{10}[n] = 1 \qquad 0 \leq n \leq N-1.$$

Observe that $\mathbf{D}_{10}$ corresponds to the measurements without using any mask, $\mathbf{D}_8$ corresponds to the measurements where the right half of the signal is blocked by the mask so that the autocorrelation of the left half of the signal is measured,

and $\mathbf{D}_9$ corresponds to the measurements where the left half is blocked and the autocorrelation of the right half is measured (see Fig. 4.3b for an example). Let $\mathbf{x}_1 = (x[0], x[1], \ldots, x[\frac{N}{2} - 1])^T$ and $\mathbf{x}_2 = (x[\frac{N}{2}], x[\frac{N}{2} + 1], \ldots, x[N-1])^T$.

Theorem 4.4.1. *Consider the set of signals* $\mathbf{x}_0 = (\mathbf{x}_1; \mathbf{x}_2)$ *such that* $z^{\frac{N}{2}-1}X_1(z)$ *and* $z^{\frac{N}{2}-1}X_2(z)$ *do not have any common factors, and* $x_0[0], x_0[\frac{N}{2}] \neq 0$. *Suppose oversampled measurements are taken with the masks defined by* $\mathbf{D}_8$, $\mathbf{D}_9$, *and* $\mathbf{D}_{10}$. *The convex program (4.9) has a unique feasible point, namely,* $\mathbf{x}_0\mathbf{x}_0^\star$, *almost surely.*

Proof. The intuition behind this mask construction is the following: The measurements corresponding to $\mathbf{D}_8$ and $\mathbf{D}_9$ provide $X_1(z)X_1^\star(z^{-\star})$ and $X_2(z)X_2^\star(z^{-\star})$ respectively. The measurement corresponding to $\mathbf{D}_{10}$ provides

$$\left(X_1(z) + z^{-\frac{N}{2}}X_2(z)\right)\left(X_1^\star(z^{-\star}) + z^{\frac{N}{2}}X_2^\star(z^{-\star})\right).$$

By subtracting out the known quantities $X_1(z)X_1^\star(z^{-\star})$ and $X_2(z)X_2^\star(z^{-\star})$, we can infer $z^{-\frac{N}{2}}X_2(z)X_1^\star(z^{-\star}) + z^{\frac{N}{2}}X_1(z)X_2^\star(z^{-\star})$. The first quantity only has terms which involve negative powers of z, and the second quantity only has terms which involve positive powers of z. Due to this, we can infer $X_2(z)X_1^\star(z^{-\star})$ and $X_1(z)X_2^\star(z^{-\star})$.

Essentially, the measurements provide the knowledge of the autocorrelations of $\mathbf{x}_1$, $\mathbf{x}_2$ and their cross-correlation. If the polynomials $X_1(z)$ and $X_2(z)$ do not have any common factors, then they can be uniquely reconstructed by looking at the common factors of $X_1(z)X_1^\star(z^{-\star})$ and $X_1(z)X_2^\star(z^{-\star})$, and $X_2(z)X_2^\star(z^{-\star})$ and $X_2(z)X_1^\star(z^{-\star})$ respectively.

The proof of this theorem is identical to the proof of Theorem 4.3.1.

The range space of $\mathcal{A}^\star$, with these measurements, is the set of all symmetric $N \times N$ matrices which are such that the submatrix corresponding to the $0 \leq n \leq \frac{N}{2} - 1$ rows and columns, $\frac{N}{2} \leq n \leq N-1$ rows and columns, $0 \leq n \leq \frac{N}{2} - 1$ rows and $\frac{N}{2} \leq n \leq N-1$ columns, and $\frac{N}{2} \leq n \leq N-1$ rows and $0 \leq n \leq \frac{N}{2} - 1$ columns are all Toeplitz.

We first show that the measurement operator $\mathcal{A}$ satisfies Condition 1. Since $\mathbf{X} \in T$, we can write $\mathbf{X} = \mathbf{x}_0\mathbf{w}^\star + \mathbf{w}\mathbf{x}_0^\star$ for some $\mathbf{w} = (w[0], w[1], \ldots, w[N-1])^T$. Therefore, $\mathcal{A}(\mathbf{X}) = 0$ can be equivalently written as $\mathbf{T}(\mathbf{x}_0)(\mathrm{Re}(\mathbf{w}^T), \mathrm{Im}(\mathbf{w}^T))^T = 0$, where $\mathbf{T}$ is a $2N \times 2N$ matrix which is a function of $\mathbf{x}_0$. The determinant of any submatrix of $\mathbf{T}$ is a rational function of $\mathrm{Re}(\mathbf{x}_0)$ and $\mathrm{Im}(\mathbf{x}_0)$. Therefore, it is either identically 0, or almost always nonzero. By substituting $\mathbf{x}_0 = (1, 0, \ldots, 0)^T$, one can see that

there exists an $N-1 \times N-1$ submatrix with nonzero determinant almost always. Since $\mathbf{w} = ic\mathbf{x}_0$ is in the null space of $\mathbf{T}$, we infer that the rank of $\mathbf{T}$ is almost always $N-1$. Consequently, we almost surely have $\mathbf{w} = ic\mathbf{x}_0$ as the only solution, which corresponds to $\mathbf{X} = -ic\mathbf{x}_0\mathbf{x}_0^\star + ic\mathbf{x}_0\mathbf{x}_0^\star = 0$.

We now show that Condition 2 is satisfied. We construct a dual certificate based on Sylvester matrices [Bit+78], which often come up in problems involving finding common roots of two polynomials [DBD12].

Let $\mathbf{S}$ be an $N \times N$ Sylvester matrix corresponding to the two polynomials $z^{\frac{N}{2}}X_1(z)$ and $z^{\frac{N}{2}}X_2(z)$, i.e.,

$$\mathbf{S} = \begin{bmatrix} x_2[0] & 0 & . & . & 0 & -x_1[0] & 0 & . & . & 0 \\ x_2[1] & x_2[0] & . & . & 0 & -x_1[1] & -x_1[0] & . & . & 0 \\ . & x_2[1] & . & . & . & . & -x_1[1] & . & . & . \\ . & . & . & . & . & . & . & . & . & . \\ x_2[\frac{N}{2}-1] & . & . & . & x_2[0] & -x_1[\frac{N}{2}-1] & . & . & . & -x_1[0] \\ 0 & x_2[\frac{N}{2}-1] & . & . & x_2[1] & 0 & -x_1[\frac{N}{2}-1] & . & . & -x_1[1] \\ 0 & 0 & . & . & . & 0 & 0 & . & . & . \\ . & 0 & . & . & . & . & 0 & . & . & . \\ . & . & . & . & x_2[\frac{N}{2}-1] & . & . & . & . & -x_1[\frac{N}{2}-1] \\ 0 & 0 & . & . & 0 & 0 & 0 & . & . & 0 \end{bmatrix}. \quad (4.16)$$

The rank of the Sylvester matrix is known to be $N-c$, where c is the number of common roots between the two polynomials [Bit+78]. In our setup, we have $\text{rank}(\mathbf{S}) = N-1$, as $z=0$ is the only common root between $z^{\frac{N}{2}}X_1(z)$ and $z^{\frac{N}{2}}X_2(z)$. By construction, we have $\mathbf{S}\mathbf{x}_0 = 0$. We consider the following dual certificate:

$$\mathbf{W} = \mathbf{S}^\star\mathbf{S}. \quad (4.17)$$

Clearly, $\mathbf{W}$ is positive semidefinite, has rank $N-1$ and $\mathbf{W}\mathbf{x}_0 = 0$. Also, $\mathbf{S}^\star\mathbf{S}$ is a "block" Toeplitz matrix which is in the range space of $\mathcal{A}^\star$. This completes the proof. □

Remark 1: The number of phaseless measurements considered by Design #1 is $4N$ ($2N$ measurements per mask). In fact, the effective number of phaseless measurements considered by Design #2 is $4N$ too, and not $6N$. This is due to the fact that the second and the third masks measure the autocorrelation of a signal of length $\frac{N}{2}$, and therefore use only N phaseless measurements (either the entire N point DFT or N measurements from the $2N$ point DFT).

Remark 2: Similar to Theorem 4.3.2, Theorem 4.4.1 can be extended to the N point DFT setup by making use of $2 \times 3 + 1 = 7$ masks.

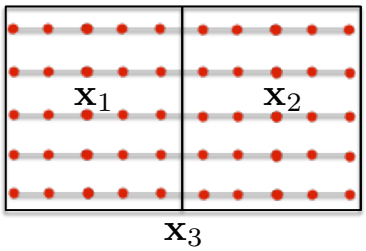


Figure 4.4: A $2D$ example of measurements using Design #2. The $2D$ autocorrelation of signals $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ are obtained as measurements.

4.5 Extension to $2D$

The results in this chapter can be immediately extended to $2D$ by making use of the trick described in Section 3.5. Let $\mathbf{x}$ be a two-dimensional signal with N_1 rows and N_2 columns, and $\mathbf{a}$ be its two-dimensional autocorrelation with $2N_1 - 1$ rows and $2N_2 - 1$ columns. As earlier, let $\mathbf{x}_{1D} = \text{vec}(\mathbf{x})$ denote the one-dimensional vector constructed by stacking the columns of $\mathbf{x}$ on top of each other. In Section 3.5, we showed that the one-dimensional autocorrelation of $\mathbf{x}_{1D}$ can be inferred from $\mathbf{a}$.

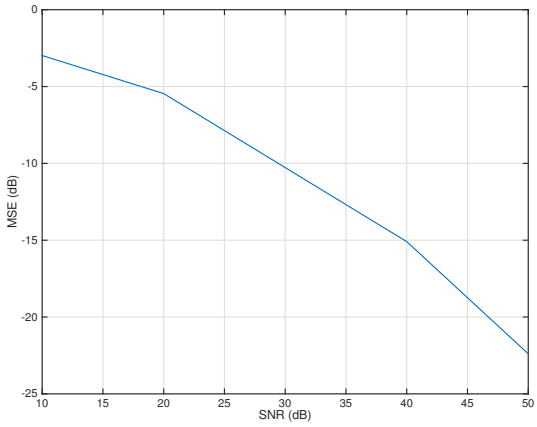
Therefore, the results of Design #1 can be generalized to $2D$ if the first mask is chosen such that its diagonal entries are equal to 1 everywhere, and the second mask is chosen such that $d[0, 0]$ is zero and the remaining diagonal entries are equal to 1. Similarly, the results of Design #2 can be generalized to $2D$ if the first mask is chosen such that its diagonal entries are equal to 1 everywhere, the second mask is chosen such that $d[n, m] = 0$ when $m \geq \frac{N_2}{2}$ and 1 otherwise, and the third mask is chosen such that $d[n, m] = 1$ when $m \geq \frac{N_2}{2}$ and 0 otherwise (see Fig. 4.4 for a pictorial example).

Remark: Due to the assumption $x_0[0] \neq 0$ in Theorem 4.3.1 and $x_0[0], x_0[\frac{N}{2}] \neq 0$ in Theorem 4.4.1, there are no ambiguities due to time-shift.

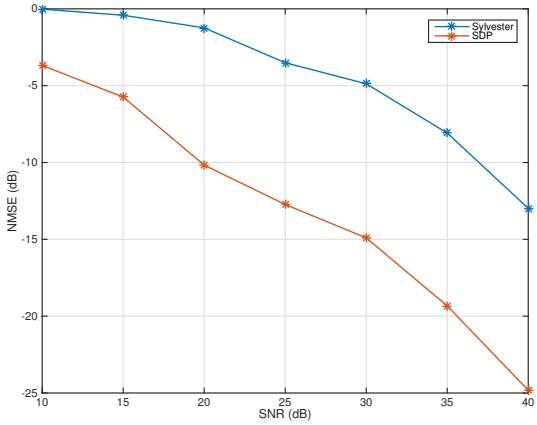
4.6 Numerical Simulations

In this section, we demonstrate the performance of the proposed methods in the noisy setting. Signals of length $N = 32$ are chosen from an i.i.d. complex normal distribution, and the normalized mean squared-error is plotted as a function of SNR.

The performance of the SDP method when measurements are obtained using Design #1 and #2 is plotted in Fig. 4.5a and 4.5b respectively. The plots show that the reconstruction is stable. For Design #2, we also evaluate the performance of the Sylvester matrix based common factor finding approach proposed in [Ton+95],



(a)



(b)

Figure 4.5: NMSE vs SNR of the SDP method for (a) Design #1 (b) Design #2.

where the authors aim to do blind channel estimation (the mathematical problem encountered is the same). The simulations clearly demonstrate the superior ability of the SDP method.

4.7 Conclusions and Future Work

In this chapter, we showed that oversampled measurements from two specific masks or three easy-to-implement masks are enough for the SDP method to provably reconstruct most signals. For both these designs, we showed that the total number of phaseless measurements considered is $4N$. Further, simulations showed that the reconstruction is stable in the noisy setting.

A natural direction for future study is stability analysis. In [JEH15d], we provided

loose bounds on the reconstruction error in the noisy setting. Given the practicality of the proposed designs, tight bounds on the reconstruction error would be very insightful in various applications.

Providing theoretical guarantees for $2D$ signals by viewing them as $2D$ signals instead of vectorizing them into $1D$ signals is another interesting problem. The approach used in Section 4.5 is suboptimal due to the fact that information from a lot of $2D$ measurements are not considered.

Chapter 5

STFT PHASE RETRIEVAL

In this chapter, we consider STFT phase retrieval, which is the problem of recovering a signal from its STFT magnitude. The STFT of a signal is defined as follows:

Let $\mathbf{w} = (w[0], w[1], \ldots, w[W-1])^T$ be a window of length W such that it has nonzero values only within the interval $[0, W-1]$. The STFT of $\mathbf{x}$ with respect to $\mathbf{w}$, denoted by $\mathbf{Y}_w$, is defined as

$$Y_w[m,r] = \sum_{n=0}^{N-1} x[n]w[rL-n]e^{-i2\pi\frac{mn}{N}} \quad \text{for} \quad 0 \le m \le N-1 \quad \text{and} \quad 0 \le r \le R-1, \tag{5.1}$$

where the parameter L denotes the separation in time between adjacent short-time sections and the parameter $R = \lceil \frac{N+W-1}{L} \rceil$ denotes the number of short-time sections considered.

The STFT can be interpreted as follows: Suppose $\mathbf{w}_r$ denotes the signal obtained by shifting the flipped window $\mathbf{w}$ by rL time units (i.e., $w_r[n] = w[rL-n]$) and $\circ$ is the Hadamard (element-wise) product operator. The rth column of $\mathbf{Y}_w$, for $0 \le r \le R-1$, corresponds to the N point DFT of $\mathbf{x} \circ \mathbf{w}_r$. In essence, the window is flipped and slid across the signal (see Figure 5.1 for a pictorial representation), and $\mathbf{Y}_w$ corresponds to the Fourier transform of the windowed signal recorded at regular intervals. This interpretation is known as the *sliding window* interpretation.

Let $\mathbf{Z}_w$ be the $N \times R$ measurements corresponding to the magnitude-square of the STFT of $\mathbf{x}$ with respect to $\mathbf{w}$ so that $Z_w[m,r] = |Y_w[m,r]|^2$. Let $\mathbf{W}_r$, for $0 \le r \le R-1$, be the $N \times N$ diagonal matrix with diagonal elements $(w_r[0], w_r[1], \ldots, w_r[N-1])$. STFT phase retrieval can be mathematically stated as:

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && Z_w[m,r] = |\langle \mathbf{f}_m, \mathbf{W}_r\mathbf{x} \rangle|^2 \end{aligned} \tag{5.2}$$

for $0 \le m \le N-1$ and $0 \le r \le R-1$, where $\mathbf{f}_m$ is the conjugate of the mth column of the N point DFT matrix and $\langle ., . \rangle$ is the inner product operator. In fact, STFT phase retrieval can be equivalently stated by only considering the measurements corresponding to $0 \le r \le R-1$ and $1 \le m \le M$, for any parameter M satisfying

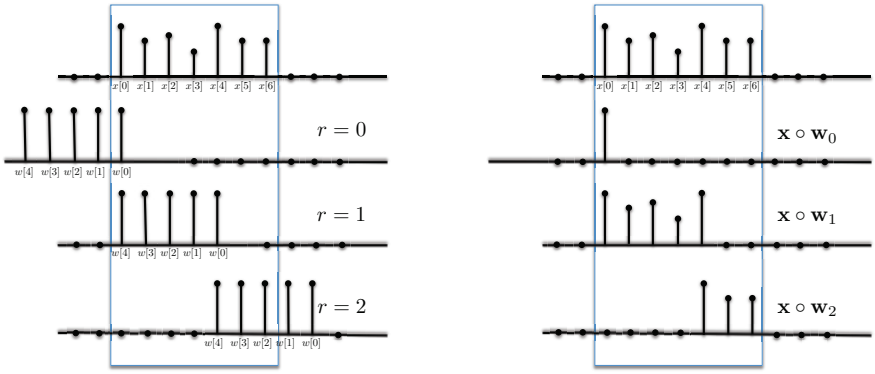


Figure 5.1: Sliding window interpretation of the STFT for $N = 7$, $W = 5$ and $L = 4$. The shifted window overlaps with the signal for 3 shifts, and hence $R = 3$ short-time sections are considered.

$2W \leq M \leq N$ (see Section 9.1 for details). This equivalence significantly reduces the number of measurements obtained when $W \ll N$[1].

The motivation for STFT phase retrieval is two-fold: First, a lot of redundancy can be introduced in the magnitude-only measurements by maintaining a substantial overlap between adjacent short-time sections. As we will see, the redundancy offered by the STFT enables unique, efficient and robust recovery in many cases. Second, it is possible to obtain such measurements in many phase retrieval applications by introducing certain modifications in the measurement systems. In the following, we first describe the origin of STFT phase retrieval in two such applications, and then present our results.

5.1 Ptychography/ Fourier Ptychography

Ptychography is a technology invented by Walter Hoppe [Fra+12] as a means to obtain additional information about the underlying signal, in order to overcome the uniqueness issues of phase retrieval in diffraction imaging. Ptychography, along with developments in detector and computing technologies, have resulted in X-ray, optical and electron microscopes with increased spatial resolution without the need for advanced lenses. A typical ptychography setup, courtesy of [Nas+14], is described in Fig. 5.2.

Let $\psi(x, y)$ denote the object, centered at the origin. Also, let the direction of light be parallel to the z-axis, and the plane of the object and the two-dimensional detector

[1] We further reduce the number of measurements per short-time section through super-resolution. In particular, we consider the setup with $2L \leq W \leq \frac{N}{2}$ and $4L \leq M \leq N$.

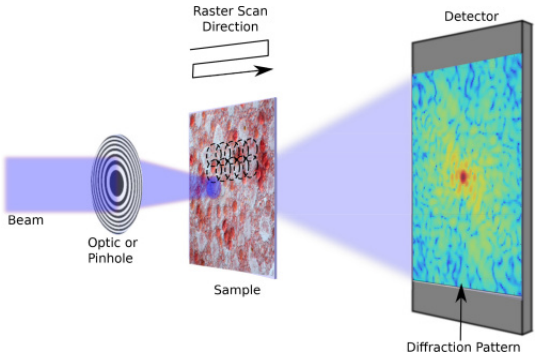


Figure 5.2: A typical ptychography setup (courtesy of [Nas+14]).

be perpendicular to the z-axis such that $z = 0$ and $z = z'$ respectively. In a standard diffraction imaging setup, the entire object is illuminated and the diffraction patterns are recorded. In ptychography, only a small part of the sample is illuminated (as seen in Fig. 5.2) in each recording. This can be done by using focusing lenses, or by physically blocking the light source using masks. Multiple diffraction patterns are collected by moving the object in step-sizes much smaller than the window size, so that there is a substantial overlap between adjacent measurements.

Let $\mathcal{W}_r(x, y)$ be the indicator function, such that it has a value 1 if (x, y) is illuminated during the rth measurement, and 0 otherwise. By using Huygens principle (same arguments as in Section 1.1), if $\psi_{trans,r}(x, y)$ denotes the secondary source at $(x, y, 0)$ produced by $\psi(x, y)$, then we have $\psi_{trans,r}(x, y) = \mathcal{W}_r(x, y)\psi(x, y)$.

Also, let $I_r(x', y')$ denote the intensities recorded during the rth measurement at (x', y', z') for various (x', y'). Through the Fraunhofer approximation (see Section 1.1), we have

$$\begin{aligned} I_r(x', y') &\propto \left| \iint \psi_{trans,r}(x, y) e^{i\frac{2\pi}{\lambda}\frac{-x'x-y'y}{z'}} dxdy \right|^2 \\ &\propto \left| \iint \mathcal{W}_r(x, y)\psi(x, y) e^{i\frac{2\pi}{\lambda}\frac{-x'x-y'y}{z'}} dxdy \right|^2 . \end{aligned} \tag{5.3}$$

If the measurements are recorded at regular intervals, i.e., the step size between adjacent measurements remains constant, then these measurements are precisely the two-dimensional STFT magnitude-squares of $\psi(x, y)$.

Observe that the ptychography setup involves moving parts, which is suboptimal, as it requires precision control over actuation, optical alignment and motion tracking. More recently, a related technology called Fourier ptychography was proposed in [ZHY13; Ou+13] as a solution to this problem. The setup is detailed in Fig. 5.3.

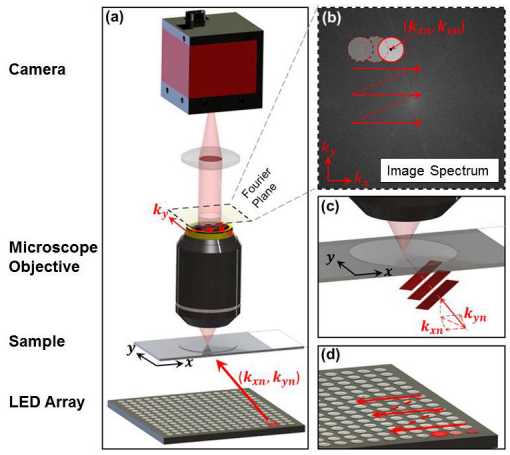


Figure 5.3: Fourier ptychography setup (courtesy of [ZHY13]).

In Fourier ptychography, instead of recording measurements from a single source, measurements are recorded from a two-dimensional array of sources (say, located at $z = -f$).

Suppose the source at $(x_r, y_r, -f)$ is the only source which is on. If $(x_r, y_r) \neq (0, 0)$, then different points (x, y) in the object are at a different distance from this source. Due to this, while the amplitude of the illumination is uniform, there is a phase delay proportional to the distance from this source. In particular, the illumination of the sample at (x, y) is proportional to $e^{i\frac{2\pi}{\lambda}\frac{x_r x + y_r y}{f}}$ (see [HY14] for details). Consequently, the secondary source $\psi_{trans,r}(x, y)$ is equal to $e^{i\frac{2\pi}{\lambda}\frac{x_r x + y_r y}{f}}\psi(x, y)$.

A lens, with focal length f, is used to Fourier transform the wave due to this secondary source (see Section 1.2, Fourier plane in Fig. 5.3). In the Fourier plane, the wave is given by $\hat{\psi}(\frac{x'-x_r}{\lambda f}, \frac{y'-y_r}{\lambda f})$, where $\hat{\psi}$ is the two-dimensional Fourier transform of the object.

For different values of (x_r, y_r), the wave in the Fourier plane corresponds to different shifts of $\hat{\psi}$. Only a small part of $\hat{\psi}(\frac{x'-x_r}{\lambda f}, \frac{y'-y_r}{\lambda f})$ is allowed to pass through and the diffraction pattern is recorded using CCD cameras (a fixed mask physically blocks most of the wave). If $\mathcal{W}(x', y')$ denotes the indicator function, such that it has a value 1 if the wave at (x', y') is allowed to pass through, and 0 otherwise, then the measurements are the magnitude-squares of the Fourier transform of $\mathcal{W}(x', y')\hat{\psi}(\frac{x'-x_r}{\lambda f}, \frac{y'-y_r}{\lambda f})$. Observe that the role of the object space and Fourier space are permuted.

If the sources are placed uniformly in the cartesian grid, then these measurements

are precisely the two-dimensional STFT magnitude-squares of $\psi(x, y)$.

5.2 Contributions

In this chapter, our contribution is two-fold:

(i) Uniqueness guarantees: Researchers have previously developed conditions under which the STFT magnitude uniquely identifies signals up to a global phase. However, either prior information on the signal is assumed in order to provide the guarantees, or the guarantees are limited. For instance, the results provided in [NQL83] require exact knowledge of a small portion of the underlying signal. In [Eld+15], the guarantees developed are for the setup in which adjacent short-time sections differ in only one index. These limitations are primarily due to a small number of adversarial signals which cannot be uniquely identified from their STFT magnitude. Here, in contrast, we develop conditions under which the STFT magnitude is an *almost surely* unique signal representation. In particular, we show that, with the exception of a set of signals of measure zero, non-vanishing signals can be uniquely identified up to a global phase from their STFT magnitude if adjacent short-time sections overlap (Theorem 5.3.1). We then extend this result to incorporate sparse signals which have a limited number of consecutive zeros (Corollary 5.3.1).

(ii) Recovery algorithms: Researchers have previously developed efficient iterative algorithms based on classic optimization frameworks to solve the STFT phase retrieval problem. Examples include the Griffin-Lim (GL) algorithm [GL84] and STFT-GESPAR for sparse signals [Eld+15]. While these techniques work well in practice, they do not have theoretical guarantees. In [JEH15b] and [SS12], a semidefinite relaxation based STFT phase retrieval algorithm, called STliFT (see Algorithm 6 below), was proposed. In this work, we conduct extensive numerical simulations and provide theoretical guarantees for STliFT. In particular, we conjecture that STliFT can recover most non-vanishing signals up to a global phase from their STFT magnitude if adjacent short-time sections differ in at most half the indices (Conjecture 5.4.1). When this condition is satisfied, we argue that one can super-resolve (i.e., discard high frequency measurements) and reduce the number of measurements to $(4 + o(1))N$, where N is the length of the complex signal. Therefore, STliFT recovers most non-vanishing signals uniquely, efficiently and robustly, using an order-wise optimal number of phaseless measurements.

We prove this conjecture for the setup in which the exact knowledge of a small portion of the underlying signal is available (Theorem 5.4.1). For particular choices

of STFT parameters, this portion vanishes asymptotically, due to which this setup is asymptotically reasonable. We also prove this conjecture for the case in which adjacent short-time sections differ in only one index (Theorem 5.4.2). We then extend these results to incorporate sparse signals which have a limited number of consecutive zeros (Corollary 5.4.1).

5.3 Uniqueness

In this section, we first review the existing results regarding uniqueness of STFT phase retrieval, and then present our results. A signal $\mathbf{x}$ is non-vanishing if $x[n] \neq 0$ for each $0 \leq n \leq N-1$. Similarly, a window $\mathbf{w}$ is called non-vanishing if $w[n] \neq 0$ for all $0 \leq n \leq W-1$. These results are summarized in Table 5.1.

First, we argue that, for $W < N$ (which is typically the case), $L < W$ is a necessary condition in order to be able to uniquely identify most signals: If $L > W$, then the STFT magnitude does not contain any information from some locations of the signal, because of which most signals cannot be uniquely identified. If $L = W$, then the adjacent short-time sections do not overlap and hence STFT phase retrieval is equivalent to a series of non-overlapping phase retrieval problems. Consequently, as in the case of phase retrieval, most 1D signals are not uniquely identifiable. For higher dimensions (2D and above), almost all windowed signals corresponding to each of the short-time sections are uniquely identified up to trivial ambiguities if $\mathbf{w}$ is non-vanishing. However, since there is no way of establishing relative phase, time-shift or conjugate-flip between the windowed signals corresponding to the various short-time sections, most signals cannot be uniquely identified. For example, suppose we choose $L = W = 2$ and $w[n] = 1$ for all $0 \leq n \leq W-1$. Consider the signal $\mathbf{x}_1 = (1, 2, 3)^T$ of length $N = 3$. Signals $\mathbf{x}_1$ and $\mathbf{x}_2 = (1, -2, -3)^T$ have the same STFT magnitude. In fact, more generally, signals $\mathbf{x}_1$ and $(1, e^{i\phi}2, e^{i\phi}3)^T$, for any ϕ, have the same STFT magnitude.

For some specific choices of $\{\mathbf{w}, L\}$, it has been shown that all non-vanishing signals can be uniquely identified from their STFT magnitude up to a global phase. In [Eld+15], it is proven that the STFT magnitude uniquely identifies non-vanishing signals up to a global phase for $L = 1$ if the window $\mathbf{w}$ is chosen such that the N point DFT of $(|w[0]|^2, |w[1]|^2, \ldots, |w[N-1]|^2)$ is non-vanishing, $2 \leq W \leq \frac{N+1}{2}$ and $W - 1$ is coprime with N. In [NQL83], the authors prove that if the first L samples are known a priori, then the STFT magnitude can uniquely identify non-vanishing signals for any L if the window $\mathbf{w}$ is chosen such that it is non-vanishing

Non-vanishing signals $\{x[n] \neq 0$ for all $0 \leq n \leq N-1\}$	Uniqueness if the first L samples are known a priori, $2L \leq W \leq \frac{N}{2}$ and $\mathbf{w}$ is non-vanishing [NQL83]
	Uniqueness up to a global phase if $L = 1$, $2 \leq W \leq \frac{N+1}{2}$, $W-1$ coprime with N and mild conditions on $\mathbf{w}$ [Eld+15]
	Uniqueness up to a global phase for almost all signals if $L < W \leq \frac{N}{2}$ and $\mathbf{w}$ is non-vanishing [JEH15c]
Sparse signals $\{x[n] = 0$ for at least one $0 \leq n \leq N-1\}$	Uniqueness for signals with at most $W - 2L$ consecutive zeros if the first L samples, starting from the first nonzero sample, are known a priori, $2L \leq W \leq \frac{N}{2}$ and $\mathbf{w}$ is non-vanishing [NQL83]
	No uniqueness for most signals with W consecutive zeros [Eld+15]
	Uniqueness up to a global phase and time-shift for almost all signals with less than $\min\{W-L, L\}$ consecutive zeros if $L < W \leq \frac{N}{2}$ and $\mathbf{w}$ is non-vanishing [JEH15c]

Table 5.1: Uniqueness results for STFT phase retrieval ($2W \leq M \leq N$).

and $2L \leq W \leq \frac{N}{2}$.

We prove the following result for non-vanishing signals:

Theorem 5.3.1. *Almost all non-vanishing signals can be uniquely identified up to a global phase from their STFT magnitude if* $\{\mathbf{w}, L, M\}$ *satisfy*

(i) $\mathbf{w}$ *is non-vanishing*

(ii) $L < W \leq \frac{N}{2}$

(iii) $2W \leq M \leq N$.

Proof. The proof is based on a technique commonly known as *dimension counting*. The outline is as follows (see Section 9.2 for details):

Consider the short-time sections r and $r+1$. Since adjacent short-time sections overlap (due to $L < W$), there exists at least one index, say n_0, where both $\mathbf{x} \circ \mathbf{w}_r$ and $\mathbf{x} \circ \mathbf{w}_{r+1}$ have nonzero values.

Since $W \leq \frac{N}{2}$, there can be at most 2^W distinct windowed signals $\mathbf{x} \circ \mathbf{w}_r$ up to a phase that have the same Fourier magnitude [Hof64]. Consequently, $|x[n_0]|$ is restricted to 2^W values by the rth column of the STFT magnitude (let $\mathcal{S}_r$ denote the set of these

values). Similarly, $|x[n_0]|$ is restricted to 2^W values by the $r+1$th column of the STFT magnitude (denote the set of these values by $\mathcal{S}_{r+1}$).

By construction, $\mathcal{S}_r \cap \mathcal{S}_{r+1} \neq \phi$ as the STFT magnitude is generated by an underlying signal $\mathbf{x}_0$, i.e., $|x_0[n_0]| \in \mathcal{S}_r \cap \mathcal{S}_{r+1}$. Using Lemma 9.2.1 and Theorem 9.2.1, we show that, for almost all non-vanishing signals, $\mathcal{S}_r \cap \mathcal{S}_{r+1}$ has cardinality one. In other words, $\mathbf{x} \circ \mathbf{w}_r$ is uniquely identified up to a phase almost surely.

Since adjacent short-time sections overlap, non-vanishing signals are uniquely identified up to a *global* phase from the knowledge of $\mathbf{x} \circ \mathbf{w}_r$ (up to a phase) for $0 \leq r \leq R-1$ if $\mathbf{w}$ is non-vanishing. □

Sparse Signals

While the aforementioned results provide guarantees for non-vanishing signals, they do not say anything about sparse signals. Reconstruction of sparse signals involves certain challenges which are not encountered in the reconstruction of non-vanishing signals.

The following example is provided in [Eld+15] to show that the time-shift ambiguity cannot be resolved for some classes of sparse signals and some choices of $\{\mathbf{w}, L\}$: Suppose $\{\mathbf{w}, L\}$ is chosen such that $L \geq 2$, W is a multiple of L and $w[n] = 1$ for all $0 \leq n \leq W-1$. Consider a signal $\mathbf{x}_1$ of length $N \geq L+1$ such that it has nonzero values only within an interval of the form $[(t-1)L+1, (t-1)L+L-p] \subset [0, N-1]$ for some integers $1 \leq p \leq L-1$ and $t \geq 1$. The signal $\mathbf{x}_2$ obtained by time-shifting $\mathbf{x}_1$ by $q \leq p$ units (i.e., $x_2[i] = x_1[i-q]$) has the same STFT magnitude. The issue with this class of sparse signals is that the STFT magnitude is identical to the Fourier magnitude because of which the time-shift and conjugate-flip ambiguities cannot be resolved.

It is also shown that some sparse signals cannot be uniquely recovered even up to trivial ambiguities for some choices of $\{\mathbf{w}, L\}$ using the following example: Consider two non-overlapping intervals $[u_1, v_1], [u_2, v_2] \subset [0, N-1]$ such that $u_2 - v_1 > W$, and take a signal $\mathbf{x}_1$ supported on $[u_1, v_1]$ and $\mathbf{x}_2$ supported on $[u_2, v_2]$. The magnitude-square of the STFT of $\mathbf{x}_1 + \mathbf{x}_2$ and of $\mathbf{x}_1 - \mathbf{x}_2$ are equal for any choice of L. The difficulty with this class of sparse signals is that the two intervals with nonzero values are separated by a distance greater than W because of which there is no way of establishing relative phase using a window of length W.

These examples demonstrate the fact that sparse signals are harder to recover than non-vanishing signals in this setup. Since the aforementioned issues are primarily

due to a large number of consecutive zeros, the uniqueness guarantees for non-vanishing signals have been extended to incorporate sparse signals with limits on the number of consecutive zeros. In [NQL83], it was shown that if L consecutive samples, starting from the first nonzero sample, are known a priori, then the STFT magnitude can uniquely identify signals with less than $W - 2L$ consecutive zeros for any L if the window $\mathbf{w}$ is chosen such that it is non-vanishing and $2L \leq W \leq \frac{N}{2}$.

Below, we extend Theorem 5.3.1 to prove the following result for sparse signals:

Corollary 5.3.1. *Almost all sparse signals with less than* $\min\{W-L, L\}$ *consecutive zeros can be uniquely identified up to a global phase and time-shift from their STFT magnitude if* $\{\mathbf{w}, L, M\}$ *satisfy*

(i) $\mathbf{w}$ *is non-vanishing*

(ii) $L < W \leq \frac{N}{2}$

(iii) $2W \leq M \leq N$.

Proof. The $\min\{W - L, L\}$ bound on consecutive zeros ensures the following: For sufficient pairs of adjacent short-time sections, there is at least one index among the overlapping and non-overlapping indices respectively, where the underlying signal has a nonzero value. We refer the readers to Section 9.3 for details. □

5.4 STliFT

In this section, we first discuss the alternating projection based algorithm, and then present our results for the SDP algorithm.

Alternating Projections

The classic alternating projection algorithm to solve phase retrieval [GS72] has been adapted to solve STFT phase retrieval by Griffin and Lim [GL84]. To this end, STFT phase retrieval is reformulated as the following least-squares problem:

$$\min_{\mathbf{x}} \quad \sum_{r=0}^{R-1} \sum_{m=0}^{N-1} \left(\sqrt{Z_w[m,r]} - |\langle \mathbf{f}_m, \mathbf{W}_r \mathbf{x} \rangle| \right)^2. \tag{5.4}$$

The Griffin-Lim (GL) algorithm attempts to minimize this objective by starting with a random initialization and imposing the time domain and STFT magnitude constraints alternately using projections. The details of the various steps are summarized in Algorithm 5.

The objective is shown to be monotonically decreasing as the iterations progress. An important feature of the GL algorithm is its empirical ability to converge to the global minimum when there is substantial overlap between adjacent short-time sections [BE15]. However, no theoretical recovery guarantees are available. To establish such guarantees, we rely on a semidefinite relaxation approach.

Algorithm 5 Griffin-Lim (GL) Algorithm

Input: STFT magnitude-square measurements $\mathbf{Z}_w$ and window $\mathbf{w}$
Output: Estimate $\hat{\mathbf{x}}$ of the underlying signal
Initialize: Choose a random input signal $\mathbf{x}^{(0)}$, $\ell = 0$
while halting criterion false **do**
$\ell \leftarrow \ell + 1$
Compute the STFT of $\mathbf{x}^{(\ell-1)}$: $Y_w^{(\ell)}[m,r] = \sum_{n=0}^{N-1} x^{(\ell-1)}[n] w[rL-n] e^{-i2\pi \frac{mn}{N}}$
Impose STFT magnitude constraints: $Y_w'^{(\ell)}[m,r] = \frac{Y_w^{(\ell)}[m,r]}{\left|Y_w^{(\ell)}[m,r]\right|} \sqrt{Z_w[m,r]}$
Compute the inverse DFT of $\mathbf{Y}_w'^{(\ell)}$ for each short-time section to obtain windowed signals $\mathbf{x}_r'^{(\ell)}$
Impose time domain constraints to obtain $\mathbf{x}^{(\ell)}$: $x^{(\ell)}[n] = \frac{\sum_r x_r'^{(\ell)}[n] w^\star[rL-n]}{\sum_r |w[rL-n]|^2}$
end while
return $\hat{\mathbf{x}} \leftarrow \mathbf{x}^{(\ell)}$

Semidefinite Relaxation

A semidefinite relaxation based STFT phase retrieval algorithm (STliFT) was explored in [SS12; JEH15b; Hor+15]. The details of the algorithm are provided in Algorithm 6. In the following, we develop conditions on $\{\mathbf{x}_0, \mathbf{w}, L, M\}$ which ensure that the convex program (5.5) has $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$ as the unique solution. Consequently, under these conditions, STliFT uniquely recovers the underlying signal up to a global phase.

Based on extensive numerical simulations, we conjecture the following:

Conjecture 5.4.1. *The convex program (5.5) has a unique solution* $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$*, for most non-vanishing signals* $\mathbf{x}_0$*, if*

(i) $\mathbf{w}$ *is non-vanishing*

(ii) $2L \leq W \leq \frac{N}{2}$

(iii) $4L \leq M \leq N$.

Algorithm 6 STliFT

Input: STFT magnitude measurements $Z_w[m, r]$ for $1 \leq m \leq M$ and $0 \leq r \leq R-1$, $\{\mathbf{w}, L\}$.

Output: Estimate $\hat{\mathbf{x}}$ of the underlying signal $\mathbf{x}_0$.

- Obtain $\hat{\mathbf{X}}$ by solving:

$$\begin{aligned} &\text{minimize} && \operatorname{trace}(\mathbf{X}) \\ &\text{subject to} && Z_w[m, r] = \operatorname{trace}(\mathbf{W}_r^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{W}_r \mathbf{X}) \\ & && \mathbf{X} \succcurlyeq 0 \end{aligned} \tag{5.5}$$

  for $1 \leq m \leq M$ and $0 \leq r \leq R-1$.

- Return $\hat{\mathbf{x}}$, where $\hat{\mathbf{x}}\hat{\mathbf{x}}^\star$ is the best rank-one approximation of $\hat{\mathbf{X}}$.

The number of phaseless measurements considered can be calculated as follows: The total number of short-time sections is $\lceil \frac{N+W-1}{L} \rceil$. For each short-time section, $M = 4L$ phaseless measurements are sufficient. Hence, the total number of phaseless measurements is $\lceil \frac{N+W-1}{L} \rceil \times 4L \leq 4\,(N+W)+2W$. Consequently, when $W = o(N)$, this number is $(4+o(1))N$, which is order-wise optimal. In fact, in generalized phase retrieval, it is conjectured that $(4-o(1))N$ phaseless measurements are necessary [Bal+09; BCE06].

The proof techniques used in [Can+15] and [CSV13] are not applicable in the STFT setup. In [Can+15] and [CSV13], the measurement vectors are chosen from a random distribution such that they satisfy the restricted isometry property. Furthermore, the randomness in the measurement vectors is used to construct approximate dual certificates based on concentration inequalities. In the STFT setup, testing whether the given measurement vectors satisfy the restricted isometry property is difficult. Also, due to the lack of randomness in the measurement vectors, a different approach is required to construct dual certificates.

In the following, we develop a proof technique for the STFT setup, and use it to prove Conjecture 5.4.1, with additional assumptions.

Theorem 5.4.1. *The convex program (5.5) has a unique feasible matrix* $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$, *for almost all non-vanishing signals* $\mathbf{x}_0$, *if*

(i) $\mathbf{w}$ *is non-vanishing*

(ii) $2L \leq W \leq \frac{N}{2}$

(iii) $4L \leq M \leq N$

(iv) $x_0[n]$ *for* $0 \leq n \leq \left\lfloor \frac{L}{2} \right\rfloor$ *is known a priori.*

Proof. See Section 9.4. □

While it is sufficient to show that (5.5) has a unique solution $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$, observe that Theorem 5.4.1 ensures that (5.5) has a unique feasible matrix. This is a stronger condition, and as a consequence, the choice of the objective function does not matter in the noiseless setting. While this might suggest that the requirements of the setup are strong, we argue that it is not the case. In fact, this phenomenon is also observed in generalized phase retrieval (Section 1.3 in [CSV13]) and phase retrieval using random masks (Theorem 1.1 in [Can+15]).

Theorem 5.4.1 assumes prior knowledge of the first $\lceil \frac{L}{2} \rceil$ samples, i.e., half of the second short-time section is required to be known a priori. This is not a lot of prior information if $W \ll N$, which is typically the case. When $W = o(N)$, the fraction of the signal that is required to be known a priori is less than $\frac{W}{N}$, which tends to 0 as $N \to \infty$.

Theorem 5.4.2. *The convex program (5.5) has a unique feasible matrix* $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$*, for almost all non-vanishing signals* $\mathbf{x}_0$*, if*

(i) $\mathbf{w}$ *is non-vanishing*

(ii) $2 \leq W \leq \frac{N}{2}$

(iii) $4 \leq M \leq N$

(iv) $L = 1$.

Proof. This is a direct consequence of Theorem 5.4.1. The value of $|x_0[0]|$ (and hence $x_0[0]$, without loss of generality) can be inferred from the STFT magnitude if $L = 1$. □

When $L = 1$, the number of phaseless measurements is $4(N + W)$, which is again order-wise optimal. For example, when $W = 2$, at most $4N + 8$ phaseless measurements are considered. Unlike Theorem 5.4.1, no prior information is necessary.

Theorems 5.4.1 and 5.4.2 can be seamlessly extended to incorporate sparse signals:

Corollary 5.4.1. *The convex program (5.5) has a unique feasible matrix* $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$, *for almost all sparse signals* $\mathbf{x}_0$ *which have at most* $W - 2L$ *consecutive zeros, if*

(i) $\mathbf{w}$ *is non-vanishing*

(ii) $2L \leq W \leq \frac{N}{2}$

(iii) $4L \leq M \leq N$

(iv) Either $L = 1$ *or* $x_0[n]$ *for* $i_0 \leq n < i_0 + L$ *is known a priori, where* i_0 *is the smallest index such that* $x_0[i_0] \neq 0$.

Proof. See Section 9.4. □

5.5 Stability

In practice, the measurements are contaminated by additive noise, i.e., the measurements are of the form

$$Z_w[m, r] = \left|\langle \mathbf{f}_m, \mathbf{W}_r\mathbf{x}\rangle\right|^2 + z[m, r]$$

for $1 \leq m \leq M$ and $0 \leq r \leq R - 1$, where $\mathbf{z}_r = (z[0, r], z[1, r], \ldots, z[M - 1, r])^T$ is the additive noise corresponding to the rth short-time section and $4L \leq M \leq N$. STliFT, in the noisy setting, can be implemented as follows: Suppose $\|\mathbf{z}_r\|_2 \leq \eta$ for all $0 \leq r \leq R - 1$. The constraints in the convex program (5.5) can be replaced by

$$\sum_{m=1}^{M} \left(Z_w[m, r] - \text{trace}(\mathbf{W}_r^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{W}_r \mathbf{X}) \right)^2 \leq \eta^2 \tag{5.6}$$

for $0 \leq r \leq R - 1$. We recommend the use of trace minimization as the objective function. Numerical simulations strongly suggest that the reconstruction is stable in the noisy setting.

5.6 Extension to $2D$

We refer the readers to Sections 3.5 and 4.5. In summary, the arguments in this chapter can be directly extended to $2D$, if the adjacent short-time sections are such that they have identical rows and differ in the appropriate number of columns, or if they have identical columns and differ in the appropriate number of rows. For example, for STliFT to work, the adjacent short-time sections should be such that they share the same rows and differ in at most 50% of the columns, or share the same columns and differ in at most 50% of the rows.

5.7 Numerical Simulations

In this section, we demonstrate the empirical abilities of STliFT using numerical simulations.

In the first set of simulations, we evaluate the performance of STliFT and GL algorithm as a function of window and shift lengths. We choose $N = 32$, and vary $\{L, W\}$. For each choice of $\{L, W\}$, we consider $M = 4L$ phaseless measurements and perform 100 trials. In every trial, we choose a random signal such that the values in each location are drawn from an i.i.d. standard complex normal distribution. We select the window $\mathbf{w}$ such that $w[n] = 1$ for all $0 \leq n \leq W - 1$. The probability of successful recovery as a function of $\{L, W\}$ is plotted in Fig. 5.4a and 5.4b respectively.

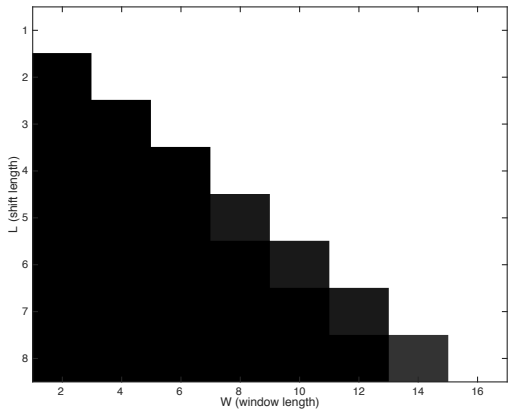
Observe that STliFT successfully recovers the underlying signal with very high probability when $2L \leq W \leq \frac{N}{2}$ and fails with very high probability when $2L > W$. The choice of $\{L, W\} = \{\frac{N}{4}, \frac{N}{2}\}$ uses only *six* short-time sections and STliFT recovers the underlying signal with very high probability, which, given the limited success of semidefinite relaxation based algorithms in the Fourier phase retrieval setup, is very encouraging. Also, the superior recovery ability of STliFT when compared to the alternating projection based GL algorithm can be clearly seen.

In the second set of simulations, we evaluate the performance of STliFT as a function of shift length and measurements per short-time section. We choose $N = 32$ and $W = 16$, and vary $\{L, M\}$. For each choice of $\{L, M\}$, we perform 100 trials as before. The probability of successful recovery as a function of $\{L, M\}$ is plotted in Fig. 5.5. Observe that recovery is successful in the $4L \leq M < 2W$ regime, which demonstrates super-resolution.

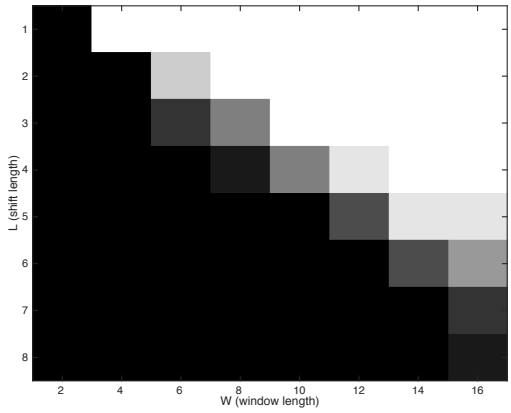
In the third set of simulations, we evaluate the performance of STliFT and GL algorithm in the noisy setting. We choose $M = 2W$, the rest of the parameters are the same as the first set of simulations. The normalized mean-squared error, given by

$$NMSE = \min_{|c|=1} \frac{\|\mathbf{x}_0 - c\hat{\mathbf{x}}\|^2}{\|\mathbf{x}_0\|^2}, \tag{5.7}$$

is plotted as a function of SNR in Fig. 5.6a and 5.6b. The linear relationship between them in Fig. 5.6a shows that STliFT stably recovers the underlying signal in the presence of noise. Further, the plots demonstrate the superior reconstruction ability of STliFT when compared to GL algorithm. It can also be observed that the choices of $\{W, L\}$ which correspond to significant overlap between adjacent short-



(a) STliFT.



(b) GL algorithm.

Figure 5.4: Probability of successful recovery, for $N = 32$, $M = 4L$, and various choices of $\{L, W\}$, in the noiseless setting (white region: success with probability 1, black region: success with probability 0).

time sections tend to recover signals more stably compared to values of $\{W, L\}$ which correspond to less overlap, which is not surprising.

5.8 Conclusions and Future Work

In this chapter, we considered the STFT phase retrieval problem. We showed that, if $L < W \leq \frac{N}{2}$, then almost all non-vanishing signals can be uniquely identified from their STFT magnitude up to a global phase, and extended this result to incorporate sparse signals which have less than $\min\{W - L, L\}$ consecutive zeros.

For $2L \leq W \leq \frac{N}{2}$, we conjectured that most non-vanishing signals can be recovered

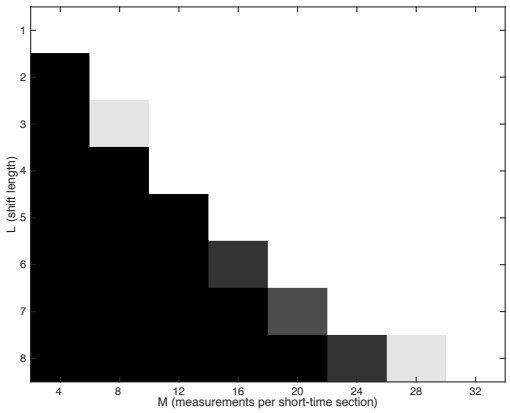
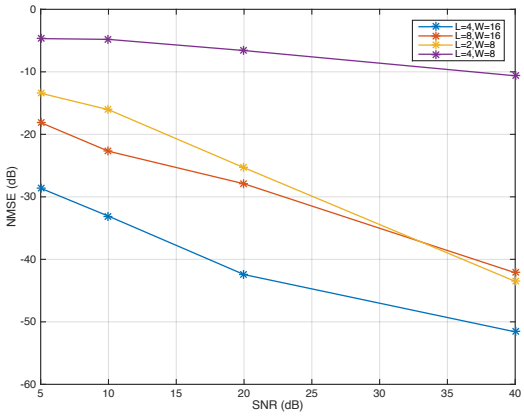


Figure 5.5: Probability of successful recovery using STliFT for $N = 32$, $W = 16$, and various choices of $\{L, M\}$ (white region: success with probability 1, black region: success with probability 0).

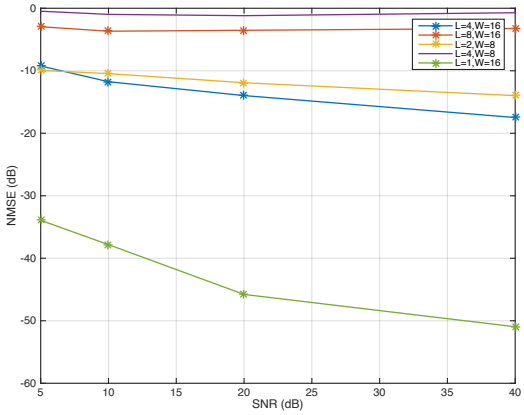
up to a global phase by a semidefinite relaxation based algorithm (STliFT). When $W = o(N)$, through super-resolution, we reduced the number of phaseless measurements to $(4 + o(1))N$. We proved this conjecture for the setup in which the first $\left\lfloor \frac{L}{2} + 1 \right\rfloor$ samples are known, and for the case in which $L = 1$. We argued that the additional assumptions are asymptotically reasonable when $W \ll N$, which is typically the case in practical methods. We then extended these results to incorporate sparse signals which have at most $W - 2L$ consecutive zeros.

Directions for future study include a proof of this conjecture without any additional assumptions, and a stability analysis in the noisy setting.

Also, a thorough analysis of the phase transition at $2L = W$ will provide a more complete characterization of STliFT. In particular, showing why the recovery fails when the overlap is less than 50% is an interesting direction. Observe that when $2L = W$, the total number of phaseless measurements considered is approximately $4N$, a quantity which comes up in many phase retrieval problems. Hence, an analysis of this phase transition would improve our fundamental understanding of recovery from phaseless measurements in general.



(a) STliFT.



(b) GL algorithm.

Figure 5.6: NMSE (dB) vs SNR (dB) in the noisy setting for $N = 32$, $M = 2W$.

Chapter 6

PHASELESS SUPER-RESOLUTION

Before introducing phaseless super-resolution, we briefly discuss super-resolution. The problem of recovering a signal from its low-frequency Fourier measurements is referred to as super-resolution. It is a fundamental problem in signal processing, and comes up in applications where it is difficult to obtain high-frequency measurements due to physical limitations. For example, in optical systems, there is a fundamental resolution limit due to diffraction [BW00]. In radar systems, there is a limit due to the size of the detectors [OBP94].

Due to the absence of high-frequency information, super-resolution is an ill-posed problem. There is a fundamental limit, called the Rayleigh criterion [Ray79], on the minimum separation between two locations with nonzero values, in order to be able to stably recover the signal. A pictorial representation of this criterion is shown in Fig. 6.1. In other words, signals which have two locations with nonzero values closer than the Rayleigh criterion cannot be resolved stably. If M low-frequencies from the N point DFT are available as measurements, then this limit turns out to be $\frac{N}{M}$ [Don92].

On the algorithmic front, the classic super-resolution algorithms include MUSIC [Sch86] and ESPRIT [RK89]. More recently, a convex programming based algorithm has been proposed in [CF14; Tan+13]. In particular, the authors show that signals with minimum separation at least $4 \times \frac{N}{M}$ can be stably reconstructed through ℓ_1 minimization (total variation minimization for the continuous model).

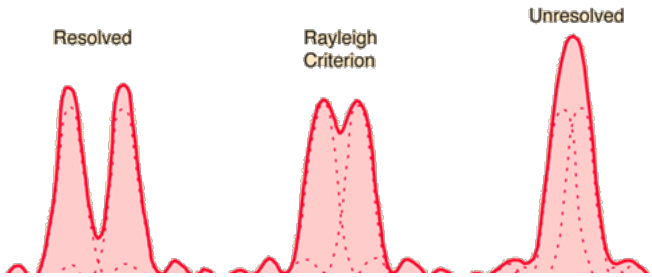


Figure 6.1: Rayleigh criterion: If two locations with nonzero values are located such that the first zero of one sinc coincides with the maximum of the other, then the signal is barely resolvable (courtesy of [Hyp]).

In this chapter, we consider phaseless super-resolution, which is the problem of recovering a signal from its low-frequency Fourier magnitude-square measurements. Mathematically, it is the following recovery problem:

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && Z[m] = \left|\langle \mathbf{f}_m, \mathbf{x}\rangle\right|^2 \quad \text{for} \quad 0 \le m \le M-1, \end{aligned} \tag{6.1}$$

where $\mathbf{f}_m$ is the conjugate of the mth column of the N point DFT matrix $\mathbf{F}$, and $\mathbf{Z} = (Z[0], Z[1], \ldots, Z[M-1])^T$ is the $M \times 1$ vector corresponding to the low-frequency magnitude-square measurements.

Phaseless super-resolution is the combination of two classic signal processing problems: phase retrieval and super-resolution. Consequently, in order to solve phaseless super-resolution, it is necessary to overcome the uniqueness issues of phase retrieval and super-resolution. To this end, we consider the use of additional magnitude-only measurements and impose a minimum separation prior on the signal.

The minimum separation of $\mathbf{x}_0$, denoted by $\Delta(\mathbf{x}_0)$, is defined as the closest distance between any two nonzero entries in $\mathbf{x}_0$, i.e,

$$\Delta(\mathbf{x}_0) = \min_{n \neq m, x_0[n] \neq 0, x_0[m] \neq 0} (n-m) \bmod N. \tag{6.2}$$

Here, the distance is defined in a cyclic manner. For example, if $N = 100$, then the distance between $n = 90$ and $m = 10$ is 20.

The masks proposed in Chapter 4 do not work in the super-resolution setup as they require the knowledge of the autocorrelation of the signal, which is available only when the entire Fourier magnitude-square information is available. In this chapter, we focus on "structured illuminations", a technique where the illuminating beam is made to hit the sample at specific angles [Far+10]. A schematic representation is provided in Fig. 6.2.

Keeping in mind practical applications (described in the next section), for $-R \le r \le R$, let $\mathbf{D}_r$ be an $N \times N$ diagonal matrix, corresponding to the rth structured illumination, with diagonal entries given by $(d_r[0], d_r[1], \ldots, d_r[N-1])$, such that

$$d_r[n] = \begin{cases} 1 & \text{if} \quad r = 0 \\ 1 + e^{i2\pi \frac{nr}{N}} & \text{if} \quad 1 \le r \le R \\ 1 - ie^{i2\pi \frac{n|r|}{N}} & \text{otherwise.} \end{cases} \tag{6.3}$$

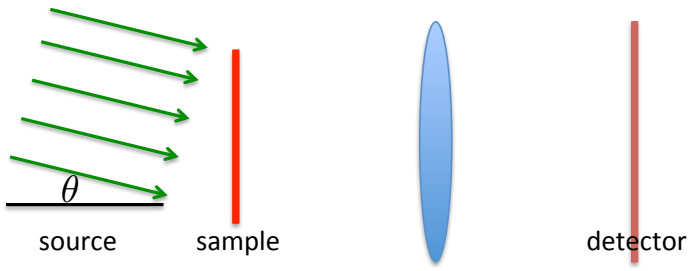


Figure 6.2: A schematic representation of structured illuminations in an optical setting.

Observe that the diagonal entries of $\mathbf{D}_0$ are equivalent to $\mathbf{f}_0$. Further, when $r > 0$ and $r < 0$, the diagonal entries of $\mathbf{D}_r$ are equivalent to $\mathbf{f}_0 + \mathbf{f}_r$ and $\mathbf{f}_0 - i\mathbf{f}_{|r|}$ respectively. Our objective is to solve the following reconstruction problem:

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} && Z[m,r] = \left|\langle \mathbf{f}_m, \mathbf{D}_r \mathbf{x}\rangle\right|^2 \\ &\quad \text{for} && 0 \le m \le M-1 \quad \text{and} \quad -R \le r \le R, \end{aligned} \tag{6.4}$$

where $\mathbf{Z}$ is an $M \times (2R+1)$ matrix, such that $Z[m,r]$ corresponds to the magnitude-square of the mth low-frequency measurement using the rth structured illumination.

Notation: $\mathbf{F}_M$ denotes the $M \times N$ submatrix of $\mathbf{F}$ constructed using the first M rows. The signal of interest is represented by $\mathbf{x}_0$ as before, and the vector $\mathbf{y}_0 = (y_0[0], y_0[1], \ldots, y_0[N-1])^T$ represents the N point Fourier transform of $\mathbf{x}_0$.

In what follows, we describe the strategy for obtaining such measurements in two applications. Then, we present our results.

6.1 X-ray Crystallography/ Coherent Diffraction Imaging

Let $\psi(x,y)$ denote the two-dimensional object, centered at the origin. Also, let the plane of the object and the two-dimensional detector be perpendicular to the z-axis such that $z = 0$ and $z = z'$ respectively. In a standard diffraction imaging setup, the entire object is illuminated using a single source placed at $(0,0,-d)$, and the diffraction patterns are recorded. Alternately, consider the setup where the object is illuminated using two sources, one placed at $(0,0,-d)$ and the other placed at $(x_r, y_r, -d)$.

Suppose the two sources are coherent. By using Huygens principle (same arguments as in Section 1.1), if $\psi_{trans,r}(x,y)$ denotes the secondary source at $(x,y,0)$

produced by $\psi(x, y)$, then we have $\psi_{trans,r}(x, y) = \psi_{trans,1}(x, y) + \psi_{trans,2}(x, y)$, where $\psi_{trans,1}(x, y)$ and $\psi_{trans,2}(x, y)$ are the secondary sources at $(x, y, 0)$ produced by $\psi(x, y)$ due to the two sources individually. Using the calculations from Section 5.1, we have $\psi_{trans,1}(x, y) = \psi(x, y)$ and $\psi_{trans,2}(x, y) = e^{i\frac{2\pi}{\lambda}\frac{x_r x + y_r y}{d}}\psi(x, y)$.

Consequently, the diffraction pattern intensities measured are proportional to the magnitude-square of the Fourier transform of

$$\psi_{trans,r}(x, y) = \left(1 + e^{i\frac{2\pi}{\lambda}\frac{x_r x + y_r y}{d}}\right)\psi(x, y). \tag{6.5}$$

In other words, the measurements are such that

$$\begin{aligned} I_r(x', y') &\propto \left|\iint \psi_{trans,r}(x, y) e^{i\frac{2\pi}{\lambda}\frac{-x'x - y'y}{z'}} dxdy\right|^2 \\ &\propto \left|\iint \left(1 + e^{i\frac{2\pi}{\lambda}\frac{x_r x + y_r y}{d}}\right)\psi(x, y) e^{i\frac{2\pi}{\lambda}\frac{-x'x - y'y}{z'}} dxdy\right|^2 \\ &\propto \left|\hat{\psi}\left(\frac{x'}{\lambda z'}, \frac{y'}{\lambda z'}\right) + \hat{\psi}\left(\frac{x'}{\lambda z'} - \frac{x_r}{\lambda d}, \frac{y'}{\lambda z'} - \frac{y_r}{\lambda d}\right)\right|^2. \end{aligned}$$

For instance, in the one-dimensional setup, if the distance between adjacent detectors is given by δ and $x_r = r\delta\frac{d}{z'}$, then the diagonal entries of $\mathbf{D}_r$ are $\mathbf{f}_0 + \mathbf{f}_r$.

If the two sources are ninety-degrees out of phase, then we have

$$\psi_{trans,r}(x, y) = \left(1 - ie^{i\frac{2\pi}{\lambda}\frac{x_r x + y_r y}{d}}\right)\psi(x, y), \tag{6.6}$$

and the corresponding diagonal entries of $\mathbf{D}_r$ are $\mathbf{f}_0 - i\mathbf{f}_r$.

6.2 Direction of Arrival Estimation

In Section 1.4, we considered the classic direction of arrival estimation setup, which involves one transmitter and M receivers placed uniformly. Alternately, consider the setup where two transmitters, one placed at the origin and the other placed at $x = -\frac{\lambda}{2}$, are used (see Fig. 6.3).

Suppose the two sources are coherent. The received vector is the sum of the received signals due to the two transmitters individually, i.e.,

$$\begin{aligned} y^{(\omega)}[m] &= \sum_{k=1}^{K} \hat{s}(\omega - \omega_c) e^{-i\omega_c \frac{2r_k - \frac{m\lambda}{2}\sin\theta_k}{c}} \rho_k + \sum_{k=1}^{K} \hat{s}(\omega - \omega_c) e^{-i\omega_c \frac{2r_k + \frac{\lambda}{2}\sin\theta_k - \frac{m\lambda}{2}\sin\theta_k}{c}} \rho_k \\ &\propto \sum_{k=1}^{K} (1 + e^{-i\pi\sin\theta_k}) e^{i\pi m \sin\theta_k} \left(\rho_k e^{-\frac{i2\omega_c r_k}{c}}\right), \end{aligned} \tag{6.7}$$

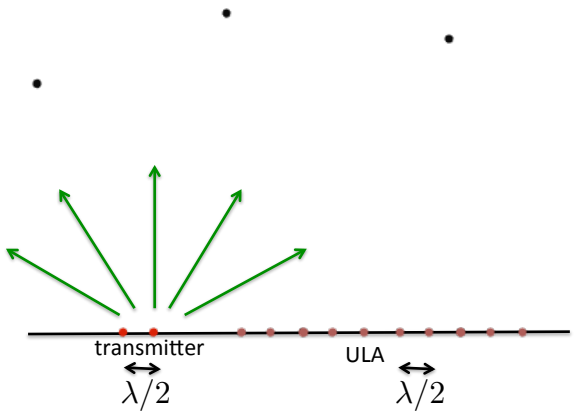


Figure 6.3: Implementation of the proposed additional magnitude-only measurements in the Direction of Arrival Estimation setup (ULA=Uniform Linear Array).

which is precisely the measurements due to the matrix with diagonal entries $\mathbf{f}_0 + \mathbf{f}_{-1}$. Here, we used the fact that the total distance travelled by the wave transmitted by the source at $x = -\frac{\lambda}{2}$, reflected by object k onto receiver m is well-approximated by $2r_k + \frac{\lambda}{2}\sin\theta_k - \frac{m\lambda}{2}\sin\theta_k$. In order to change the diagonal entries to $\mathbf{f}_0 + \mathbf{f}_r$, the second transmitter has to be placed at $x = \frac{r\lambda}{2}$ instead. Also, if the two sources are ninety-degrees out of phase, then the diagonal entries become $\mathbf{f}_0 - i\mathbf{f}_r$.

Remark: When the M receivers are placed uniformly, the number of objects that can be detected is well-known to be $O(M)$. Recently, in [PV11; VP11], the authors proposed "coprime arrays", where two uniform arrays with M_1 and M_2 receivers and spacings $\frac{M_2\lambda}{2}$ and $\frac{M_1\lambda}{2}$ respectively are considered. The number of objects that can be detected is shown to be $O(M_1M_2)$. Suppose three transmitters, one placed at the origin and the other two placed at $x = -\frac{M_1\lambda}{2}$ and $x = -\frac{M_2\lambda}{2}$, are used. If the first and the third transmitters are on while the second is off, then the setting is similar to two transmitters and M_1 uniformly placed receivers. Similarly, if the first and the second transmitters are on while the third is off, then the setting is similar to two transmitters and M_2 uniformly placed receivers. Consequently, using three transmitters instead of two, the theory in this section can be applied to coprime arrays. The same idea also applies to "nested arrays", which are concatenations of two uniform arrays [PV10].

6.3 Contributions

In this chapter, we show that the SDP algorithm can provably recover most signals when measurements are obtained using 3 simple structured illuminations (i.e., $R = 1$

is sufficient). As shown in the previous section, these structured illuminations can be implemented by using just two sources in several applications. We also show that the reconstruction is stable in the noisy setting. Indeed, the stability parameter depends on the number of sources and structured illuminations considered.

6.4 Methodology

Two-Stage Combinatorial Reconstruction

In the noiseless setting, phase can be uniquely resolved up to a global factor if $R \geq 1$. In [Can+15], the following arguments are provided: The measurements obtained provide the knowledge of $|y_0[n]|^2$, $|y_0[n] + y_0[n-1]|^2$ and $|y_0[n] - iy_0[n-1]|^2$ for $0 \leq n \leq M-1$. Writing $y_0[n] = |y_0[n]|\, e^{i\phi_0[n]}$ for $0 \leq n \leq M-1$, we have

$$|y_0[n] + y_0[n-1]|^2 = |y_0[n]|^2 + |y_0[n-1]|^2 + 2|y_0[n]||y_0[n-1]|\,\mathrm{Re}(e^{i(\phi_0[n-1]-\phi_0[n])}),$$

$$|y_0[n] - iy_0[n-1]|^2 = |y_0[n]|^2 + |y_0[n-1]|^2 + 2|y_0[n]||y_0[n-1]|\,\mathrm{Im}(e^{i(\phi_0[n-1]-\phi_0[n])}).$$

Consequently, if $y_0[n] \neq 0$ for $0 \leq n \leq M-1$, then the measurements provide the relative phases $\phi_0[n-1] - \phi_0[n]$ for $0 \leq n \leq M-1$. By setting $\phi_0[0] = 0$ without loss of generality, $\phi_0[n]$ can be inferred for $1 \leq n \leq M-1$.

Once phase is resolved, phaseless super-resolution reduces to reconstructing $\mathbf{x}_0$ from $e^{i\phi}\mathbf{F}_M\mathbf{x}_0$, which is the super-resolution problem. Hence, by using such measurements, any super-resolution algorithm can be extended to solve phaseless super-resolution if $R \geq 1$. It is well-known that a k-sparse signal $\mathbf{x}_0$ can be reconstructed from $\mathbf{F}_M\mathbf{x}_0$ if $M \geq 2k+1$ (e.g., matrix pencil method [HS90]). Consequently, k-sparse signals can be reconstructed from $(2k+1) \times 3 = 6k+3$ low-frequency Fourier magnitude measurements.

However, such an approach cannot be used in the noisy setting due to error propagation issues. In the next subsection, we consider the SDP algorithm.

Semidefinite Relaxation

The matrix $\mathbf{X}_0 = \mathbf{x}_0\mathbf{x}_0^\star$ we are interested in recovering is both sparse and low-rank. The most natural objective function to recover such a matrix is a linear combination of the ℓ_1 norm and the nuclear norm (which is equal to the *trace* norm for positive semidefinite matrices). Since the measurement corresponding to $r = 0$ and $m = 0$ fixes $\mathrm{trace}(\mathbf{X})$, we consider the following convex algorithm:

Algorithm 7 SDP based Phaseless Super-resolution

Input: Masked low-frequency Fourier magnitude measurements $\mathbf{Z}$
Output: Sparse signal $\hat{\mathbf{x}}$ satisfying the measurements

- Solve for $\hat{\mathbf{X}}$:

$$\begin{aligned} &\text{minimize} && \|\mathbf{X}\|_1 \\ &\text{subject to} && Z[m,r] = \text{trace}(\, \mathbf{D}_r^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{D}_r \mathbf{X} \,) \\ &\quad \text{for} \quad 0 \le m \le M-1 \quad \text{and} && \quad -R \le r \le R, \\ &&& \mathbf{X} \succcurlyeq 0. \end{aligned} \tag{6.8}$$

- Return $\hat{\mathbf{x}}$, where $\hat{\mathbf{x}}$ is the best rank-one approximation of $\hat{\mathbf{X}}$.

Theorem 6.4.1. *The matrix $\mathbf{X}_0 = \mathbf{x}_0 \mathbf{x}_0^\star$ is the unique optimizer of (6.8), and therefore $\mathbf{x}_0$ can be uniquely reconstructed up to a global phase if*

(i) $R \ge 1$

(ii) $\Delta(\mathbf{x}_0) \ge \frac{4.76N}{M}$

(iii) $y_0[0], y_0[1], \ldots, y_0[M-1] \neq 0$.

Proof. Analyzing convex programs with deterministic measurement vectors is a difficult task in general. In order to analyze (6.8), inspired by [BR15], we use the following approach: We first show that the affine constraints in (6.8) are of the form $\text{trace}(\mathbf{c}_{r,m}\mathbf{c}_{r,m}^T \mathbf{W})$, where $\mathbf{W} = \mathbf{F}_M \mathbf{X} \mathbf{F}_M^\star$ and $\mathbf{c}_{r,m}$ is an $M \times 1$ *sensing vector* corresponding to each measurement. For the matrices defined in (6.3), we then calculate the sensing vectors and show that they, along with the positive semidefinite condition, uniquely determine $\mathbf{W}$. We finally use the results of [CF14; Tan+13] to show that $\mathbf{X}_0$ is the unique optimizer of (6.8).

We now show the first step of the aforementioned approach. For each r, let $(s_r[0], s_r[1], \ldots, s_r[N-1])^T$ denote the N point DFT of $(d_r[0], d_r[1], \ldots, d_r[N-1])^T$ and for each l, let $\text{Diag}(\mathbf{f}_l)$ be an $N \times N$ diagonal matrix with diagonal entries $\mathbf{f}_l$. We have

$$\mathbf{f}_m^\star \mathbf{D}_r = \left(\sum_{l=0}^{N-1} s_r[l]\, \mathbf{f}_m^\star \text{Diag}(\mathbf{f}_l) \right).$$

Since $\mathbf{f}_m^\star \mathrm{Diag}(\mathbf{f}_l) = \mathbf{f}_{m-l}^\star$, for every m and r, the aforementioned expression can be rewritten as

$$\mathbf{f}_m^\star \mathbf{D}_r = \sum_{l=0}^{N-1} s_r[l] \mathbf{f}_{m-l}^\star = \sum_{l=0}^{N-1} s_r[l] \mathbf{e}_{m-l}^\star \mathbf{F},$$

where $\mathbf{e}_m$ is the mth column of the identity matrix.

The matrices defined in (6.3) satisfy $s_r[l] = 0$ for every $l > r$. Consequently, in the regime $r \leq m \leq M-1$, we have $0 \leq m-l \leq M-1$ when $s_r[l] \neq 0$. Therefore, $s_r[l]\mathbf{e}_{m-l}^\star \mathbf{F}$ is equivalent to $s_r[l]\mathbf{e}_{m-l}^\star \mathbf{F}_M$, due to which we can rewrite $\mathrm{trace}(\mathbf{D}_r^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{D}_r \mathbf{X})$ as:

$$\mathrm{trace}\left(\left(\sum_{l=0}^{M-1} s_r^\star[l] \mathbf{e}_{m-l} \right) \left(\sum_{l=0}^{M-1} s_r[l] \mathbf{e}_{m-l}^\star \right) \mathbf{F}_M \mathbf{X} \mathbf{F}_M^\star \right). \tag{6.9}$$

We now calculate the sensing vectors when measurements are obtained using the matrices defined in (6.3). The values of $s_r[l]$ are:

$$s_0[l] = \begin{cases} 1 & \text{for} \quad l = 0 \\ 0 & \text{otherwise} \end{cases} \tag{6.10}$$

$$s_r[l] = \begin{cases} 1 & \text{for} \quad l = 0 \\ -i & \text{for} \quad l = -r \quad \text{if} \quad r < 0 \\ 1 & \text{for} \quad l = r \quad \text{if} \quad r > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Substituting these values in (6.9), the measurement corresponding to $r = 0$ and any $0 \leq m \leq M-1$ fixes $W[m,m]$. Similarly, the measurements corresponding to $r = \pm 1$ and any $1 \leq m \leq M-1$ fix the values of $W[m-1,m-1] + W[m-1,m] + W[m,m-1] + W[m,m]$ and $W[m-1,m-1] + iW[m-1,m] - iW[m,m-1] + W[m,m]$. These measurements, combined with the measurements corresponding to $r = 0$, fix $W[m-1,m]$ and $W[m,m-1]$.

Hence, the diagonal and the first off-diagonal entries of every feasible $\mathbf{W}$ match the diagonal and the first off-diagonal entries of the matrix $(\mathbf{F}_M \mathbf{x}_0)(\mathbf{F}_M \mathbf{x}_0)^\star$. Since the entries are sampled from a rank one matrix with non-zero diagonal entries (the first M values of the N point DFT of $\mathbf{x}_0$ are non-zero), there is exactly one positive semidefinite completion, which is the rank one completion $(\mathbf{F}_M \mathbf{x}_0)(\mathbf{F}_M \mathbf{x}_0)^\star$.

In particular, due to the fact that $\mathbf{W} \succcurlyeq 0$ and $\mathbf{F}_M \mathbf{x}_0$ is non-vanishing, $(\mathbf{F}_M \mathbf{x}_0)(\mathbf{F}_M \mathbf{x}_0)^\star$ is the only feasible $\mathbf{W}$. The reconstruction problem is therefore reduced to

$$\begin{aligned} &\text{minimize} && \|\mathbf{X}\|_1 \\ &\text{subject to} && \mathbf{F}_M \mathbf{X} \mathbf{F}_M^\star = \mathbf{F}_M \mathbf{x}_0 \mathbf{x}_0^\star \mathbf{F}_M^\star \\ & && \mathbf{X} \succcurlyeq 0. \end{aligned} \tag{6.11}$$

This is precisely the two-dimensional super-resolution problem. Since the conditions of Theorem 1.3 in [CF14] are satisfied by $\mathbf{x}_0 \mathbf{x}_0^\star$, $\mathbf{x}_0 \mathbf{x}_0^\star$ is the unique optimizer of (6.11). □

6.5 Stability

We now consider the impact of measurement noise on the performance of the proposed SDP algorithm. Specifically, we consider measurements of the form

$$Z[m, r] = \left|\langle \mathbf{f}_m, \mathbf{D}_r \mathbf{x} \rangle\right|^2 + z[m, r] \tag{6.12}$$

for $0 \le m \le M-1$ and $-R \le r \le R$, where $\mathbf{z}_r = (z[0, r], z[1, r], \ldots, z[M-1, r])^T$ is the additive noise corresponding to the measurements from the rth structured illumination. The SDP algorithm, in the noisy setting, can be implemented as follows: Suppose the ℓ_1 norm of the noise vector is bounded by η, the affine constraints in the convex program can be replaced by

$$\sum_{r=-R}^{R} \sum_{m=0}^{M-1} \left| Z[m, r] - \text{trace}\left(\mathbf{D}_r^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{D}_r \mathbf{X} \right) \right| \le \eta.$$

In this setting, we prove the following theorem:

Theorem 6.5.1. *The optimizer $\hat{\mathbf{X}}$ of (6.8), in the noisy setting, satisfies*

$$\|\hat{\mathbf{X}} - \mathbf{X}_0\|_1 \le C \frac{\|\mathbf{w}_0\|_2^2 \gamma_{max}^2}{\gamma_{min}^3} R M^4 S R F^4 \eta$$

for some positive constant C, where $\gamma_{max} = \max\{|y_0[0]|^2, |y_0[1]|^2, \ldots, |y_0[M-1]|^2\}$ and $\gamma_{min} = \min_{0 \le m < M-R} \max\{|y_0[m]|^2, \ldots, |y_0[m+R-1]|^2\}$, if

(i) $R \ge 1$

(ii) $\Delta(\mathbf{x}_0) \ge \frac{4.76N}{M}$

(iii) $\gamma_{min} > 0$.

Proof. See Appendix 10.1. □

6.6 Extension to $2D$

Consider the one-dimensional example with $R = 1$, where two transmitters, one placed at 0 and the other placed at x_1, are used. We showed that three structured illuminations from these transmitters are sufficient for the SDP method to work: The first transmission involves just the transmitter at 0, and the second and the third involve coherent and ninety-degrees out of phase transmissions from the two transmitters respectively.

The theory developed in this chapter can be extended to incorporate two-dimensional signals as follows: Three transmitters are used, where one is placed at $(0, 0)$ and the other two are placed at $(x_1, 0)$ and $(0, x_1)$. A total of five structured illuminations are used: The first transmission involves just the transmitter at $(0, 0)$, the second and the third involve coherent and ninety-degrees out of phase transmissions from the first and the second transmitter respectively, and the fourth and the fifth involve coherent and ninety-degrees out of phase transmissions from the first and the third transmitter respectively. The correctness of this method can be seen by using a vectorizing operator on $\mathbf{F}_M \mathbf{x} \mathbf{F}_M^\star$, similar to the arguments in Sections 3.5 and 4.5.

6.7 Numerical Simulations

In this section, we demonstrate the performance of the proposed method using numerical simulations.

In the first set of simulations, we choose $N = 32$, $R = 1$, and evaluate the performance for various choices of minimum-spacing $\Delta(\mathbf{x}_0)$ and number of low-frequency measurements M. For each choice of $\Delta(\mathbf{x}_0)$ and M, we perform 25 trials using the parser YALMIP and the solver SeDuMi. For each trial, a complex signal is randomly generated as follows: Starting from an empty support, 100 indices in the range 0 and $N - 1$ are generated uniformly at random (with repetition) and sequentially added as long as the minimum-spacing criterion is not violated. The signal values in the support are drawn from a standard complex Gaussian distribution independently. The probability of successful recovery is plotted in Fig. 6.4. The black region corresponds to a success probability of 0 and the white region corresponds to a success probability of 1. If $M \gtrsim \frac{2N}{\Delta(\mathbf{x}_0)}$, then the SDP algorithm recovers the signal with very high probability.

In the second set of simulations, we evaluate the performance of the SDP algorithm in the noisy setting. We choose $N = 32$, $M = 10$, $R = 1, 2, 3, 4$ and $\Delta(\mathbf{x}_0) = 8$. The lifted signal $\mathbf{X}_0$ is normalized so that $\|\mathbf{X}_0\|_1 = 1$, and the value of $\|\mathbf{X}_0 - \hat{\mathbf{X}}\|_1$ is

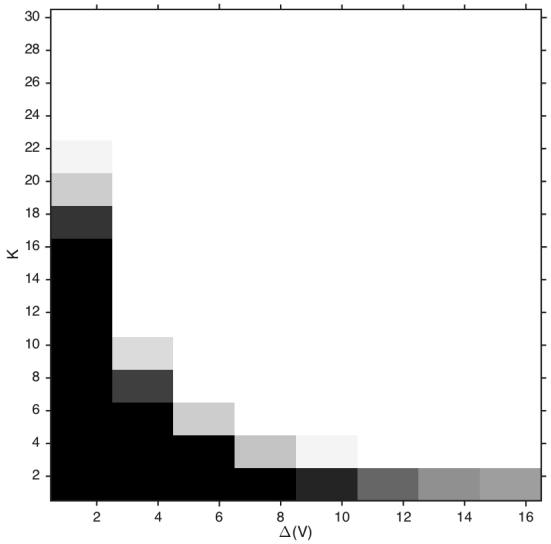


Figure 6.4: Probability of successful recovery for $N = 32$, $R = 1$, and various choices of M and $\Delta(\mathbf{x}_0)$.

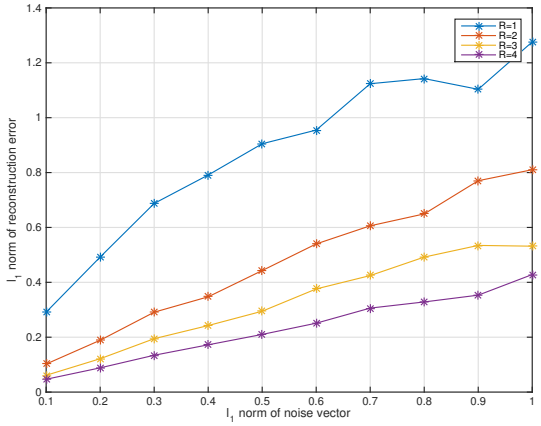


Figure 6.5: Stability of the SDP algorithm in the noisy setting for $N = 32$, $M = 10$, $\Delta(\mathbf{x}_0) = 8$, and various choices of R.

plotted as a function of the ℓ_1 norm of the noise vector. The results of the simulations are shown in Fig. 6.5. The plots demonstrate that the reconstruction is stable in the noisy setting.

6.8 Conclusions and Future Work

We considered the problem of phaseless super-resolution. We showed that any super-resolution algorithm can be extended to solve phaseless super-resolution, when certain additional magnitude-only measurements are available. We also described

the practical implementation of such measurements in various applications.

We then focused our attention on the SDP algorithm, and provided theoretical guarantees. In particular, we extended the super-resolution results of [CF14] to incorporate phaseless super-resolution.

We have identified the following set of questions as potential directions for future study:

- Numerical simulations suggest that the stability analysis provided in this chapter is not tight. Given the practicality of the proposed designs, it would be very useful to have tight bounds on the reconstruction error.
- Suppose the structured illuminations are implemented using exactly T transmitters. Where should the transmitters be placed, and what structured illuminations should be used, so that the SDP method provably works, and is most stable? The answer to this question would be very helpful from a practical point of view in applications like optics, diffraction imaging, crystallography and radar. While a tight stability analysis would help answer this question, a loose stability analysis which mirrors the simulations in terms of trends would also be very useful.

Chapter 7

CONCLUDING REMARKS AND FUTURE DIRECTIONS

In this work, we studied several variants of the phase retrieval problem using convex optimization theory.

With focus on applications like astronomy and crystallography, we studied the sparse phase retrieval problem in Chapter 3. We first showed that most $(N-1)$-sparse signals can be uniquely recovered from oversampled measurements. Then, we developed the TSPR algorithm and provided theoretical guarantees. In particular, we showed that TSPR can provably recover most $O(N^{\frac{1}{2}-\epsilon})$-sparse signals, and stably reconstruct most $O(N^{\frac{1}{4}-\epsilon})$-sparse signals.

In Chapter 4, we considered the problem of phase retrieval using masks, with focus on applications like crystallography, diffraction imaging and optics. We showed that oversampled measurements using two specific masks, or three simple-to-implement masks, are sufficient for the SDP method to provably recover most signals. If oversampled measurements are unavailable, then we argued that five specific masks, or seven simple-to-implement masks, are sufficient.

Keeping in mind applications like ptychography and Fourier ptychography, we studied the STFT phase retrieval problem in Chapter 5. We showed that almost all signals can be uniquely identified if $L < W \leq \frac{N}{2}$ (i.e., adjacent short-time sections overlap). We also showed that the SDP method provably recovers most signals if $2L \leq W \leq \frac{N}{2}$ (i.e., adjacent short-time sections overlap by 50%) under asymptotically reasonable assumptions.

In Chapter 6, we considered the problem of phaseless super-resolution, with focus on diffraction imaging, optics and radar applications. We showed that measurements using three structured illuminations are sufficient for the SDP method to provably recover most signals. We also described the implementation of these structured illuminations in various applications.

These results establish that convex optimization is a powerful tool to study phase retrieval type problems. Furthermore, the comparative numerical study of the SDP and the alternating projection based methods clearly demonstrated the superior abilities of the SDP based methods.

Below, we describe several natural directions for future research.

7.1 Precise Stability Analysis

In this work, we provided strong theoretical recovery guarantees for the SDP method in the noiseless setting, along with a weak stability analysis in many cases. Empirical evidence shows that the reconstruction is very stable in the noisy setting. Indeed, a natural direction for future study is precise stability analysis. Tight bounds on the reconstruction error would be very useful in several applications.

In the following, we describe the key quantities involved in the stability analysis. The semidefinite programs considered in this work can be written as

$$\begin{aligned} \text{minimize} \quad & \operatorname{trace}(\mathbf{X}) + \lambda \|\mathbf{X}\|_1 \\ \text{subject to} \quad & \mathcal{A}(\mathbf{X}) = \mathbf{b} \\ & \mathbf{X} \succcurlyeq 0. \end{aligned} \tag{7.1}$$

The method of dual certificates involved the construction of a *dual certificate* $\mathbf{W}$ with certain properties. Precise lower bounds on the following quantities can potentially result in a tight bound:

- Second smallest singular value of $\mathbf{W}$
- $\min_{\mathbf{h}} \frac{\|\mathcal{A}(\mathbf{x}_0\mathbf{h}^\star + \mathbf{h}\mathbf{x}_0^\star)\|}{\|\mathbf{x}_0\mathbf{h}^\star + \mathbf{h}\mathbf{x}_0^\star\|}$.

In a compressed sensing style analysis, concentration inequalities play a critical role in bounding these quantities. Indeed, the second quantity is reminiscent of the Restricted Isometry Property (RIP). However, when the measurements are deterministic, a different approach is required. We refer the interested readers to [JEH15d; Jag+16] for details.

7.2 Non-Convex Optimization

Recently, researchers have successfully analyzed the problem of signal recovery from random phaseless measurements using tools from non-convex optimization. In particular, measurements of the form

$$Z[m] = \left|\langle \mathbf{a}_m, \mathbf{x} \rangle\right|^2, \tag{7.2}$$

where $\mathbf{a}_m$ is sampled from a generic distribution, are assumed to be available. In [NJS13], theoretical guarantees were provided for the alternating projection based

algorithm. In [CLS15b], the authors provided theoretical guarantees for a stochastic gradient descent based algorithm (called the WF algorithm). A carefully chosen initialization and concentration inequalities are an integral part of these methods.

A natural question is whether one can extend these results to the setups considered in this work. Indeed, a different approach is required as the measurements are deterministic. In this regard, the dual certificates (4.12, 4.17, 9.8, 10.2) could potentially be very useful. If that is indeed the case, then the convex framework might provide valuable insights into our understanding of the alternating projection and the stochastic gradient descent methods. The framework could also help standardize the tools for analyzing such non-convex methods.

7.3 General Theory for QCQPs

At a very high level, the approach used in this work can be summarized as follows:

- Find an algebraic reconstruction method
- Construct a dual certificate using ideas from the algebraic method

In particular, the dual certificate (4.12) uses ideas from the algebraic method described in (4.10). The dual certificate (4.17) uses ideas from the Sylvester matrix based algebraic method (4.16). The algebraic methods described in [NQL83] and [Can+15] influenced the construction of the dual certificates (9.8) and (10.2) respectively.

The exact relationship between these algebraic reconstruction methods and dual certificates is unclear. We strongly suspect the existence of a fundamental quantity which connects these results. If that is the case, then identifying this quantity would improve our understanding of semidefinite relaxation and its role in solving general QCQPs.

BIBLIOGRAPHY

[260] 260H. "https://260h.pbworks.com/w/page/30814223/XRayCrystallography". In: ().

[Ale+14] Boris Alexeev et al. "Phase retrieval with polarization". In: *SIAM Journal on Imaging Sciences* 7.1 (2014), pp. 35–66.

[AS15] Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2015.

[B+13] Quentin Berthet, Philippe Rigollet, et al. "Optimal detection of sparse principal components in high dimension". In: *The Annals of Statistics* 41.4 (2013), pp. 1780–1815.

[Bal+09] Radu Balan et al. "Painless reconstruction from magnitudes of frame coefficients". In: *Journal of Fourier Analysis and Applications* 15.4 (2009), pp. 488–501.

[Bat82] RHT Bates. "Astronomical speckle imaging". In: *Physics Reports* 90.4 (1982), pp. 203–297.

[Bay04] Buyurman Baykal. "Blind channel estimation via combining autocorrelation and blind phase estimation". In: *Circuits and Systems I: Regular Papers, IEEE Transactions on* 51.6 (2004), pp. 1125–1131.

[BCE06] Radu Balan, Pete Casazza, and Dan Edidin. "On signal reconstruction without phase". In: *Applied and Computational Harmonic Analysis* 20.3 (2006), pp. 345–356.

[BCL02] Heinz H Bauschke, Patrick L Combettes, and D Russell Luke. "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization". In: *JOSA A* 19.7 (2002), pp. 1334–1345.

[BCM14] Afonso S Bandeira, Yutong Chen, and Dustin G Mixon. "Phase retrieval from power spectra of masked signals". In: *Information and Inference* (2014), iau002.

[BE15] Tamir Bendory and Yonina C Eldar. "A Least Squares Approach for Stable Phase Retrieval from Short-Time Fourier Transform Magnitude". In: *arXiv preprint arXiv:1510.00920* (2015).

[Ber99] Dimitri P Bertsekas. "Nonlinear programming". In: (1999).

[Bit+78] RR Bitmead et al. "Greatest common divisor via generalized Sylvester and Bezout matrices". In: *Automatic Control, IEEE Transactions on* 23.6 (1978), pp. 1043–1047.

[BR13] Quentin Berthet and Philippe Rigollet. "Complexity theoretic lower bounds for sparse principal component detection". In: *Conference on Learning Theory*. 2013, pp. 1046–1066.

[BR15] Sohail Bahmani and Justin Romberg. “Efficient compressive phase retrieval with constrained sensing vectors”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 523–531.

[Bra13] William Lawrence Bragg. “The structure of some crystals as indicated by their diffraction of X-rays”. In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 89. 610. The Royal Society. 1913, pp. 248–277.

[BW00] Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. CUP Archive, 2000.

[Can+15] Emmanuel J Candes et al. “Phase retrieval via matrix completion”. In: *SIAM Review* 57.2 (2015), pp. 225–251.

[CF14] Emmanuel J Candès and Carlos Fernandez-Granda. “Towards a Mathematical Theory of Super-resolution”. In: *Communications on Pure and Applied Mathematics* 67.6 (2014), pp. 906–956.

[Cha+12] Venkat Chandrasekaran et al. “The convex geometry of linear inverse problems”. In: *Foundations of Computational mathematics* 12.6 (2012), pp. 805–849.

[CLS15a] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. “Phase retrieval from coded diffraction patterns”. In: *Applied and Computational Harmonic Analysis* 39.2 (2015), pp. 277–299.

[CLS15b] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. “Phase retrieval via Wirtinger flow: Theory and algorithms”. In: *Information Theory, IEEE Transactions on* 61.4 (2015), pp. 1985–2007.

[CSV13] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming”. In: *Communications on Pure and Applied Mathematics* 66.8 (2013), pp. 1241–1274.

[CT05] Emmanuel J Candes and Terence Tao. “Decoding by linear programming”. In: *Information Theory, IEEE Transactions on* 51.12 (2005), pp. 4203–4215.

[CWB08] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. “Enhancing sparsity by reweighted l1 minimization”. In: *Journal of Fourier analysis and applications* 14.5-6 (2008), pp. 877–905.

[Dak00] Tamara Dakic. “On the turnpike problem”. PhD thesis. SIMON FRASER UNIVERSITY, 2000.

[DBD12] Philippe Dreesen, Kim Batselier, and Bart De Moor. “Back to the roots: Polynomial system solving, linear algebra, systems theory”. In: *Proc 16th IFAC Symposium on System Identification (SYSID 2012)*. 2012, pp. 1203–1208.

[Don92] David L Donoho. “Superresolution via sparsity constraints”. In: *SIAM Journal on Mathematical Analysis* 23.5 (1992), pp. 1309–1331.

[Dre07] Jan Drenth. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

[EK12] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[Eld+15] Yonina C Eldar et al. “Sparse phase retrieval from short-time Fourier measurements”. In: *Signal Processing Letters, IEEE* 22.5 (2015), pp. 638–642.

[EM14] Yonina C Eldar and Shahar Mendelson. “Phase retrieval: Stability and recovery guarantees”. In: *Applied and Computational Harmonic Analysis* 36.3 (2014), pp. 473–494.

[Fan12] Albert Fannjiang. “Absolute uniqueness of phase retrieval with random illumination”. In: *Inverse Problems* 28.7 (2012), p. 075008.

[Far+10] Ahmad Faridian et al. “Nanoscale imaging using deep ultraviolet digital holographic microscopy”. In: *Optics express* 18.13 (2010), pp. 14159–14164.

[FD87] C Fienup and J Dainty. “Phase retrieval and image reconstruction for astronomy”. In: *Image Recovery: Theory and Application* (1987), pp. 231–275.

[FHB03] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. “Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices”. In: *American Control Conference, 2003. Proceedings of the 2003*. Vol. 3. IEEE. 2003, pp. 2156–2162.

[Fie82] James R Fienup. “Phase retrieval algorithms: a comparison”. In: *Applied optics* 21.15 (1982), pp. 2758–2769.

[FR12] Benjamin Fine and Gerhard Rosenberger. *The fundamental theorem of algebra*. Springer Science & Business Media, 2012.

[Fra+12] Joachim Frank et al. *Computer processing of electron microscope images*. Vol. 13. Springer Science & Business Media, 2012.

[Fri66] David L Fried. “Optical resolution through a randomly inhomogeneous medium for very long and very short exposures”. In: *JOSA* 56.10 (1966), pp. 1372–1379.

[GKK15] David Gross, Felix Krahmer, and Richard Kueng. “Improved recovery guarantees for phase retrieval from coded diffraction patterns”. In: *Applied and Computational Harmonic Analysis* (2015).

[GL84] Daniel W Griffin and Jae S Lim. “Signal estimation from modified short-time Fourier transform”. In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 32.2 (1984), pp. 236–243.

[God97] Lal C Godara. “Application of antenna arrays to mobile communications: Beam-forming and direction-of-arrival considerations”. In: *Proceedings of the IEEE* 85.8 (1997), pp. 1195–1245.

[Goo05a] Joseph W Goodman. *Introduction to Fourier optics, Chapter 3*. Roberts and Company Publishers, 2005.

[Goo05b] Joseph W Goodman. *Introduction to Fourier optics, Chapter 5*. Roberts and Company Publishers, 2005.

[GS72] R W Gerchberg and W O Saxton. “A practical algorithm for the determination of the phase from image and diffraction plane pictures”. In: *Optik* 35 (1972), p. 237.

[GW95] Michel X Goemans and David P Williamson. “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. In: *Journal of the ACM (JACM)* 42.6 (1995), pp. 1115–1145.

[GZW88] Richard M Goldstein, Howard A Zebker, and Charles L Werner. “Satellite radar interferometry: Two-dimensional phase unwrapping”. In: *Radio science* 23.4 (1988), pp. 713–720.

[Har98] John W Hardy. *Adaptive optics for astronomical telescopes*. Oxford University Press on Demand, 1998.

[Hir+11] Michael Hirsch et al. “Online multi-frame blind deconvolution with super-resolution and saturation correction”. In: *Astronomy & Astrophysics* 531 (2011), A9.

[HJ12] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[HM82] Monson H Hayes and James H McClellan. “Reducible polynomials in more than one variable”. In: *Proceedings of the IEEE* 70.2 (1982), pp. 197–198.

[Hof64] Edward M Hofstetter. “Construction of time-limited functions with specified autocorrelation functions”. In: *Information Theory, IEEE Transactions on* 10.2 (1964), pp. 119–126.

[Hor+15] Roarke Horstmeyer et al. “Solving ptychography with a convex relaxation”. In: *New journal of physics* 17.5 (2015), p. 053044.

[HS70] Andras Hajnal and Endre Szemeredi. “Proof of a conjecture of Erdos”. In: *Combinatorial theory and its applications* 2 (1970), pp. 601–623.

[HS90] Yingbo Hua and Tapan K Sarkar. “Matrix pencil method for estimating parameters of exponentially damped undamped sinusoids in noise”. In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 38.5 (1990), pp. 814–824.

[Huy85] Christiaan Huygens. *Traité de la lumière: où sont expliquées les causes de ce qui luy arrive dans la reflexion, & dans la refraction, et particulierement dans l'etrange refraction du cistal d'Islande*. Chez Pierre vander Aa, 1885.

[HY14] Roarke Horstmeyer and Changhuei Yang. "A phase space model of Fourier ptychographic microscopy". In: *Optics express* 22.1 (2014), pp. 338–358.

[Hyp] HyperPhysics. "http://hyperphysics.phy-astr.gsu.edu/hbase/phyopt/raylei.html". In: ().

[Jag+16] Kishore Jaganathan et al. "Phaseless super-resolution using masks". In: *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE. 2016.

[JEH15a] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. "Phase Retrieval: An Overview of Recent Developments". In: *arXiv preprint arXiv:1510.07713* (2015).

[JEH15b] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. "Recovering signals from the Short-Time Fourier Transform magnitude". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE. 2015, pp. 3277–3281.

[JEH15c] Kishore Jaganathan, Yonina C Eldar, and Babak Hassibi. "STFT phase retrieval: Uniqueness guarantees and recovery algorithms". In: *arXiv preprint arXiv:1508.02820* (2015).

[JEH15d] Kishore Jaganathan, Yonina Eldar, and Babak Hassibi. "Phase retrieval with masks using convex optimization". In: *Information Theory (ISIT), 2015 IEEE International Symposium on*. IEEE. 2015, pp. 1655–1659.

[JH13] Kishore Jaganathan and Babak Hassibi. "Reconstruction of integers from pairwise distances". In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE. 2013, pp. 5974–5978.

[Joh+08] I Johnson et al. "Coherent diffractive imaging using phase front modifications". In: *Physical review letters* 100.15 (2008), p. 155503.

[JOH12a] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. "On robust phase retrieval for sparse signals". In: *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE. 2012, pp. 794–799.

[JOH12b] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. "Phase retrieval for sparse signals using rank minimization". In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE. 2012, pp. 3449–3452.

[JOH12c] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. “Recovery of sparse 1-D signals from the magnitudes of their Fourier transform”. In: *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium On*. IEEE. 2012, pp. 1473–1477.

[JOH13a] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. “Sparse phase retrieval: Convex algorithms and limitations”. In: *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE. 2013, pp. 1022–1026.

[JOH13b] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. “Sparse phase retrieval: Uniqueness guarantees and recovery algorithms”. In: *arXiv preprint arXiv:1311.2745* (2013).

[KEO16] Dani Kogan, Yonina C. Eldar, and Dan Oron. “On The 2D Phase Retrieval Problem”. In: (2016).

[Lab70] Antoine Labeyrie. “Attainment of diffraction limited resolution in large telescopes by Fourier analysing speckle patterns in star images”. In: *Astron. Astrophys* 6.1 (1970), pp. 85–87.

[Lov79] László Lovász. “On the Shannon capacity of a graph”. In: *Information Theory, IEEE Transactions on* 25.1 (1979), pp. 1–7.

[LP97] Erwin G Loewen and Evgeny Popov. *Diffraction gratings and applications*. CRC Press, 1997.

[LV11] Yue M Lu and Martin Vetterli. “Sparse spectral factorization: Unicity and reconstruction algorithms”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE. 2011, pp. 5976–5979.

[LV13] Xiaodong Li and Vladislav Voroninski. “Sparse signal recovery from quadratic measurements via convex programming”. In: *SIAM Journal on Mathematical Analysis* 45.5 (2013), pp. 3019–3033.

[LW88] Paul Lemke and Michael Werman. “On the Complexity of Inverting the Autocorrelation Function of a Finite Integer Sequence, and the Problem of Locating n Points on a Line, Given the Unlabelled Distances Between Them”. In: (1988).

[Mar07] Stefano Marchesini. “Invited article: A unified evaluation of iterative projection algorithms for phase retrieval”. In: *Review of scientific instruments* 78.1 (2007), p. 011301.

[Mil90] Rick P Millane. “Phase retrieval in crystallography and optics”. In: *JOSA A* 7.3 (1990), pp. 394–411.

[MS12] Subhadip Mukherjee and Chandra Sekhar Seelamantula. "An iterative algorithm for phase retrieval with sparsity constraints: application to frequency domain optical coherence tomography". In: *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE. 2012, pp. 553–556.

[Nas+14] Youssef SG Nashed et al. "Parallel ptychographic reconstruction". In: *Optics express* 22.26 (2014), pp. 32082–32097.

[NCK] NCKAS. "http://www.nckas.org/asterisms/home.htm". In: ().

[NJS13] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. "Phase retrieval using alternating minimization". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2796–2804.

[NQL83] S Hamid Nawab, Thomas F Quatieri, and Jae S Lim. "Signal reconstruction from short-time Fourier transform magnitude". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 31.4 (1983), pp. 986–998.

[OBP94] JW Odendaal, E Barnard, and CWI Pistorius. "Two-dimensional superresolution radar imaging using the MUSIC algorithm". In: *Antennas and Propagation, IEEE Transactions on* 42.10 (1994), pp. 1386–1391.

[OE14] Henrik Ohlsson and Yonina C Eldar. "On conditions for uniqueness in sparse phase retrieval". In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 1841–1845.

[OR70] James M Ortega and Werner C Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Vol. 30. Siam, 1970.

[Ou+13] Xiaoze Ou et al. "Quantitative phase imaging via Fourier ptychographic microscopy". In: *Optics letters* 38.22 (2013), pp. 4845–4848.

[Oym+15] Samet Oymak et al. "Simultaneously structured models with application to sparse and low-rank matrices". In: *Information Theory, IEEE Transactions on* 61.5 (2015), pp. 2886–2908.

[Pat34] Arthur Lindo Patterson. "A Fourier series method for the determination of the components of interatomic distances in crystals". In: *Physical Review* 46.5 (1934), p. 372.

[Pat44] A Lindo Patterson. "Ambiguities in the X-ray analysis of crystal structures". In: *Physical Review* 65.5-6 (1944), p. 195.

[PLR14] Ramtin Pedarsani, Kangwook Lee, and Kannan Ramchandran. "PhaseCode: Fast and efficient compressive phase retrieval based on sparse-graph codes". In: *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*. IEEE. 2014, pp. 842–849.

[PS82] Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1982.

[PV10] Piya Pal and PP Vaidyanathan. "Nested arrays: a novel approach to array processing with enhanced degrees of freedom". In: *Signal Processing, IEEE Transactions on* 58.8 (2010), pp. 4167–4181.

[PV11] Piya Pal and Palghat P Vaidyanathan. "Coprime sampling and the MUSIC algorithm". In: *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE*. IEEE. 2011, pp. 289–294.

[Ran+13] Juri Ranieri et al. "Phase retrieval for sparse signals: Uniqueness conditions". In: *arXiv preprint arXiv:1308.3058* (2013).

[Ray79] Lord Rayleigh. "XXXI. Investigations in optics, with special reference to the spectroscope". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 8.49 (1879), pp. 261–274.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3 (2010), pp. 471–501.

[RK89] Richard Roy and Thomas Kailath. "ESPRIT-estimation of signal parameters via rotational invariance techniques". In: *Acoustics, Speech and Signal Processing, IEEE Transactions on* 37.7 (1989), pp. 984–995.

[RSZ94] CR Rao, Chellury R Sastry, and Bin Zhou. "Tracking the direction of arrival of multiple moving targets". In: *Signal Processing, IEEE Transactions on* 42.5 (1994), pp. 1133–1144.

[SBE12] Yoav Shechtman, Amir Beck, and Yonina C Eldar. "Efficient phase retrieval of sparse signals". In: *Electrical and Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*. IEEE. 2012, pp. 1–5.

[SBE14] Yoav Shechtman, Andre Beck, and Yonina C Eldar. "GESPAR: Efficient phase retrieval of sparse signals". In: *Signal Processing, IEEE Transactions on* 62.4 (2014), pp. 928–938.

[Sch86] Ralph O Schmidt. "Multiple emitter location and signal parameter estimation". In: *Antennas and Propagation, IEEE Transactions on* 34.3 (1986), pp. 276–280.

[She+15] Yoav Shechtman et al. "Phase retrieval with application to optical imaging: a contemporary overview". In: *Signal Processing Magazine, IEEE* 32.3 (2015), pp. 87–109.

[SR15] Philip Schniter and Sundeep Rangan. "Compressive phase retrieval via generalized approximate message passing". In: *Signal Processing, IEEE Transactions on* 63.4 (2015), pp. 1043–1055.

[SS12] Dennis L Sun and Julius O Smith III. “Estimating a signal from a magnitude spectrogram via convex optimization”. In: *arXiv preprint arXiv:1209.2076* (2012).

[SSL90] Steven S Skiena, Warren D Smith, and Paul Lemke. “Reconstructing sets from interpoint distances”. In: *Proceedings of the sixth annual symposium on Computational geometry*. ACM. 1990, pp. 332–339.

[Sza+12] A Szameit et al. “Sparsity-based single-shot subwavelength coherent diffractive imaging”. In: *Nature materials* 11.5 (2012), pp. 455–459.

[Tan+13] Gongguo Tang et al. “Compressed sensing off the grid”. In: *Information Theory, IEEE Transactions on* 59.11 (2013), pp. 7465–7490.

[TF09] T Engin Tuncer and Benjamin Friedlander. *Classical and modern direction-of-arrival estimation*. Academic Press, 2009.

[Ton+95] Lang Tong et al. “Blind channel identification based on second-order statistics: A frequency-domain approach”. In: *Information Theory, IEEE Transactions on* 41.1 (1995), pp. 329–334.

[Tro15] Joel A Tropp. “Convex recovery of a structured signal from independent random linear measurements”. In: *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[VB96] Lieven Vandenberghe and Stephen Boyd. “Semidefinite programming”. In: *SIAM review* 38.1 (1996), pp. 49–95.

[Ver10] Roman Vershynin. “Introduction to the non-asymptotic analysis of random matrices”. In: *arXiv preprint arXiv:1011.3027* (2010).

[VP11] Palghat P Vaidyanathan and Piya Pal. “Sparse sensing with co-prime samplers and arrays”. In: *Signal Processing, IEEE Transactions on* 59.2 (2011), pp. 573–586.

[W+53] James D Watson, Francis HC Crick, et al. “Molecular structure of nucleic acids”. In: *Nature* 171.4356 (1953), pp. 737–738.

[Wal63] Adriaan Walther. “The question of phase retrieval in optics”. In: *Journal of Modern Optics* 10.1 (1963), pp. 41–49.

[Zha+10] Xiaofei Zhang et al. “Direction of departure (DOD) and direction of arrival (DOA) estimation in MIMO radar with reduced-dimension MUSIC”. In: *Communications Letters, IEEE* 14.12 (2010), pp. 1161–1163.

[ZHY13] Guoan Zheng, Roarke Horstmeyer, and Changhuei Yang. “Wide-field, high-resolution Fourier ptychographic microscopy”. In: *Nature photonics* 7.9 (2013), pp. 739–745.

[1]Endnotes are notes that you can use to explain text in a document.

APPENDIX

Chapter 8

SUPPLEMENTARY MATERIALS FOR CHAPTER III

8.1 Proof of Theorem 3.2.1

We use the following notation in this section: If $\mathbf{x}$ is a signal of length l_x, then $\mathbf{x} = \{x_0, x_1, \ldots, x_{l_x-1}\}$ and $\{x_0, x_{l_x-1}\} \neq 0$. $\equiv$ implies equality up to time-shift, conjugate-flip and global phase, i.e., equality up to trivial ambiguities. $\tilde{\mathbf{x}}$ denotes the signal obtained by conjugate-flipping $\mathbf{x}$, i.e., $\tilde{\mathbf{x}} = \{x^\star_{l_x-1}, x^\star_{l_x-2}, \ldots, x^\star_0\}$.

In order to characterize the set of signals with aperiodic support which cannot be uniquely recovered by (3.2), we make use of the following lemma:

Lemma 8.1.1. *If two non-equivalent signals* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *have the same autocorrelation, then there exist signals* $\mathbf{g}$ *and* $\mathbf{h}$*, of lengths* l_g *and* l_h *respectively, such that*

(i) $\mathbf{x}_1 \equiv \mathbf{g} \star \mathbf{h}$ & $\mathbf{x}_2 \equiv \mathbf{g} \star \tilde{\mathbf{h}}$

(ii) $l_g + l_h - 1 = l_x$*, and* $l_g, l_h \geq 2$

(iii) $h_0 = 1$, $h_{l_h-1} \neq 0$, $g_0 \neq 0$, $g_{l_g-1} \neq 0$

(iv) $l_g \geq l_h$.

Proof. (i) Let $X_1(z)$, $X_2(z)$, $G(z)$ and $H(z)$ be the z-transforms of the signals $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{g}$ and $\mathbf{h}$ respectively. Since $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same autocorrelation, we have

$$A(z) = X_1(z)X_1^\star(z^{-\star}) = X_2(z)X_2^\star(z^{-\star}),$$

where $A(z)$ is the z-transform of the autocorrelation of $\mathbf{x}_1$ and $\mathbf{x}_2$. If z_0 is a zero of $A(z)$, then $z_0^{-\star}$ is also a zero of $A(z)$[4]. For every such pair of zeros $(z_0, z_0^{-\star})$, z_0 can be assigned to $X_1(z)$ or $X_1^\star(z^{-\star})$, and $X_2(z)$ or $X_2^\star(z^{-\star})$. Let $P_1(z)$, $P_2(z)$ and $P_3(z)$ be the polynomials constructed from such pairs of zeros which are assigned

[4]The problem of recovering $X(z)$ from $A(z)$ is hence equivalent to the problem of assigning pairs of zeros of the form $(z_0, z_0^{-\star})$ between $X(z)$ and $X^\star(z^{-\star})$ (see [Hof64; JOH12c]). Since $A(z)$ can have at most n such pairs, this can be done in at most 2^n ways and hence for a given autocorrelation, there can be at most 2^n non-equivalent solutions.

to $(X_1(z), X_2(z))$, $(X_1(z), X_2^\star(z^{-\star}))$ and $(X_1^\star(z^{-\star}), X_2(z))$ respectively. Note that $P_3(z) \equiv P_2^\star(z^{-\star})$. We have

$$X_1(z) \equiv P_1(z)P_2(z) \;\; \& \;\; X_2(z) \equiv P_1(z)P_2^\star(z^{-\star}),$$

and hence $X_1(z)$ and $X_2(z)$ can be written as

$$X_1(z) \equiv G(z)H(z) \quad X_2(z) \equiv G(z)H^\star(z^{-\star}),$$

where $G(z) = P_1(z)$ and $H(z) = P_2(z)$, or equivalently

$$\mathbf{x}_1 \equiv \mathbf{g} \star \mathbf{h} \quad \mathbf{x}_2 \equiv \mathbf{g} \star \tilde{\mathbf{h}}$$

in the time domain.

(ii) If two signals of lengths l_g and l_h are convolved, the resulting signal (in this case $\mathbf{x}_1$ or $\mathbf{x}_2$) will be of length $l_g + l_h - 1$. l_g and l_h are greater than or equal to 2 because otherwise, $\mathbf{x}_1$ and $\mathbf{x}_2$ will be equivalent.

(iii) Since $\mathbf{g}$ and $\mathbf{h}$ are signals of lengths l_g and l_h respectively, $\{h_0, h_{l_h-1}, g_0, g_{l_g-1}\} \neq 0$ by definition. The signals $(\mathbf{g} \star \mathbf{h})$ and $(\alpha\mathbf{g} \star \mathbf{h}/\alpha)$ are the same for any constant α. Hence, without loss of generality, we can set $h_0 = 1$.

(iv) Suppose $l_g < l_h$. Since $\mathbf{x}_1$ and $\tilde{\mathbf{x}}_2$ have the same autocorrelation, we can apply part (i) of this lemma to signals $\mathbf{x}_1$ and $\tilde{\mathbf{x}}_2$ to get $\mathbf{x}_1 = \mathbf{h} \star \mathbf{g}$ and $\tilde{\mathbf{x}}_2 = \mathbf{h} \star \tilde{\mathbf{g}}$. Hence, without loss of generality, the signals $\mathbf{g}$ and $\mathbf{h}$ can be interchanged. □

First, we will prove the theorem for the $k = n - 1$ case as it is relatively easier and provides intuition for the $k < n - 1$ case.

Case I: $k = n - 1$

$\mathcal{S}_{n-1}$, i.e., the set of signals with aperiodic support and sparsity equal to $n - 1$, has $2(n-1)$ degrees of freedom (as each nonzero location can have a complex value and hence can have 2 degrees of freedom). We will show that the set of signals in $\mathcal{S}_{n-1}$ that cannot be recovered by (3.2) has degrees of freedom strictly less than $2(n - 1)$.

Suppose $\mathbf{x}_1 \in \mathcal{S}_{n-1}$ is not recoverable by (3.2), then there must exist another signal $\mathbf{x}_2$, with sparsity less than or equal to $n - 1$, which has the same autocorrelation. At least one location in both $\mathbf{x}_1$ and $\mathbf{x}_2$ have a value zero, say $x_{1,i} = 0$ and $x_{2,j} = 0$ for some $1 \leq i, j \leq n - 2$. Note that we can always find an i and j in this range as $\mathbf{x}_1$ has aperiodic support and the lengths of $\mathbf{x}_1$ and $\mathbf{x}_2$ are the same. From Lemma

8.1.1, there must exist two signals $\mathbf{g}$ and $\mathbf{h}$, of lengths l and $n-l+1$ for some $\frac{n+1}{2} \leq l \leq n-1$, such that

$$\sum_r g_r h_{i-r} = 0 \;\; \& \;\; \sum_r g_r h^{\star}_{n-l-j+r} = 0 \tag{8.1}$$

and $\{g_0, g_{l-1}, h_{n-l}\} \neq 0, h_0 = 1$.

Our strategy is the following: We will count the degrees of freedom of the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ which satisfy (8.1) for some choice of $\{l, i, j\}$ and show that it is strictly less than $2(n-1)$.

The following arguments can be made for any particular choice of $\{l, i, j\}$: the two bilinear equations in (8.1) can be represented in the matrix form as

$$\mathbf{Hg} = 0,$$

where $\mathbf{g}$ is the column vector $\{g_0, g_1, \ldots, g_{l-1}\}^T$ and $\mathbf{H}$ is the $2 \times l$ matrix containing the corresponding entries of $\mathbf{h}$ given by (8.1). For example, if $i < j < l-1$, then (8.1) can be written as

$$\begin{bmatrix} h_i & h_{i-1} & \ldots & h_0 & \ldots & 0 & 0 & \ldots \\ h^{\star}_{n-l-j} & h^{\star}_{n-l-j+1} & \ldots & \ldots & \ldots & h^{\star}_{n-l} & 0 & \ldots \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ \ldots \\ g_{l-1} \end{bmatrix} = 0.$$

The degrees of freedom of the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ which satisfy the system of equations (8.1) can be calculated as follows: Since $\mathbf{h}$ is a complex vector of length $n-l+1$ and $h_0 = 1$, $\mathbf{h}$ can have $2(n-l)$ degrees of freedom. For each $\mathbf{h}$, since each independent row of $\mathbf{H}$ restricts $\mathbf{g}$ by one dimension in the complex space, or equivalently, 2 degrees of freedom, $\mathbf{g}$ can have $2l - 2 \times \text{rank}(\mathbf{H})$ degrees of freedom.

There are two possibilities:

(i) $\text{rank}(\mathbf{H}) = 2$: This happens generically, hence $\mathbf{h}$ can have $2(n-l)$ degrees of freedom. For each choice of $\mathbf{h}$ such that $\text{rank}(\mathbf{H}) = 2$, $\mathbf{g}$ can have $2(l-2)$ degrees of freedom. Hence, the degrees of freedom of the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ in this case which satisfy (8.1) is $2(n-l) + 2(l-2) = 2(n-2)$.

(ii) $\text{rank}(\mathbf{H}) = 1$: In this case, each 2×2 submatrix of $\mathbf{H}$ must be rank 1, which could happen for some $\mathbf{h}$. The set of such $\mathbf{h}$ has degrees of freedom at most $2(n-l)-1$, as the degrees of freedom of at least one entry of $\mathbf{h}$ gets reduced by one. For example,

if the 2×2 submatrix is $[h_1, h_0; h^\star_{n-l-1}, h^\star_{n-l}]$, then $h^\star_{n-l-1} = \frac{h_1 h^\star_{n-l}}{h_0}$ and hence, once h_1 and h_{n-l} are chosen, h_1 can take precisely one value and hence 2 degrees of freedom are lost for h_1. For some 2×2 matrices, like $[h_1, h_0; h^\star_0, h^\star_1]$, the condition is $|h_1| = |h_0|$ because of which there will be a loss of one degree of freedom for h_1. For each choice of $\mathbf{h}$ such that $\text{rank}(\mathbf{H}) = 1$, $\mathbf{g}$ can have $2(l-1)$ degrees of freedom. Hence, the degrees of freedom of the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ in this case which satisfy (8.1) is at most $2(n-l) - 1 + 2(l-1) = 2(n-1) - 1$.

We have shown that for any particular choice of $\{l, i, j\}$, the degrees of freedom of the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ which satisfy (8.1) is at most $2(n-1)-1$. The degrees of freedom of the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ which satisfy (8.1) for *some* choice of $\{l, i, j\}$ can be obtained by considering each valid choice of $\{l, i, j\}$ and taking a union of the resulting $\{\mathbf{g}, \mathbf{h}\}$. Since the union of a finite number of manifolds with degrees of freedom at most $2(n-1) - 1$ is a manifold with degrees of freedom at most $2(n-1) - 1$, the set of signals in $\mathcal{S}_{n-1}$ which cannot be recovered uniquely by (3.2) is a manifold of dimension at most $2(n-1) - 1$, which is strictly less than $2(n-1)$. Hence, almost all signals in $\mathcal{S}_{n-1}$ can be uniquely recovered by solving (3.2).

(ii) **Case II**: $k \leq n - 1$

$\mathcal{S}_k$, i.e., the set of signals with aperiodic support and sparsity equal to k, has $2k$ degrees of freedom (as each nonzero location can have a complex value and hence can have 2 degrees of freedom). This can also be calculated as follows:

Consider the set of signals of length $l_x \geq 3$ which have zeros in the locations $\{i_1, i_2, \ldots, i_{l_x-k}\}$ (the indices are arranged in increasing order, $i_1 \geq 1$ and $i_{l_x-k} \leq l_x - 2$ by definition). Since any $\mathbf{x}$ of length l_x can be written as $\mathbf{g} \star \mathbf{h}$, where $\mathbf{g}$ and $\mathbf{h}$ are signals of lengths l and $l_x - l + 1$ for *any* $2 \leq l \leq l_x - 1$ (see proof of Lemma 8.1.1), there must exist two signals $\mathbf{g}$ and $\mathbf{h}$, of lengths l and $l_x - l + 1$ for *any* $2 \leq l \leq l_x - 1$ such that

$$\sum_r g_r h_{i_p - r} = 0 \ \ \forall \ \ 1 \leq p \leq l_x - k \tag{8.2}$$

and $\{g_0, g_{l-1}, h_{l_x-l}\} \neq 0, h_0 = 1$. The degrees of freedom of the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ which satisfy the system of equations (8.2) can be calculated as follows: Let $\mathcal{M}_1$ be the manifold containing the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ which satisfy the following set of equations:

$$\sum_r g_r h_{i_p - r} = 0 \ \ \forall \ \ 1 \leq p \leq z_1, \tag{8.3}$$

where z_1 is the maximum integer such that $i_{z_1} < l-1$. Also, let $\mathcal{M}_2$ be the manifold containing the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ which satisfy the following set of equations:

$$\sum_r g_r h_{i_p - r} = 0 \ \ \forall \ \ z_1 + 1 \leq p \leq l_x - k. \tag{8.4}$$

The bilinear equations in (8.3) can be represented in the matrix form as

$$\mathbf{H}_1 \mathbf{g} = 0,$$

where $\mathbf{g}$ is the column vector $\{g_0, g_1, \ldots, g_{l-1}\}^T$ and $\mathbf{H}_1$ is a $z_1 \times l$ matrix containing the corresponding entries of $\mathbf{h}$ given by (8.3). The matrix $\mathbf{H}_1$ can be obtained by considering the rows corresponding to $\{i_1, i_2, \ldots, i_{z_1}\}$ of the following matrix:

$$\begin{bmatrix} h_0 & 0 & 0 & \ldots & \ldots & 0 & 0 \\ h_1 & h_0 & 0 & \ldots & \ldots & 0 & 0 \\ h_2 & h_1 & h_0 & 0 & \ldots & 0 & 0 \\ & & & \ldots & & & \\ & & & \ldots & & & \\ & & & \ldots & h_1 & h_0 & 0 \end{bmatrix}.$$

Note that $\text{rank}(\mathbf{H}_1) = z_1$ for all choices of $\mathbf{h}$ due to the fact that the columns corresponding to $\{i_1, i_2, \ldots, i_{z_1}\}$ of $\mathbf{H}_1$ have a lower triangular structure and $h_0 = 1$. Since $\mathbf{h}$ is a vector of length $l_x - l + 1$ with $h_0 = 1$, $\mathbf{h}$ has $2(l_x - l)$ degrees of freedom. $\mathbf{g}$ is a vector of length l and for each $\mathbf{h}$, since each independent row of $\mathbf{H}$ restricts $\mathbf{g}$ by one dimension in the complex space, or equivalently, 2 degrees of freedom, $\mathbf{g}$ can have $2l - 2 \times \text{rank}(\mathbf{H}_1) = 2l - 2z_1$ degrees of freedom. Hence, the manifold $\mathcal{M}_1$ has $2l - 2z_1 + 2(l_x - l) = 2(l_x - z_1)$ degrees of freedom. In other words, the manifold $\mathcal{M}_1$ has lost $2z_1$ degrees of freedom (from the maximum possible $2l_x$ degrees of freedom).

Similarly, the bilinear equations in (8.4) can be represented in the matrix form as

$$\mathbf{H}_2 \mathbf{g} = 0,$$

where $\mathbf{g}$ is the column vector $\{g_0, g_1, \ldots, g_{l-1}\}^T$ and $\mathbf{H}_2$ is a $(l_x - k - z_1) \times l$ matrix containing the corresponding entries of $\mathbf{h}$ given by (8.4). The matrix $\mathbf{H}_2$ can be obtained by considering the rows corresponding to $\{i_{z_1+1}, i_{z_1+2}, \ldots, i_{l_x-k}\}$ of the

following matrix:

$$\begin{bmatrix} \dots & \dots & & \dots & h_2 & h_1 & h_0 \\ 0 & \dots & \dots & \dots & \dots & h_2 & h_1 \\ 0 & 0 & \dots & \dots & \dots & \dots & h_2 \\ & & & \dots & & & \\ & & & \dots & 0 & h_{l_x-l} & h_{l_x-l-1} \\ & & & \dots & 0 & 0 & h_{l_x-l} \end{bmatrix}$$

and hence $\text{rank}(\mathbf{H}_2) = l_x - k - z_1$ for all choices of $\mathbf{h}$ due to the fact that the columns corresponding to $\{i_{z_1+1}, i_{z_1+2}, \dots, i_{l_x-k}\}$ have an upper triangular structure and $h_{l_x-l} \neq 0$. Since $\mathbf{h}$ is a vector of length $l_x - l + 1$ with $h_0 = 1$, $\mathbf{h}$ can have $2(l_x - l)$ degrees of freedom, and since $\mathbf{g}$ is a vector of length l, it can have $2l - 2 \times \text{rank}(\mathbf{H}_2) = 2l - 2(l_x - k - z_1)$ degrees of freedom. Hence, the manifold $\mathcal{M}_2$ has $2l - 2(l_x - k - z_1) + 2(l_x - l) = 2(k + z_1)$ degrees of freedom. In other words, the manifold $\mathcal{M}_2$ has lost $2l_x - 2(k + z_1)$ degrees of freedom (from the maximum possible $2l_x$ degrees of freedom).

The number of degrees of freedom lost by the manifold $\mathcal{M}_1 \cap \mathcal{M}_2$ is, in this case, given by the sum of the number of degrees of freedom lost by the manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$, i.e., $2l_x - 2k$ (see [Fan12] for a proof based on codimension). This can be seen as follows: The total loss of degrees of freedom in the manifold $\mathcal{M}_1 \cap \mathcal{M}_2$ is given by the sum of the loss of degrees of freedom in each individual set minus the degrees lost due to overcounting (due to the fact that some linear combinations of the bilinear equations in $\mathcal{M}_2$ can be written as linear combinations of the bilinear equations in $\mathcal{M}_1$ for all possible $\{\mathbf{g}, \mathbf{h}\}$ considered in $\mathcal{M}_1$ (or vice versa)). Since $g_{l-1} \neq 0$, by observing the coefficients of g_{l-1} in $\mathcal{M}_1$ and $\mathcal{M}_2$, it can be seen that a necessary condition for some linear combinations of the bilinear equations in $\mathcal{M}_2$ to be written as linear combinations of the bilinear equations in $\mathcal{M}_1$ for all possible $\{\mathbf{g}, \mathbf{h}\}$ considered in $\mathcal{M}_1$ is that the corresponding linear combinations of $\{h_{i_{l_x-k}-l+1}, h_{i_{l_x-k-1}-l+1}, \dots, h_{i_{z_1+1}-l+1}\}$ must be zero for all $\mathbf{h}$ considered in $\mathcal{M}_1$. Hence, if $\{h_{i_{l_x-k}-l+1}, h_{i_{l_x-k-1}-l+1}, \dots, h_{i_{z_1+1}-l+1}\}$ were to be chosen from a manifold which has lost $2c$ degrees of freedom, at most c independent linear combinations of $\{h_{i_{l_x-k}-l+1}, h_{i_{l_x-k-1}-l+1}, \dots, h_{i_{z_1+1}-l+1}\}$ could be zero (as each independent linear combination reduces one dimension in the complex space, or equivalently, two degrees of freedom), and hence removal of c independent linear combinations of the bilinear equations in $\mathcal{M}_2$ will definitely make the two systems of bilinear equations independent. Hence, the loss of degrees of freedom of $\mathcal{M}_1$ and $\mathcal{M}_2$ is $2z_1 + 2c$ and at least $2l_x - 2(k + z_1) - 2c$ respectively (for any valid choice of c),

because of which the loss of degrees of freedom of $\mathcal{M}_1 \cap \mathcal{M}_2$ is at least $2l_x - 2k$. Since for $c = 0$, this bound is tight, the degrees of freedom of the set $\mathcal{M}_1 \cap \mathcal{M}_2$ is $2k$.

The set $\mathcal{S}_k$ is constructed by considering every possible choice of $\{i_1, \ldots, i_{l_x-k}, l, l_x\}$ and taking a union of the corresponding $\{\mathbf{g}, \mathbf{h}\}$. Since the union of a finite number of manifolds with degrees of freedom $2k$ is a manifold with degrees of freedom $2k$, $\mathcal{S}_k$ has $2k$ degrees of freedom.

Suppose $\mathbf{x}_1 \in \mathcal{S}_k$, of length l_x, is not recoverable by (3.2), then there must exist another signal $\mathbf{x}_2$ of length l_x, with sparsity less than or equal to k, which has the same autocorrelation. At least $l_x - k$ locations in both $\mathbf{x}_1$ and $\mathbf{x}_2$ have a value zero, let these locations be denoted by $\{i_1, i_2, \ldots, i_{l_x-k}\}$ and $\{j_1, j_2, \ldots, j_{l_x-k}\}$ respectively. Then from Lemma 8.1.1, there must exist two signals $\mathbf{g}$ and $\mathbf{h}$, of lengths l and $l_x - l + 1$ for some $\frac{l_x+1}{2} \leq l \leq l_x - 1$, such that

$$\sum_r g_r h_{i_p - r} = 0 \;\; \& \;\; \sum_r g_r h^{\star}_{l_x - l - j_p + r} = 0 \tag{8.5}$$

and $\{g_0, g_{l-1}, h_{l_x-l}\} \neq 0, h_0 = 1$.

Our strategy is the following: We will count the degrees of freedom of the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ which satisfy (8.5) for some choice of $\{l, l_x, i_1, \ldots, i_{l_x-k}, j_1, \ldots, j_{l_x-k}\}$ and show that it is strictly less than $2k$ if $\mathbf{x}_1$ has aperiodic support. First, we will show that the degrees of freedom of this set is strictly less than $2k$ if there is some $1 \leq p \leq l_x - k$ such that $j_p \notin \{i_1, i_2, \ldots, i_{l_x-k}\}$, i.e., when the two signals $\mathbf{x}_1$ and $\mathbf{x}_2$ have different support (as a consequence, for most signals, this proves that (3.2) correctly identifies their support, irrespective of whether they have periodic or aperiodic support). We then show that if there is no $1 \leq p \leq l_x - k$ such that $j_p \notin \{i_1, i_2, \ldots, i_{l_x-k}\}$, i.e., the two signals $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same support, the degrees of freedom is strictly less than $2k$ if the support of $\mathbf{x}_1$ (or equivalently $\mathbf{x}_2$) is aperiodic.

The following arguments hold for any particular $\{l, l_x, i_1, \ldots, i_{l_x-k}, j_1, \ldots, j_{l_x-k}\}$:

Suppose there exists at least one $1 \leq p \leq l_x - k$ such that $j_p \notin \{i_1, i_2, \ldots, i_{l_x-k}\}$. If $j_p < l - 1$, construct the manifold $\mathcal{M}_1^0$ using the $z_1 + 1$ bilinear equations corresponding to the indices $\{i_1, i_2, \ldots, i_{z_1}\}$ and j_p. In matrix notation, these bilinear equations can be represented as $\mathbf{H}_3\mathbf{g} = 0$ where $\text{rank}(\mathbf{H}_3) = z_1 + 1$ as the columns corresponding to the indices $\{i_1, i_2, \ldots, i_{z_1}, j_p\}$ (rearrange the indices in increasing order) has a lower triangular structure and $h_0 = 1, h_{l_x-l} \neq 0$. Hence, $\mathcal{M}_1^0$ has lost

$2(z_1 + 1)$ degrees of freedom. The manifold $\mathcal{M}_2^0$ can be constructed the same way as $\mathcal{M}_2$ and hence $\mathcal{M}_2^0$ has lost $2l_x - 2(k + z_1)$ degrees of freedom. Due to the same arguments as in the case of $\mathcal{M}_1 \cap \mathcal{M}_2$, $\mathcal{M}_1^0 \cap \mathcal{M}_2^0$ loses $2(z_1 + 1) + 2l_x - 2(k + z_1) = 2l_x - 2k + 2$ degrees of freedom, i.e., has $2k - 2$ degrees of freedom. If $j_p \geq l - 1$, the arguments can be repeated by incorporating the bilinear equation corresponding to j_p in $\mathcal{M}_2^0$ instead of $\mathcal{M}_1^0$ to see that $\mathcal{M}_1^0 \cap \mathcal{M}_2^0$ has strictly less than $2k$ degrees of freedom.

By considering every possible choice of $\{l_x, l, i_1, i_2, \ldots, i_{l_x-k}, j_1, j_2, \ldots, j_{l_x-k}\}$ such that there is some $1 \leq p \leq l_x - k$ such that $j_p \notin \{i_1, i_2, \ldots, i_{l_x-k}\}$, and taking a union of the corresponding $\{\mathbf{g}, \mathbf{h}\}$, we see that the set of signals $\mathbf{x}_1 \in \mathcal{S}_k$ which cannot be recovered by (3.2), due to the fact that there exists another $(\leq k)$-sparse signal which has the same autocorrelation and different support, is a manifold with degrees of freedom strictly less than $2k$.

Suppose there is no $1 \leq p \leq l_x - k$ such that $j_p \notin \{i_1, i_2, \ldots, i_{l_x-k}\}$. This is the case when $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same support and the same autocorrelation. The manifold $\mathcal{M}_1^1$ is constructed using the $2z_1$ equations corresponding to the indices $\{i_1, j_1, i_2, j_2, \ldots, i_{z_1}, j_{z_1}\}$ (the corresponding equations in matrix notation being $\mathbf{H}_5\mathbf{g} = 0$) and the manifold $\mathcal{M}_2^1$ is constructed using the $2(l_x - k - z_1)$ equations corresponding to the indices $\{i_{z_1+1}, j_{z_1+1}, \ldots i_{l_x-k}, j_{l_x-k}\}$ (the corresponding equations in the matrix notation being $\mathbf{H}_6\mathbf{g} = 0$).

In this case, $\text{rank}(\mathbf{H}_5) \geq z_1$ for all choices of $\mathbf{h}$. For the choices of $\mathbf{h}$ with $\text{rank}(\mathbf{H}_5) \geq z_1 + 1$, the manifold $\mathcal{M}_1^1$ loses at least $2(z_1 + 1)$ degrees of freedom, because of which $\mathcal{M}_1^1 \cap \mathcal{M}_2^1$ will have at most $2k - 2$ degrees of freedom due to the same arguments as in the case of $\mathcal{M}_1 \cap \mathcal{M}_2$. For the choices of $\mathbf{h}$ with $\text{rank}(\mathbf{H}) = z_1$, we will show that the degrees of freedom corresponding to the entry h_{l_x-l} will go down by at least one if $\mathbf{x}_1$ has aperiodic support, because of which $\mathcal{M}_2^1$ will lose $2z_1 + 1$ degrees of freedom, and hence $\mathcal{M}_1^1 \cap \mathcal{M}_2^1$ will have at most $2k - 1$ degrees of freedom.

Consider every 2×2 submatrix involving the first two rows of $\mathbf{H}_5$. If even one of them is full rank, then the rank of $\mathbf{H}_5$ would be at least $z_1 + 1$. If the rank of all such submatrices are 1, then they have to satisfy equations of the form $h^\star_{l_x-l} h_1 = h_0 h^\star_{l_x-l-1}$, and so on. This equation, for example, removes at least one degree of freedom for h_{l_x-l} unless $h_1 = h_{l_x-l-1} = 0$. By considering every 2×2 submatrix involving the first two rows of $\mathbf{H}_5$ and involving the column corresponding to i_1, we can conclude that there is a loss in the degrees of freedom of h_{l_x-l} unless $h_{i_1} = h_{i_1-1} = \ldots = h_1 = 0$. In this event, the first two rows become equivalent to the

condition $g_{i_1} = 0$ as $h_0 = 1$. By considering the third and fourth row and repeating the same arguments, we can conclude that there is a loss in the degrees of freedom of $h_{l_x - l}$ unless $g_{i_2} = 0$. Continuing similarly, we see that a necessary condition for the degrees of freedom of $h_{l_x - l}$ to not go down when $\text{rank}(\mathbf{H}_5) = z_1$ is: $g_{i_p} = 0$ for all $1 \leq p \leq z_1$. The arguments can be repeated exactly the same way using $\mathbf{H}_6$ (going from last row to first as it is upper triangular) to get a further necessary condition $h_{i_p - l + 1} = 0$ for all $z_1 + 1 \leq p \leq l_x - k$.

We have established that there is a loss of degrees of freedom of $h_{l_x - l}$ unless $\mathbf{h}$ has $l_x - k - z_1$ particular entries with value 0 and $\mathbf{g}$ has z_1 particular entries with value 0. Consider the set of all possible $\{\mathbf{g}, \mathbf{h}\}$ such that $g_{i_p} = 0$ for all $1 \leq p \leq z_1$ and $h_{i_p - l + 1} = 0$ for all $z_1 + 1 \leq p \leq l_x - k$. The set of the signals $\mathbf{g} \star \mathbf{h}$ obtained from such $\mathbf{g}$ and $\mathbf{h}$ is a manifold with $2k$ degrees of freedom. We will show that most of these signals have a sparsity strictly greater than k if the support of $\mathbf{x}_1$ is aperiodic, which will complete the proof as the degrees of freedom of the set of such $\{\mathbf{g}, \mathbf{h}\}$ which satisfy (8.5) has to further reduce by at least one in order to meet the sparsity constraints.

Consider the set of all $\mathbf{g}$ that have nonzero entries in the indices $\{u_0 = 0, u_1, \ldots, u_{a-1}\}$ (and zero in other indices) and the set of all $\mathbf{h}$ that have nonzero entries in the indices $\{v_0 = 0, v_1, \ldots, v_{b-1}\}$ (and zero in other indices). Then, almost surely (the set of violations is measure zero), the set of all possible $\mathbf{g} \star \mathbf{h}$ will have nonzero entries in the following $a + b - 1$ locations: $\{u_0 = 0, u_1, \ldots, u_{a-1}, u_{a-1} + v_1, \ldots, u_{a-1} + v_{b-1}\}$. If there has to be no more locations with nonzero entries almost surely: Consider the terms of the form $u_{a-2} + v_p$ for $0 \leq p \leq b - 1$. Since there can be b such terms, and all of them are greater than u_{a-3} and lesser than $u_{a-1} + v_{b-1}$, they have to precisely be equal to the following b terms in the same order: $\{u_{a-2}, u_{a-1}, u_{a-1} + v_1, \ldots, u_{a-1} + v_{b-2}\}$. This gives the condition that $v_p - v_{p-1}$ is equal to $u_{a-1} - u_{a-2}$ for all $1 \leq p \leq b - 1$. Similarly, by observing that the following $a + b - 1$ locations $\{v_0 = 0, v_1, \ldots, v_{b-1}, v_{b-1} + u_1, \ldots, v_{b-1} + u_{a-1}\}$ almost surely have nonzero values and considering terms of the form $v_{b-2} + u_p$ for $0 \leq p \leq a - 1$, we get the condition that $u_p - u_{p-1}$ is equal to $v_{b-1} - v_{b-2}$ for all $1 \leq p \leq a - 1$. Hence, if the signal has aperiodic support, then almost all $\mathbf{g} \star \mathbf{h}$ have strictly greater than $a + b - 1$ nonzero entries. Substituting $a = l - z_1$ and $b = k + z_1 - l + 1$, we see that $a + b - 1 = k$ and hence, almost always, the resulting convolved signal has sparsity strictly greater than k.

By considering every possible choice of $\{l_x, l, i_1, i_2, \ldots, i_{l_x - k}, j_1, j_2, \ldots, j_{l_x - k}, \}$ such

that there is no $1 \leq p \leq l_x - k$ such that $j_p \notin \{i_1, i_2, \ldots, i_{l_x-k}\}$ and taking the union of the corresponding $\{\mathbf{g}, \mathbf{h}\}$, we conclude that the set of signals $\mathbf{x}_1 \in \mathcal{S}_k$ which cannot be recovered by (3.2), due to the fact that there exists another signal with the same autocorrelation and same support, is a manifold with degrees of freedom strictly less than $2k$.

Hence, we have shown that (3.2) can recover almost all sparse signals with aperiodic support for every sparsity k such that $k \leq n - 1$.

8.2 Proof of Theorem 3.3.2

In this section, V is a subset of $\{0, 1, \ldots, n-1\}$, constructed as follows: For each $0 \leq i \leq n-1$, i belongs to the support independently with probability $\frac{s}{n}$. In order to resolve the trivial ambiguity due to time-shift, we will shift the set so that $i = 0$ belongs to the support. Let these entries be denoted by $V = \{v_0, v_1, \ldots, v_{k-1}\}$. We have $v_0 = 0$, which will ensure $V \subseteq W$. The distribution of V (if the time-shift was c units) is as follows: $0 \in V$ with probability 1. For all $0 < i < n - c$, $i \in V$ with probability $\frac{s}{n}$ independently, and for all $i \geq n - c$, $i \in V$ with probability 0. Hence, irrespective of the value of the time-shift c, the following bound can be used: For any $i > 0$, $i \in V$ with probability less than or equal to $\frac{s}{n}$ independently.

Instead of resolving the trivial ambiguity due to flipping, we will use the following proof strategy (as the distribution of V is easier to work with compared to U): We will show that if the steps of the support recovery algorithm are done using entries of the form v_{0i}, the failure probability can be bounded by $O(n^{-0.1\epsilon})$. The *same* arguments can be used to show that if the steps are done using entries of the form $v_{k-i-1,k-1}$, the failure probability can be bounded by $O(n^{-0.1\epsilon})$. Since u_{0i} is either equal to v_{0i} or $v_{k-i-1,k-1}$, this would imply that if the steps are done using entries of the form u_{0i}, the support recovery algorithm will succeed with the desired probability.

Lemma 8.2.1. *The probability that an integer $l > 0$, which does not belong to V, belongs to W is bounded by $O\left(\frac{s^2}{n}\right)$.*

Proof. For $l \in W$ to happen, there must exist at least one g such that $g, g + l \in V$. Hence,

$$Pr\{l \in W\} = Pr\Big\{ \bigcup_{g=0}^{n-l-1} g, g + l \in V \Big\}.$$

There can be two cases:

(i) $g = 0$: In this case, $Pr\{g, g + l \in V\} = Pr\{l \in V\}$.

(ii) $g > 0$: In this case, for each g, $Pr\{g, g + l \ \in V\} \leq \left(\frac{s}{n}\right)^2$ due to independence. Also, since g can take at most $n - l$ distinct values, by union bound, we have:

$$Pr\{l \in W\} = \sum_{g=0}^{n-l-1} Pr\Big\{\{g, g + l\} \in V\Big\} \leq Pr\{l \in V\} + \frac{s^2}{n}.$$

$Pr\{l \in W\}$ can be written as:

$$Pr\{l \in W | l \in V\}Pr\{l \in V\} + Pr\{l \in W | l \notin V\}Pr\{l \notin V\}.$$

Since $Pr\{l \in W | l \in V\} = 1$ (as $0, l \in V$), we have

$$Pr\{l \in W | l \notin V\} = \frac{Pr\{l \in W\} - Pr\{l \in V\}}{Pr\{l \notin V\}}.$$

Using the fact that $Pr\{l \notin V\} \geq 1 - \frac{s}{n}$, we can obtain the following bound:

$$Pr\{l \in W | l \notin V\} \leq \frac{\frac{s^2}{n}}{1 - \frac{s}{n}} = O\left(\frac{s^2}{n}\right). \tag{8.6}$$

In fact, since $s \leq n^{\frac{1}{2}}$, we have $\frac{s}{n} \leq \frac{1}{n^{\frac{1}{2}}}$ which is less than $\frac{1}{2}$ for $n \geq 4$. Consequently, we have the upper bound $\frac{2s^2}{n}$ for $n \geq 4$. □

Lemma 8.2.2 (Intersection Step)**.** *The probability that an integer $l > v_{01}$, which does not belong to V, belongs to $W \cap (W + v_{01})$ is bounded by $O\left(\frac{s^4}{n^2}\right)$.*

Proof. We can write

$$Pr\{l \in W \cap (W + v_{01})\} = Pr\{\ l, l - v_{01} \ \in W\}$$

$$= \sum_d Pr\{v_{01} = d\} \times Pr\{\ l, l - d \ \in W | v_{01} = d\}.$$

For the two events $l, l - d \in W$ to happen, there has to be some g such that $g, g + l \in V$ (this explains the event $l \in W$) and some h such that $h, h + l - d \in V$ (this explains the event $l - d \in W$). Since we are conditioning on $v_{01} = d$, note that $0, d \in V$, and $1, 2, \ldots, d - 1 \notin V$ and for all $i > d$, $i \in V$ with probability less than or equal to $\frac{s}{n}$ independently.

Case I: $d \neq \frac{l}{2}$

The events $l, l - d \in W$ can happen due to one of the following cases:

(i) There exists some integer g whose presence in V, using $0, d \in V$, explains both the events $l \in W$ and $l - d \in W$. $g = l$ is the only integer which comes under this case (i.e., if $l \in V$, then both the events can be explained). The probability of this case happening is $Pr\{l \in V\}$.

(ii) There exists some distinct pair of integers $\{g, h\}$ whose presence in V, using $0, d \in V$, explains both the events $l \in W$ and $l - d \in W$. There are at most three possibilities: $\{g, h\} = \{\{l - d, l + d\}, \{l, l - d\}, \{l, l + d\}\}$ (possibilities involving $l - d$ can happen only for $l > 2d$, hence there is only one possibility for $l < 2d$). The probability of each of these possibilities can be bounded $\frac{s^2}{n^2}$, hence the probability of this case happening is bounded by $\frac{3s^2}{n^2}$.

(iii) There exists some integer g whose presence in V, using $0, d \in V$, explains exactly one of the events $l \in W$ or $l - d \in W$. There are two possibilities as g can be $\{l - d, l + d\}$ (possibilities involving $l - d$ can happen only for $l > 2d$, there is only one possibility for $l < 2d$), and hence the probability of this happening is less than or equal to $\frac{2s}{n}$. Consider the possibility where $l + d \in V$ (happens with probability at most $\frac{s}{n}$ and the event $l - d \in W$ has to be explained). This can happen if $2l \in V$ or $2d \in V$ as they are separated from $l + d$ by $l - d$ (the probability of this happening is bounded by $\frac{2s}{n}$) or there exists an integer h such that $h, h + l \in V$, where both $\{h, h + l\}$ are distinct from $\{0, d, l + d\}$ (h can be chosen in at most n different ways and for each h, the probability is bounded by $\frac{s^2}{n^2}$, the probability of this happening can hence be bounded by $\frac{s^2}{n}$). The same arguments hold for the $l - d$ case too. Hence, the probability of this case happening is upper bounded by $2 \times \left(\frac{s}{n}\right)\left(\frac{s^2}{n} + \frac{2s}{n}\right) = \frac{2s^3}{n^2} + \frac{4s^2}{n^2}$.

(iv) Both the events $l \in W$ and $l - d \in W$ are explained by integers in V not involving $0, d \in V$. This can happen in two ways:

(a) There exists integers g and h such that $g, g + l, h, h + l - d$ are distinct and belong to V. In this case, g can be chosen in at most n different ways and for each g, the probability of $g, g + l \in V$ is bounded by $\frac{s^2}{n^2}$. Similarly, h can be chosen in at most n different ways and for each h, the probability of $h, h + l - d \in V$ is bounded by $\frac{s^2}{n^2}$. The probability of this case is hence upper bounded by $n^2 \times \frac{s^4}{n^4} = \frac{s^4}{n^2}$.

(b) There exists integers g and h such that $\{g, g + l, h, h + l - d\}$ belong to V and only three of them are distinct (there is an overlap). This overlap can happen in four ways: $g = h$, $g + l = h$, $g = h + l - d$ or $g + l = h + l - d$. g can be chosen in n different ways as in the previous case, and for each g, the probability of $g, g + l \in V$

is bounded by $\frac{s^2}{n^2}$. However, for each g, only 4 choices of h are valid as there are four ways of overlap. Also, the probability of $h, h+l-d \in V$ conditioned on $g, g+l \in V$ is bounded by $\frac{s}{n}$ as one of $\{h, h+l-d\}$ already belongs to V due to overlap, and the other can belong to V with probability at most $\frac{s}{n}$ due to independence. The probability of this case is hence upper bounded by $4 \times n \times \frac{s^3}{n^3} = \frac{4s^3}{n^2}$. Note that the overlap requirement has reduced the choice of h from n to 4 and increased the bound on the probability of $h, h+l-d \in V$ from $\frac{s^2}{n^2}$ to $\frac{s}{n}$.

Case II: $d = \frac{l}{2}$

In this case, the event $d \in W$ is already explained by $0, d \in V$ and hence only $2d \in W$ has to be explained. This can happen due to one of the following cases:

(i) There exists some integer g whose presence in V, using $0, d \in V$, can explain $2d \in W$. $g = \{2d, 3d\}$ are the two possibilities, hence the probability of this case is upper bounded by $2 \times \frac{s}{n} = \frac{2s}{n}$.

(ii) The event $2d \in W$ is explained by integers not involving $0, d \in V$. This can happen when there is an integer g such that $g, g+2d \in V$. As earlier, the probability of this event can be bounded by $\frac{s^2}{n}$ as g can take at most n distinct values and for each value of g, the probability is less than or equal to $\frac{s^2}{n^2}$.

$Pr\{l, l-d \in W | v_{01} = d\}$ can be upper bounded, by summing all the aforementioned probabilities. For $d \neq \frac{l}{2}$, we have the bound $\frac{s^4}{n^2} + \frac{6s^3}{n^2} + \frac{7s^2}{n^2} + Pr\{l \in V\}$. For $d = \frac{l}{2}$, similarly, we have the upper bound $\frac{s^2}{n} + \frac{2s}{n}$. Since $Pr\{v_{01} = l/2\} \leq \frac{s}{n}$ and $\sum_{d \neq l/2} Pr\{v_{01} = d\} \leq 1$, we have

$$Pr\{l \in W \cap W_1\} \leq \frac{c_2 s^4}{n^2} + Pr\{l \in V\}$$

for some constant c_2. By using the same arguments as (8.6), we get

$$Pr\{l \in W \cap W_1 | l \notin V\} = O\left(\frac{s^4}{n^2}\right).$$

□

Lemma 8.2.3. *$V = 0 \cup (W \cap (W + v_{01}))$ holds with probability at least $1 - O(n^{-0.1\epsilon})$ when $0 < \theta \leq \frac{1}{5}$.*

Proof. Since all nonzero $l \in V$ also belong to $(W \cap (W + v_{01}))$ by construction (Intersection Step), it suffices to bound the probability that some $l \notin V$ belongs to $(W \cap (W + v_{01}))$.

Let T be a random variable defined as the number of integers, that do not belong to V, that belong to the set $(W \cap (W + v_{01}))$. $Pr\{T \geq 1\}$ can be bounded as follows:

$$E[T] = \sum_{l} Pr\{l \in (W \cap (W + v_{01}))|l \notin V\}.$$

From Lemma 8.2.2, we have

$$E[T] \leq n \times O\left(\frac{s^4}{n^2}\right) = O(n^{-0.2})$$

when $s \leq n^{1/5}$. Using Markov inequality, we get

$$Pr\{T \geq 1\} \leq \frac{E[T]}{1} = O(n^{-0.2})$$

and hence T is 0 with probability at least $1 - O(n^{-0.1\epsilon})$. □

Lemma 8.2.4 (Multiple Intersection Step)**.** *The probability that an integer $l > v_{0t}$, which does not belong to V, belongs to $\left(\bigcap_{p=0}^{t}(W + v_{0p})\right)$ is bounded by $O\left(s\left(\frac{s^2}{n}\right)^{\frac{1}{4}t^{\frac{1}{2}}}\right)$ for $t = \sqrt[3]{\log(s)}$.*

Proof. This lemma, which takes into account multiple intersections, is a generalization of Lemma 8.2.2. The bounds derived in this lemma are very loose, but sufficient for the proof of Theorem 3.3.2.

As in Lemma 8.2.2, we have $Pr\{l \in \left(\bigcap_{p=0}^{t}(W + v_{0p})\right)\}$

$$= \sum_{d_1,d_2\ldots,d_t} (Pr\{v_{0p} = d_p : 0 \leq p \leq t\}\times$$

$$Pr\{\ \{l - d_p : 0 \leq p \leq t\}\ \in W|v_{0p} = d_p : 0 \leq p \leq t\}),$$

where $d_0 = 0$. Note that the integers $\{d_p : 0 \leq p \leq t\}$ have unique pairwise distances with probability at least $1 - O(n^{-\frac{1}{4}})$. This can be seen as follows: For some $\{i_1, j_1\}$ and $\{i_2, j_2\}$ (without loss of generality $j_2 > j_1$) (i) If $i_2 > j_1$ (the intervals do not overlap), $Pr\{d_{i_2j_2} = d_{i_1j_1}\} \leq \frac{s}{n}$ due to independence. (ii) If $i_2 < j_1$, $d_{i_2j_2} = d_{i_1j_1}$ can equivalently be written as $d_{j_1j_2} = d_{i_1i_2}$ which involves non-overlapping intervals. Hence the probability can still be bounded by $\frac{s}{n}$. Since there are $(t + 1)^4$ ways of choosing $\{i_1, j_1, i_2, j_2\}$, the probability that the pairwise distances of $\{d_p : 0 \leq p \leq t\}$ are not distinct can be upper bounded by $\frac{(t+1)^4 s}{n} = O(n^{-\frac{1}{4}})$. Since this probability is less than $O(n^{-\frac{\epsilon}{5}})$, which is the error probability we are aiming for, the pairwise

distances of $\{d_p : 0 \leq p \leq t\}$ are assumed to be distinct in the rest of the proof without loss of generality.

We will bound the probability with which the $t+1$ events $\{l-d_p : 0 \leq p \leq t\} \in W$ can happen, conditioned on $\{v_{0p} = d_p : 0 \leq p \leq t\}$, or equivalently, $0, d_1, d_2, \ldots, d_t \in V$, no other $\{0 \leq i \leq d_t\}$ belong to V. For $i > d_t$, $i \in V$ happens with probability less than or equal to $\frac{s}{n}$ independently.

Since $0, d_1, d_2, \ldots, d_t \in V$, they can explain some of the $t+1$ events due to pairwise distances among themselves. These integers cannot explain more than $\frac{t+1}{2}$ events due to pairwise distances among themselves, which can be seen as follows: Suppose there exists a $0 \leq p \leq t$ such that $d_p - d_{i_1} = l - d_{j_1}$ and $d_p - d_{i_2} = l - d_{j_2}$ for some $\{i_1, j_1, i_2, j_2\}$ (where i_1 and i_2 are distinct), by subtracting, we get $d_{i_2} - d_{i_1} = d_{j_2} - d_{j_1}$, which is a contradiction. Hence, for each d_p, there can be at most one i such that $d_p - d_i$ can explain one of the $t+1$ events. Consider a graph with $t+1$ nodes such that each term d_p for $0 \leq p \leq t$ corresponds to a node. Draw an edge between two nodes $\{p, i\}$ in this graph if $d_p - d_i$ can explain one of the $t+1$ events. Since no vertex in this graph can have a degree greater than 1, this graph can have at most $\frac{t+1}{2}$ edges, because of which $0, d_1, d_2, \ldots, d_t \in V$ can explain at most $\frac{t+1}{2}$ events due to pairwise distances among themselves.

Hence, at least $\frac{t+1}{2}$ events must be explained by other integers greater than d_t in V. This can happen due to one of the following cases:

(i) There exists some integer g whose presence in V, using $0, d_1, d_2, \ldots, d_t \in V$, explains at least two of the $t+1$ events $\{l - d_p : 0 \leq p \leq t\} \in W$. $g = l$ is the only integer which comes under this case, which can be seen as follows: If for some g, we have $g - d_{i_1} = l - d_{j_1}$ and $g - d_{i_2} = l - d_{j_2}$ for some $\{i_1, j_1, i_2, j_2\}$, then by subtracting, we get $d_{i_1} - d_{i_2} = d_{j_1} - d_{j_2}$ which is a contradiction unless $i_1 = i_2$ and $j_1 = j_2$. Hence, $l \in V$ is the only possibility, the probability of this case is given by $Pr\{l \in V\}$.

Let $\mathcal{G}_1$ be the set of integers g whose presence in V, using only integers from $0, d_1, d_2, \ldots, d_t \in V$, can explain exactly one of the $t+1$ events. The size of this set is less than or equal to $(t+1)^2$: For any g to belong to this set, it has to be a distance $l - d_j$ away from some integer d_i, where $0 \leq i, j \leq t$. Hence, there can be at most $(t+1) \times (t+1)$ such integers.

(ii) Consider the case where at least $\frac{t+1}{4}$ of the events are explained by integer pairs in V such that one integer is in $\mathcal{G}_1$ and the other is in $\{0, d_1, d_2, \ldots, d_t\}$. Since the

number of ways in which c integers in $\mathcal{G}_1$ can be chosen is bounded by $(t+1)^{2c}$, the probability of this case is bounded by

$$\sum_{c=\frac{t+1}{4}}^{(t+1)^2} (t+1)^{2c} \left(\frac{s}{n}\right)^c \leq (t+1)^{2+2(t+1)^2} \left(\frac{s}{n}\right)^{\frac{t+1}{4}},$$

as each term involved in the summation can be bounded by $(t+1)^{2(t+1)^2} \left(\frac{s}{n}\right)^{\frac{t+1}{4}}$. More integers might be required to be present in V to explain all the events, which might decrease the probability of this case further. However, this bound is sufficient. Since $t = \sqrt[3]{\log(s)}$, we can write $(t+1)^{2+2(t+1)^2} = O(s)$. The probability of this case is hence bounded by $O(s \left(\frac{s}{n}\right)^{\frac{t+1}{4}})$.

If less than $\frac{t+1}{4}$ events are explained by integer pairs in V such that one integer is in $\mathcal{G}_1$ and the other is in $\{0, d_1, d_2, \ldots, d_t\}$: Since the integers in $\mathcal{G}_1$ can explain at most $\frac{t+1}{4}$ events using $0, d_1, d_2, \ldots, d_t \in V$, at least $\frac{t+1}{4}$ events must be explained by integer pairs in V such that both the integers in the pair are greater than v_{0t}.

(iii) At least $\frac{t+1}{4}$ events are explained by pairs of integers not involving $\{0 \leq i \leq d_t\} \in V$. This can happen in two ways:

(a) There exist integers $\{g_1, g_2, \ldots, g_{\frac{t+1}{4}}\}$ such that $\{g_1, g_1 + l - d_{p_1}, g_2, g_2 + l - d_{p_2}, \ldots, g_{\frac{t+1}{4}}, g_{\frac{t+1}{4}} + l - d_{p_{\frac{t+1}{4}}}\}$ are distinct and belong to V. In this case, each g_i can be chosen in n ways and the probability of $g_i, g_i + l - d_{p_i} \in V$ is bounded by $\frac{s^2}{n^2}$. The probability of this case is hence bounded by $\left(\frac{s^2}{n}\right)^{\frac{t+1}{4}}$.

(b) There exist integers $\{g_1, g_2, \ldots, g_{\frac{t+1}{4}}\}$ such that $\{g_1, g_1 + l - d_{p_1}, g_2, g_2 + l - d_{p_2}, \ldots, g_{\frac{t+1}{4}}, g_{\frac{t+1}{4}} + l - d_{p_{\frac{t+1}{4}}}\}$ are not distinct. The following steps are a generalization of this case in Lemma 8.2.2: Consider a graph of $\frac{t+1}{4}$ vertices where each node corresponds to a pair $\{g_i, g_i + l - d_{p_i}\}$. An edge is drawn between vertices $\{i, j\}$ if $\{g_i, g_i + l - d_{p_i}\}$ and $\{g_j, g_j + l - d_{p_j}\}$ overlap, i.e., have an integer in common. This can happen due to 4 different cases, as in Lemma 8.2.2. Hence, between each pair $\{i, j\}$, there are at most 5 possibilities, which bounds the total number of possibilities by 5^{t^2}.

For this graph, the following can be said: (i) The number of distinct integers in $\{g_1, g_1 + l - d_{p_1}, g_2, g_2 + l - d_{p_2}, \ldots, g_{\frac{t+1}{4}}, g_{\frac{t+1}{4}} + l - d_{p_{\frac{t+1}{4}}}\}$, say c, must be at least $\frac{1}{2}t^{\frac{1}{2}}$. (ii) The number of forests in the graph is less than or equal to $\frac{c}{2}$ (as each forest must have at least two distinct integers).

Since the number of g_i which can be chosen in n different ways is equal to the number of forests in the graph and the rest of the g_i get fixed due to overlap, the probability of this case can be bounded by

$$\sum_{c=\frac{1}{2}t^{\frac{1}{2}}}^{2t} 5^{t^2} n^{\frac{c}{2}} \left(\frac{s}{n}\right)^c \leq 2t(5^{t^2}) \left(\frac{s^2}{n}\right)^{\frac{1}{4}t^{\frac{1}{2}}}.$$

Since $2t(5^{t^2}) = O(s)$, the probability of this case can be bounded by $O(s\left(\frac{s^2}{n}\right)^{\frac{1}{4}t^{\frac{1}{2}}})$.

Since the expressions are independent of $\{d_{0p} : 1 \leq p \leq t\}$, we have the following bound for $Pr\{l \in \left(\cap_{p=0}^{t}(W + v_{0p})\right)\}$:

$$Pr\{l \in V\} + O\left(s\left(\frac{s}{n}\right)^{\frac{t+1}{4}} + s\left(\frac{s^2}{n}\right)^{\frac{1}{4}t^{\frac{1}{2}}} + \left(\frac{s^2}{n}\right)^{\frac{t+1}{4}}\right),$$

which can be simplified as $O\left(s\left(\frac{s^2}{n}\right)^{\frac{1}{4}t^{\frac{1}{2}}}\right) + Pr\{l \in V\}$.

Conditioning on $l \notin V$, using the same argument as (8.6), we have the following bound:

$$Pr\{l \in \left(\cap_{p=0}^{t}(W + v_{0p})\right) \mid l \notin V\} = O\left(s\left(\frac{s^2}{n}\right)^{\frac{1}{4}t^{\frac{1}{2}}}\right).$$

□

Lemma 8.2.5. *$V = \{v_{00}, v_{01}, \ldots, v_{0,t-1}\} \cup \left(\cap_{p=0}^{t}(W + v_{0p})\right)$ holds with probability at least $1 - O(n^{-0.1\epsilon})$ when $t = \sqrt[3]{\log(s)}$ and $0 < \theta \leq \frac{1}{2} - \epsilon$.*

Proof. The proof is identical to Lemma 8.2.3. Let T be a random variable defined as the number of integers, that do not belong to V, that belong to the set $\left(\cap_{p=0}^{t}(W + v_{0p})\right)$. $Pr\{T \geq 1\}$ can be bounded as follows:

$$E[T] = \sum_{l} Pr\{l \in \left(\cap_{p=0}^{t}(W + v_{0p})\right) | l \notin V\}$$
$$= O\left(ns\left(\frac{s^2}{n}\right)^{\frac{1}{4}t^{\frac{1}{2}}}\right) \in O\left(ns\left(\frac{s^2}{n}\right)^{\frac{1}{4}\sqrt[6]{\log(n^\epsilon)}}\right).$$

This can be further upper bounded by $O(n^{-0.1\epsilon})$, since the expression is of the form $O(n^{\frac{3}{2}-\frac{\epsilon}{2}\sqrt[6]{\epsilon\log(n)}})$. Using Markov inequality, we get

$$Pr\{T \geq 1\} \leq \frac{E[T]}{1} = O(n^{-0.1\epsilon}),$$

which completes the proof. □

Lemma 8.2.6 (Graph Step)**.** *In the graph $G(\{0\} \cup (W \cap (W + v_{01})), W)$, integers $\{v_{0p} : 1 \leq p \leq t = \sqrt[3]{\log(s)}\}$ have an edge with $v_{0,k-1}$ with probability at least $1 - O(n^{-0.1\epsilon})$ when $\frac{1}{5} \leq \theta \leq \frac{1}{2} - \epsilon$.*

Proof. For any p such that $1 \leq p \leq t$, the terms v_{0p} and $v_{0,k-1}$ have a difference $v_{p,k-1}$. For there to be no edge between v_{0p} and $v_{0,k-1}$, another integer pair in $\{0\} \cup (W \cap (W + v_{01}))$ should have the same difference. For this to happen, at least one of the integers in this integer pair should be greater than $v_{p,k-1}$. The only integers greater than $v_{p,k-1}$ in W can be terms of the form $\{v_{ij} : 0 \leq i \leq p-1, j > i\}$. These terms can be split into two cases:

(i) $j \leq k - s^{\frac{\epsilon}{2}}$: Note that $Pr\{v_{0t} > v_{k-s^{\frac{\epsilon}{2}},k-1}\} = O(s^{-\frac{\epsilon}{2}})$ if $t = \sqrt[3]{\log(s)}$. This can be shown as follows: v_{0t} concentrates around its mean $\frac{tn}{s}$ with a variance bounded by $\frac{2tn^2}{s^2}$. $v_{k-s^{\frac{\epsilon}{2}},k-1}$ concentrates around its mean $\frac{s^{\frac{\epsilon}{2}}n}{s}$ with a variance bounded by $\frac{2s^{\frac{\epsilon}{2}}n^2}{s^2}$. Chebyshev's inequality completes the proof. Using this, we see that

$$v_{ij} \leq v_{0t} + v_{pj} < v_{k-s^{\frac{\epsilon}{2}},k-1} + v_{p,k-s^{\frac{\epsilon}{2}}} = v_{p,k-1}$$

with probability at least $1 - O(s^{-\frac{\epsilon}{2}})$. Hence, with probability $O(n^{-0.1\epsilon})$, one or more of these terms can be the greater term in an integer pair which can produce a difference $v_{p,k-1}$.

(ii) $k - s^{\frac{\epsilon}{2}} < j$: There are at most $ts^{\frac{\epsilon}{2}}$ such terms and p can be chosen in t different ways. For each of these terms and each choice of p, $v_{ij} - v_{p,k-1} = v_{ip} - v_{j,k-1}$ can belong to $W \cap (W + v_{01})$ with a probability at most $O\left((t^2 + s^\epsilon)(\frac{1}{s} + \frac{s^2}{n})\right)$ (Lemma 8.2.7 and Corollary 8.2.1), hence the probability that at least one of these terms will belong to $W \cap (W + v_{01})$ can be union bounded by multiplying this probability by $t^2 s^{\frac{\epsilon}{2}}$. This probability is therefore bounded by $O\left(\frac{s^{1.6\epsilon}}{s} + \frac{s^{2+1.6\epsilon}}{n}\right)$ using $t^2 = O(s^{0.1\epsilon})$. Since $n^{\frac{1}{5}} < s \leq n^{\frac{1}{2}-\epsilon}$, this is simplified as $O\left(n^{-\frac{1}{5}(1-1.6\epsilon)} + n^{(1+0.8\epsilon)(1-2\epsilon)-1}\right) = O(n^{-0.1\epsilon})$ due to $\epsilon \leq 0.3$.

Hence, with probability at least $1 - O(n^{-0.1\epsilon})$, there will be no other integer pair in $\{0\} \cup (W \cap (W + v_{01}))$ with a difference $v_{p,k-1}$ for each $1 \leq p \leq t$, because of which there will be an edge between v_{0p} and $v_{0,k-1}$ for each $1 \leq p \leq t$. □

Lemma 8.2.7. *The probability that* $\{v_{0p} - v_{k-1-q,k-1}\} \in W$*, for any* $0 < p, q < \frac{s}{4}$*, is bounded by* $O\left((p^2+q^2)\left(\frac{1}{s}+\frac{s^2}{n}\right)\right)$.

Proof. We can write $Pr\{\{v_{0p} - v_{k-1-q,k-1}\} \in W\}$ as

$$\sum_{d_1,d_2,l} Pr\{\{v_{0p}, v_{k-1-q,k-1}, v_{0,k-1}\} = \{d_1, d_2, l\}\}$$
$$\times Pr\{d_1 - d_2 \in W | v_{0p} = d_1, v_{k-1-q,k-1} = d_2, v_{0,k-1} = l\}.$$

The distribution of V, conditioned by $v_{0p} = d_1, v_{k-1-q,k-1} = d_2$, and $v_{0,k-1} = l$ is as follows (note that we are not conditioning on the exact value of k as $k \geq \frac{s}{2}$ is sufficient for this proof): $v_{0p} = d_1$ ensures that there are $p-1$ integers in between 1 and $d_1 - 1$ (call this region $\mathcal{R}_1$), the $p-1$ elements will be uniformly distributed in $\mathcal{R}_1$. Similarly, $v_{k-1-q,k-1} = d_2$ ensures that there will be $q-1$ integers uniformly distributed in the range $l - d_2 + 1$ to $l - 1$ (call this region $\mathcal{R}_3$). Since we have not fixed k to any particular value, the probability that an i in the range $d_1 + 1$ to $l - d_2 - 1$ (call this region $\mathcal{R}_2$) will belong to V can be bounded using an independent $Bern(\frac{cs}{n})$ distribution, for some constant c.

For $d_1 - d_2$ to belong to W, there must be a pair of integers $g, g + d_1 - d_2 \in V$. This can happen in the following ways:

(i) If both $g, g + d_1 - d_2 \in V$ are in (a) $\mathcal{R}_2$: the probability of this happening (using arguments similar to Lemma 8.2.1) can be upper bounded by $\frac{c^2s^2}{n^2} \times l \leq \frac{c^2s^2}{n}$. (b) $\mathcal{R}_1$: the probability of this is bounded by $\frac{\binom{p-1}{2}}{\binom{d_1-1}{2}} \times (d_1 - 2) \leq \frac{(p-1)^2}{(d_1-1)}$. (c) $\mathcal{R}_3$: this probability is, similarly, bounded by $\frac{(q-1)^2}{(d_2-1)}$.

(ii) If $g, g + d_1 - d_2 \in V$ are such that (a) one of them is in $\mathcal{R}_1$ and the other is in $\mathcal{R}_2$: the probability of this is bounded by $\frac{(p-1)}{(d_1-1)} \times \frac{cs}{n} \times (d_1 - 1)$ which can be upper bounded by $\frac{pcs}{n}$. (b) one of them is in $\mathcal{R}_2$ and the other is in $\mathcal{R}_3$: this probability is similarly upper bounded by $\frac{qcs}{n}$. (c) one of them is in $\mathcal{R}_1$ and the other is in $\mathcal{R}_3$: the probability of this is bounded by $\frac{(q-1)}{(d_2-1)}\frac{(p-1)}{(d_1-1)} \times (d_2 - 1)$ or $\frac{(q-1)}{(d_2-1)}\frac{(p-1)}{(d_1-1)} \times (d_1 - 1)$.

(iii) If one of g or $g + d_1 - d_2$ is in $\{0, d_1, l - d_2, l\}$: the other can be chosen in at most six ways, this probability can be upper bounded by $O(\frac{s}{n} + \frac{p-1}{d_1-1} + \frac{q-1}{d_2-1})$.

The summation of the probabilities can be bounded by $O(\frac{s^2}{n} + \frac{(p-1)^2}{d_1-1} + \frac{(q-1)^2}{d_2-1})$. The term $\frac{s^2}{n}$ doesn't depend on $\{d_1, d_2, l\}$ and since $\sum_{d_1,d_2,l} Pr\{\{v_{0p}, v_{k-1-q,k-1}, v_{k-1}\} = d_1, d_2, l\} \leq 1$, this bound remains the same after the summation. The term $\frac{(p-1)^2}{d_1-1}$

depends on d_1 and the summation can be bounded as follows:

$$\sum_{2\leq d_1\leq \frac{n}{s^2}} Pr\{v_{0p} = d_1\}\frac{1}{d_1-1} + \sum_{d_1>\frac{n}{s^2}} Pr\{v_{0p} = d_1\}\frac{1}{d_1-1}.$$

In the first sum, $Pr\{v_{01} = d_1\}$ can be bounded by $O(\frac{s}{n})$ and $\frac{1}{d_1-1}$ can be bounded by 1. In the second sum, $\frac{1}{d_1-1}$ can be bounded by $\frac{s^2}{n}$ and $\sum_{d_1>\frac{n}{s^2}} Pr\{v_{0p} = d_1\}$ can be bounded by 1, to bound the total summation by $p^2O\left(\frac{1}{s} + \frac{s^2}{n}\right)$. Similarly, the term involving d_2 can be bounded by $q^2O\left(\frac{1}{s} + \frac{s^2}{n}\right)$.

(iv) Both g and $g + d_1 - d_2$ are in $\{0, d_1, l - d_2, l\}$. This can happen only when: $l = 2d_1$ or $l = 2d_1 - d_2$ or $d_1 = 2d_2$. The probability of each of these happening is bounded by $O(\frac{s}{n})$.

Hence, the total probability can be upper bounded by $O\left((p^2 + q^2)\left(\frac{1}{s} + \frac{s^2}{n}\right)\right)$. □

Corollary 8.2.1. *The probability that* $\{v_{r_1p} \pm v_{k-1-q,k-1-r_2}\} \in W$*, for some* $0 \leq r_1 < p$*,* $0 \leq r_2 < q$ *and any* $0 < p, q < \frac{s}{4}$ *is bounded by* $O\left((p^2 + q^2)\left(\frac{1}{s} + \frac{s^2}{n}\right)\right)$.

Remark: The proof also works for the case when the k locations of the support are chosen uniformly at random, if $k \leq n^{\frac{1}{2}-\epsilon}$. This is due to the fact that all the probability upper bounds derived in this section still hold true up to a constant scaling. For example, the probability that $g, g + l \in V$ for $l > 0$ can be bounded by $\frac{\binom{k}{2}}{\binom{n}{2}} \leq \left(\frac{k}{n/2}\right)^2 = \left(\frac{2k}{n}\right)^2$. This probability is bounded by $\left(\frac{s}{n}\right)^2$ in the $\mathcal{BN}(n, \theta)$ setting. Even though the events $i \in V$ are no longer independent, the bounds will be identical up to a constant scaling.

8.3 Proof of Theorem 3.3.3

The lemmas in this section assume that the sparse signal is drawn from the $\mathcal{BN}(n, \theta)$ distribution, where the parameter θ satisfies $0 < \theta \leq \frac{1}{2} - \epsilon$ for some positive constant ϵ. The events in this section are conditioned with respect to a fixed $k \geq \log^6(s)$. Consequently, the probability that $i \in V$ is bounded by $O(\frac{k}{n})$, and the probability that $i, j \in V$ is bounded by $O(\frac{k^2}{n^2})$ (see Remark at the end of the proof of Theorem 3.3.2).

Consider the following matrix completion problem: Let $\mathbf{R}_0 = \mathbf{r}\mathbf{r}^\star$ be a positive semidefinite $t \times t$ matrix with all the off-diagonal components known, where $\mathbf{r} = (r_0, r_1, \ldots, r_{t-1})$ is a $t \times 1$ vector. The objective is to recover the diagonal components

(robustly) by solving a convex program. Since $\mathbf{R}$ is positive semidefinite, any 2×2 submatrix of $\mathbf{R}$ is also positive semidefinite. Consider the convex program

$$\begin{aligned} &\text{minimize} && \text{trace}(\mathbf{R}) \\ &\text{subject to} && R_{ii}R_{jj} \geq (|r_i||r_j|)^2 \quad \forall\ i \neq j \\ & && R_{ii} \geq 0 \quad \forall\, i. \end{aligned} \tag{8.7}$$

Lemma 8.3.1. $\mathbf{R}_0 = \mathbf{r}\mathbf{r}^\star$ *is the unique optimizer of (8.7) with probability at least* $1 - O(te^{-\frac{t}{4}})$.

Proof. Suppose $\mathbf{R}_0 = \mathbf{r}\mathbf{r}^\star$ is not the unique optimizer of (8.7). If $\mathbf{R}^\dagger \neq \mathbf{R}_0$ is the optimizer, then there exists at least one i such that $R^\dagger_{ii} < |r_i|^2$. For this i, $R^\dagger_{ii}$ can then be expressed as $(1-\gamma)|r_i|^2$ for some $\gamma > 0$. The constraints of (8.7) corresponding to R_{ii} (i.e., $R_{ii}R_{jj} \geq (|r_i||r_j|)^2$ for all $j \neq i$) will ensure that all other diagonal components $R_{jj}, j \neq i$ be greater than or equal to $\frac{1}{1-\gamma}|r_j|^2$, which also implies that R_{jj} is greater than $(1+\gamma)|r_j|^2$ (as $\frac{1}{1-\gamma} > 1+\gamma$) . The objective function value at the optimum can be written as

$$\text{trace}(\mathbf{R}^\dagger) = \sum_{i=1}^{i=t} R^\dagger_{ii} > \sum_j |r_j|^2 + \gamma(\sum_{j\neq i} |r_j|^2 - |r_i|^2).$$

If we can ensure that $(\sum_{j\neq i} |r_j|^2 - |r_i|^2) > 0$ for all i, we are through because $\text{trace}(\mathbf{R}^\dagger)$ is greater than $\sum_j r_j^2$, which is a contradiction. Since the signal values in the support are chosen from an i.i.d. standard complex normal distribution, the quantity $\sqrt{\sum_{j\neq i} |r_j|^2}$ concentrates around $\sqrt{t-1}$ (see the proof of Lemma 5.33 in [Ver10]), the probability that $|r_i| \geq \sqrt{\frac{t-1}{2}}$ is bounded by $O(e^{-\frac{t}{4}})$. Union bounding over all i, we obtain the bound $O(te^{-\frac{t}{4}})$. □

Lemma 8.3.2. *The probability that there exists an edge between any two particular vertices in* $H(U)$ *is at least* $1 - O\left(\frac{k^2}{n}\right)$.

Proof. Consider any pair of integers $\{i, j\}$. There will be no edge between their corresponding vertices if there exists another pair of integers $g, g+j-i \in V$. For any particular g such that $\{g, g+j-i\}$ are distinct from both i and j, $g, g+j-i \in V$ happens with probability at most $O\left(\frac{k^2}{n^2}\right)$. Since g can be chosen in at most n distinct ways, this probability can be bounded by $O(\frac{k^2}{n})$. If one of $\{g, g+j-i\}$ is equal to i or j, then there are two possibilities and the probability of each of the possibilities can be bounded by $O(\frac{k}{n})$. Hence, the probability that there is no edge between any two particular vertices can be bounded by $O(\frac{k^2}{n} + \frac{2k}{n}) = O(\frac{k^2}{n})$. □

Lemma 8.3.3. *Suppose $d_{min}(H(U))$ denotes the minimum degree of the graph $H(U)$, then $d_{min}(H(U)) \geq k(1-1/t)$ where $t = \log^2(s)$, with probability at least $1 - O(n^{-1})$. Such a graph also has a Hamiltonian cycle.*

Proof. Consider a vertex u_i. Construct a graph H_i from $H(U)$ by removing all the edges which do not involve the vertex u_i. Let us consider the vertex exposure martingale [AS15; JOH12a] on this graph H_i with the graph function $d(u_i)$, where $d(u)$ denotes the degree of the vertex u. Let F_j be the induced subgraph of H_i formed by exposed vertices after j exposures. We define a martingale $X_0, X_1, \ldots, X_{k-1}$ as follows:

$$X_j = E[d(u_i)|F_j]$$

We have $X_0 = E[d(u_i)] \geq k(1 - O(\frac{k^2}{n}))$ and $X_{k-1} = d(u_i)$. Note that $|X_{j+1} - X_j| \leq 1 \quad \forall \quad 0 \leq j \leq k-2$. Azuma's inequality [AS15] gives us

$$Pr\{d(u_i) < E[d(u_i)] - \lambda\} \leq 2e^{-\lambda^2/2k}$$

for $\lambda > 0$. Choosing $\lambda = k\left(\frac{1}{t} - O(\frac{k^2}{n})\right)$ and $t = \log^2(s)$, we get

$$Pr\{d(u_i) < k\left(1 - \frac{1}{t}\right)\} \leq 2e^{-\frac{k}{2}\left(\frac{1}{t} - O(\frac{k^2}{n})\right)^2}.$$

Using union bound to accommodate all the vertices u_i for $i = \{0, 1, \ldots, k-1\}$, we get

$$Pr\{\exists i : d(u_i) < k\left(1 - \frac{1}{t}\right)\} \leq 2ke^{-\frac{k}{2}\left(\frac{1}{t} - O(\frac{k^2}{n})\right)^2} = O(n^{-1}).$$

Every vertex in this graph has a degree at least $\frac{k}{2}$ (even $t = 2$ is sufficient for this). Dirac's theorem states that such graphs have a Hamiltonian cycle. □

8.4 Proof of Theorem 3.4.1

The lemmas in this section assume that the sparse signal is drawn from the $\mathcal{BN}(n, \theta)$ distribution, where the parameter θ satisfies $0 < \theta \leq \frac{1}{4} - \epsilon$.

Lemma 8.4.1. *The output of the first step is a term of the form $v_{i_0 j_0}$ or $v_{k-1-j_0, k-1-i_0}$, where $0 \leq i_0 < j_0 \leq 2c+1$, with probability at least $1 - O(n^{-4\epsilon})$.*

Proof. Consider the terms of the form $\{v_{0i} : 1 \leq i \leq 2c+1\}$. Since at most c of them belong to W_{del}, at least $c+1$ of them belong to $W^{\dagger}$. Similarly, at least $c+1$ terms of the form $\{v_{i,k-1} : 1 \leq i \leq 2c+1\}$ belong to $W^{\dagger}$. Hence, there exists at least

one integer (denote the minimum of them by l_1) which satisfies $1 \leq l_1 \leq 2c+1$ and $v_{0l_1}, v_{l_1,k-1} \in W^\dagger$.

Since $v_{0,k-1} \in W^\dagger$, we have $v_{l_1,k-1}, v_{0,k-1} \in T^\dagger_{sub}$ as both the conditions are satisfied:

(i) They have a difference v_{0l_1}, which belongs to $W^\dagger$.

(ii) The integer pairs of the form $\{v_{0i}, v_{l_1 i}\}$ for $2c+2 \leq i \leq k-1$ have a difference v_{0l_1}, and since at most $2c$ of the terms involved belong to W_{del}, at least $\frac{\sqrt{K^\dagger}}{4}$ such pairs belong to $W^\dagger$.

Similarly, we have $v_{0,k-1-l_2}, v_{0,k-1} \in T^\dagger_{sub}$ for some $1 \leq l_2 \leq 2c+1$. Hence, the first step chooses a value of $w^\dagger_{min}$ which is at least $\max\{v_{l_1,k-1}, v_{0,k-1-l_2}\}$, which results in a value of v_{0j_0} or $v_{k-1-j_0,k-1}$ for some $0 < j_0 \leq 2c+1$.

If $w^\dagger_{min}$ is a value higher than $\max\{v_{l_1,k-1}, v_{0,k-1-l_2}\}$, one of the following two cases must happen:

(i) $w^\dagger_{min} = v_{ij}$ for some $0 \leq i \leq 2c+1$ and $k-1-(2c+1) \leq j \leq k-1$: For each such v_{ij}, this can happen in two ways: (a) The integer pair involving $w^\dagger_{min}$ which satisfies both the conditions contains another (strictly greater) term of the form $v_{i'j'}$ which belongs to W. This can happen only if $v_{i'j'} - v_{ij} \in W$ or $v_{i'j'} - v_{ij} \in W_{ins}$ for some $v_{i'j'} \in W$. If $i' = i$ or $j' = j$, the resulting value is either $v_{jj'}$ or $v_{i'i}$ respectively, which is within the requirements of this step. If $i' \neq i$ and $j' \neq j$, the probability of $v_{i'j'} - v_{ij} \in W$ can be bounded, using Corollary 8.2.1, by $c_0\left(\frac{1}{s} + \frac{s^2}{n}\right)$ for some constant c_0 and the probability of $v_{i'j'} - v_{ij} \in W_{ins}$ can be bounded by $p \leq \frac{s^2}{n}$ due to the independence of W_{ins} and W. The total number of ways in which $\{i', j'\}$ can be chosen is bounded by a constant. (b) The integer pair involving $w^\dagger_{min}$ which satisfies both the conditions contains a (strictly greater) term g which belongs to W_{ins}. This can happen if $g - v_{ij} \in W$ or $g - v_{ij} \in W_{ins}$ for some $g \in W_{ins}$. The event $g - v_{ij} \in W$ is equivalent to the event $v_{i'j'} + v_{ij} \in W_{ins}$ for some $\{i', j'\}$, the probability of which can be bounded by $O(\frac{s^4}{n})$ as the probability for each $\{i', j'\}$ is bounded by $p \leq \frac{s^2}{n}$ due to independence and $\{i', j'\}$ can have at most $O(s^2)$ different values. The probability of $g - v_{ij} \in W_{ins}$ for a particular $\{i, j\}$ can be bounded as follows: Two integers in W_{ins} must be separated by v_{ij}, i.e., $g, g - v_{ij} \in W_{ins}$. This can be bounded by $p^2 n \leq \frac{s^4}{n}$ (using the same arguments as Lemma 8.2.1).

Since the total number of ways in which $\{i, j\}$ can be chosen is bounded by a constant, the probability of this case happening can be bounded by $O(n^{-4\epsilon})$ due to $s \leq n^{\frac{1}{4}-\epsilon}$.

(ii) $w^\dagger_{min} = g$ for some $g \in W_{ins}$: For each such g, this can happen in two ways:

(a) The integer pair involving g which satisfies both the conditions contains a (strictly greater) term of the form $v_{i'j'}$ which belongs to W. This can happen only if $v_{i'j'} - g \in W$ or $v_{i'j'} - g \in W_{ins}$ for some $v_{i'j'} \in W$. The event $v_{i'j'} - g \in W$ is equivalent to $v_{i'j'} - v_{i''j''} = g$ for some $\{i'', j''\}$. This probability is bounded by $O(\frac{s^3}{n})$ as the probability is bounded by $\frac{s}{n}$ for each $\{i'', j''\}$ and the total number of ways in which $\{i'', j''\}$ can be chosen is bounded by $O(s^2)$, and the total number of ways in which $\{i', j'\}$ can be chosen is bounded by a constant. The probability of $v_{i'j'} - g \in W_{ins}$ can be bounded by $p \leq \frac{s^2}{n}$ for every $\{i', j'\}$ and the total number of ways in which $\{i', j'\}$ can be chosen is bounded by a constant. (b) The integer pair involving g which satisfies both the conditions contains another (strictly greater) term g' which belongs to W_{ins}. For each such g', the probability of $g' - g \in W$ can be bounded by $\frac{2s^2}{n}$ (Lemma 8.2.1) and the probability of $g' - g \in W_{ins}$ can be bounded by $p \leq \frac{s^2}{n}$ due to independence. The number of such g' in W_{ins} can be calculated as follows: Since g' has to be greater than $\max\{v_{l_1,k-1}, v_{0,k-1-l_2}\}$, the range of values it can take is limited by $\min\{v_{0,l_1}, v_{k-1-l_2,k-1}\}$. Hence, the expected number of such g' is less than or equal to $(2c+1)\frac{n}{s}p = (2c+1)o(s)$. Hence, the number of such g' is $O(s)$ (Markov inequality). The probability of this event can hence be bounded by $O(\frac{s^3}{n})$.

Since the total number of such g is similarly $O(s)$, the probability of this case happening can be bounded by $O(n^{-4\epsilon})$ when $s \leq n^{\frac{1}{4}-\epsilon}$.

Hence, the output of the first step is $v_{i_0j_0}$ or $v_{k-1-j_0,k-1-i_0}$, where $0 \leq i_0 < j_0 \leq 2c+1$ with the desired probability. □

To resolve the flip ambiguity, we aim to recover the support U such that $u_{i_0j_0}$ is the output of the first step. We will provide the details for the case where $u_{i_0j_0} = v_{i_0j_0}$, the calculations are identical for the case where $u_{i_0j_0} = v_{k-1-j_0,k-1-i_0}$.

Consider the set $W^\dagger \cap (W^\dagger + v_{i_0j_0})$. At least $2c+2$ terms of the form $\{v_{i_0j} : (k-1)-(3c+1) \leq j \leq k-1\}$ belong to $W^\dagger$ and at least $2c+2$ terms of the form $\{v_{j_0j} : (k-1)-(3c+1) \leq j \leq k-1\}$ belong to $W^\dagger$ (which, when added by $v_{i_0j_0}$, gives v_{i_0j}). Hence, at least $c+2$ terms of the form $\{v_{i_0j} : (k-1)-(3c+1) \leq j \leq k-1\}$ belong to $W^\dagger \cap (W^\dagger + v_{i_0j_0})$.

Consider the integers in between $v_{(k-1)-(3c+1)}$ and v_{k-1}. For any integer, not in V, to belong to $W^\dagger \cap (W^\dagger + v_{i_0j_0})$ in this region, one of the following cases has to happen:

(i) The integer has to belong to $W \cap (W + v_{i_0j_0})$: The probability of this happening can be bounded by $\frac{c_0 s^4}{n^2}$ (Lemma 8.4.2). Hence, the probability that some integer

which is not in V, in this region, belongs to $W \cap (W + v_{i_0 j_0})$ is union bounded by $O(\frac{s^4}{n^2}) \times n = O(\frac{s^4}{n})$.

(ii) The integer has to belong to $W_{ins} \cap (W_{ins} + v_{i_0 j_0})$: For this to happen, there must exist two integers, say g_1 and g_2, in W_{ins} which are separated by $v_{i_0 j_0}$. This probability is bounded by $p^2 n \leq \frac{s^4}{n}$.

(iii) The integer has to belong to $W_{ins} \cap (W + v_{i_0 j_0})$ or $W \cap (W_{ins} + v_{i_0 j_0})$: For each $v_{ij} \in W$, the probability of $\{v_{ij} \pm v_{i_0 j_0}\} \in W_{ins}$ can be bounded by $2p \leq \frac{2s^2}{n}$. Therefore, this probability can be bounded by $O(\frac{s^4}{n})$.

Hence, the largest $c + 2$ integers in $W^\dagger \cap (W^\dagger + v_{i_0 j_0})$ correspond to the pairwise distance between v_{i_0} and v_j for some $(k-1) - (3c+1) \leq j \leq k-1$ (denote these integers by $\{v_{q_0}, v_{q_1}, \ldots, v_{q_{c+1}}\}$) with the desired probability.

For every $0 \leq p \leq \frac{k}{2}$, there exist at least two terms, say q_i and q_j such that $v_{pq_i}, v_{pq_j} \in W^\dagger$ and hence $v_{pq_{c+1}}$ will belong to the intersection $\left(W^\dagger \cap (W^\dagger + v_{q_i q_j})\right) + v_{q_j q_{c+1}}$. Hence, by considering intersections with each of the $\binom{c+2}{2}$ pairs $\{v_{q_i}, v_{q_j}\}$ and taking a union of the resulting integer sets, we can ensure that all the terms of the form $v_{pq_{c+1}}$, where $0 \leq p \leq \frac{k}{2}$, belong to the resulting integer set. The probability that some other integer in this range will belong to the integer set can be union bounded by $(c+2)^2$ times the probability calculated above in the case of intersection with $v_{i_0 j_0}$.

Using the fact that $v_{0p} = v_{0q_{c+1}} - v_{pq_{c+1}}$, v_{0p} for $0 \leq p \leq \frac{k}{2}$ can be recovered. Using the first $c+2$ of these terms, by considering intersections with each of the $\binom{c+2}{2}$ pairs $\{v_i, v_j\}$ and taking a union of the resulting integer sets, we can similarly recover all the terms of the form v_{0p}, where $\frac{k}{2} \leq p \leq k-1$.

Lemma 8.4.2. *For any integer $l > v_{0\frac{k}{2}}$ such that l does not belong to V, the probability that $l - v_{i_0}$ belongs to $W \cap (W + v_{i_0 j_0})$, where $0 \leq i_0 < j_0 \leq c$ for some constant c, is bounded by $O(\frac{s^4}{n^2})$.*

Proof. This is a generalization of Lemma 8.2.2, the events are conditioned on $v_{i_0} = d_1$ and $v_{i_0 j_0} = d_2$ instead. The conditional distribution of V is as follows: $v_{i_0} = d_1$ ensures that there are $i_0 - 1$ integers in the range 1 to $d_1 - 1$ (these integers will be uniformly distributed in this range). $v_{i_0 j_0} = d_2$ ensures that there are $j_0 - i_0 - 1$ integers in the range $d_1 + 1$ to $d_1 + d_2 - 1$ (these integers will be uniformly distributed in this range). Any integer greater than $d_1 + d_2$ will belong to V with a probability at most $\frac{s}{n}$ independently.

If $s \leq n^{\frac{1}{4}-\epsilon}$, then $H(U)$ is a clique[1] with probability at least $1 - O(\frac{s^4}{n})$ (Lemma 8.3.2 bounds the probability of each edge missing by $O(\frac{s^2}{n})$, a simple union bound completes the proof), from which we have $d_1 \neq d_2$. Also, if $i_0 \neq 0$, then $d_1 \geq \frac{n}{s^2}$ holds with probability at least $1 - O(\frac{s}{n})$ (see proof of Lemma 8.2.7).

For two events $l - d_1, l - d_1 - d_2 \in W$ to happen, there has to be some g such that $g, g + l - d_1 \in V$ and some h such that $h, h + l - d_1 - d_2 \in V$. Note that $g + l - d_1$ and $g + l - d_1 - d_2$ are greater than $d_1 + d_2$ due to $l \geq v_{0\frac{k}{2}}$ and $i_0 < j_0 \leq c$.

Lemma 8.2.2 provides the bound $O(\frac{s^4}{n^2})$ for all the cases by which the two events can be explained except for four cases, the bounds for which are provided here (see Remark at the end of the proof of Theorem 3.3.2 for relationship between calculations of independent Bernoulli and uniform distributions):

(i) There exists one integer g in the range 1 to $d_1 - 1$ or $d_1 + 1$ to $d_1 + d_2 - 1$, whose presence in V, using another integer in V greater than $d_1 + d_2$, explains exactly one event. Since the probability of $g \in V$ in this range can be bounded by $\frac{i_0-1}{d_1-1}$ or $\frac{j_0-i_0-1}{d_2-1}$ respectively, and the number of ways in which g can be chosen is bounded by $d_1 - 1$ or $d_2 - 1$ respectively, the probability of this happening can be bounded by $\frac{i_0-1}{d_1-1}\frac{s}{n}(d_1 - 1) + \frac{j_0-i_0-1}{d_2-1}\frac{s}{n}(d_2 - 1) = O(\frac{s}{n})$. The probability of this case can be bounded, using the same arguments as that of the third case (under Case I) in Lemma 8.2.2, by $2 \times O(\frac{s}{n}) \times \left(\frac{s^2}{n} + \frac{2s}{n}\right) = O(\frac{s^3}{n^2})$.

(ii) There exists one integer g in the range 1 to $d_1 - 1$ or $d_1 + 1$ to $d_1 + d_2 - 1$, whose presence in V, using two integers in V greater than $d_1 + d_2$, explains both the events. The probability of $g \in V$ can be bounded the same way as in the first case. Hence, the probability of this case can be bounded by $\frac{i_0-1}{d_1-1}\left(\frac{s}{n}\right)^2(d_1 - 1) + \frac{j_0-i_0-1}{d_2-1}\left(\frac{s}{n}\right)^2(d_2 - 1) = O(\frac{s^2}{n^2})$.

(iii) There exist two integers $\{g, h\}$ in the range 1 to d_1-1 or d_1+1 to d_1+d_2-1, whose presence in V, using one another integer in V greater than $d_1 + d_2$, explains both the events. For this to happen, there must exist two integers $g, g + d_2 \in V$ in this range. If both of them are in the range 1 to $d_1 - 1$, the probability of $g, g + d_2 \in V$ can be bounded by $4\left(\frac{i_0-1}{d_1-1}\right)^2$ and the number of ways in which g can be chosen is bounded by $d_1 - 1$. If one of them is in the range 1 to $d_1 - 1$ and the other is in the range $d_1 + 1$ to $d_1 + d_2 - 1$, the probability of $g, g + d_2 \in V$ can be bounded by $\frac{i_0-1}{d_1-1}\frac{j_0-i_0-1}{d_2-1}$ and the number of ways in which g can be chosen is bounded by d_2-1. Hence, the probability of this case is bounded by $4\left(\frac{i_0-1}{d_1-1}\right)^2\frac{s}{n}(d_1 - 1) + \frac{i_0-1}{d_1-1}\frac{j_0-i_0-1}{d_2-1}\frac{s}{n}(d_2 - 1) = O(\frac{s^3}{n^2})$.

[1]The collision-free property in [Ran+13] is equivalent to $H(U)$ being a clique, in our notation.

(iv) There exist two integers $\{g, h\}$ in the range 1 to $d_1 - 1$ or $d_1 + 1$ to $d_1 + d_2 - 1$, whose presence in V, using two other integers in V greater than $d_1 + d_2$, explains both the events. This probability can be bounded by $O\left(\frac{s}{n}\right)^2$, as the probability to explain each event can be bounded by $O(\frac{s}{n})$ using the same arguments as that of the first case. □

In order to analyze the error in the recovered signal values, we use a technique similar to the proof of Theorem 3.3.3. If $s \leq n^{\frac{1}{4}-\epsilon}$, then the graph $H(U)$ is a clique with probability at least $1 - O(\frac{s^4}{n})$ (see proof of Lemma 8.4.2). Hence, we analyze

$$\begin{aligned} &\text{minimize} \quad \text{trace}(\mathbf{X}) \\ &\text{subject to} \quad |X_{u_i u_j} - a_{|u_i - u_j|}| \leq \eta \quad \text{if } u_i \leftrightarrow u_j \text{ in } H(U) \\ &\qquad\qquad X_{ij} = 0 \quad \text{if} \quad i, j \notin U, \quad \mathbf{X} \succcurlyeq 0 \end{aligned} \tag{8.8}$$

as follows: Let $\mathbf{R}_0 = \mathbf{r}\mathbf{r}^\star$ be a $k \times k$ matrix whose off-diagonal components are measured with additive noise, i.e., $Q_{ij} = r_i r_j^\star + z_{ij} \quad \text{for} \quad 0 \leq i \neq j \leq k-1$, where $\mathbf{r} = (r_0, r_1, \ldots, r_{k-1})^T$ and the noise satisfies $|z_{ij}| \leq \eta$. The objective is to recover the diagonal components robustly. Consider the program

$$\begin{aligned} &\text{minimize} \quad \text{trace}(\mathbf{R}) \\ &\text{subject to} \quad |Q_{ij} - R_{ij}| \leq \eta \quad \text{for} \quad 0 \leq i \neq j \leq k-1 \\ &\qquad\qquad \mathbf{R} \succcurlyeq 0. \end{aligned} \tag{8.9}$$

If $\mathbf{R}^\dagger$ is the optimizer of (8.9), for all $0 \leq i \neq j \leq k-1$,

$$|R_{ij}^\dagger - r_i r_j^\star| \leq |R_{ij}^\dagger - Q_{ij}| + |Q_{ij} - r_i r_j^\star| \leq 2\eta.$$

By using AM-GM inequality, we get $|R_{ij}^\dagger|^2 \geq (|r_i|^2 - 2\eta)(|r_j|^2 - 2\eta)$ for all $i \neq j$. Since for all off-diagonal components, we have $|R_{ij}^\dagger|^2 \geq (|r_i|^2 - 2\eta)(|r_j|^2 - 2\eta)$, at most one of the diagonal terms (say i) is such that $R_{ii}^\dagger < (|r_i|^2 - 2\eta)$. If $R_{ii}^\dagger < (|r_i|^2 - k\eta)$, then the 2×2 positive semidefinite constraints would ensure that for all $j \neq i$, $R_{jj}^\dagger > (|r_j|^2 + \alpha_j \eta)$, where $\alpha_j \geq (k-4)\frac{|r_j|^2}{|r_i|^2} - 4$. The optimum value would, similar to the proof of Theorem 3.3.3, strictly increase with the desired probability. Hence, the optimizer has diagonal components $R_{jj}^\dagger \geq |r_j|^2 - 2\eta$ for $0 \leq j \neq i \leq k-1$ and $R_{ii}^\dagger \geq |r_i|^2 - k\eta$.

Since the objective function value at the optimizer is less than or equal to $\sum_j |r_j|^2$, we have the bound $\sum_j \left(R_{jj}^\dagger - |r_j|^2\right)^2 \leq (2\eta)^2(k-1) + (\eta k)^2 + (3\eta k)^2 \leq 12k^2\eta^2$.

Since there are at most k^2 off-diagonal entries and each of them are measured with an error of at most 2η, we have

$$\|\mathbf{X}^\dagger - \mathbf{x}_0\mathbf{x}_0^\star\|_2^2 \leq 12k^2\eta^2 + 4\eta^2k^2 \leq 16k^2\eta^2,$$

which concludes the proof.

Chapter 9

SUPPLEMENTARY MATERIALS FOR CHAPTER V

9.1 Equivalent Definition of STFT Phase Retrieval

Since we consider an N point DFT and W satisfies $W \leq \frac{N}{2}$, STFT phase retrieval can be equivalently stated in terms of the short-time autocorrelation $\mathbf{a}_w$ [Hof64]:

$$\begin{aligned} &\text{find} && \mathbf{x} \\ &\text{subject to} \\ &a_w[m,r] = \sum_{n=0}^{N-1-m} x[n]w[rL-n]x^\star[n+m]w^\star[rL-(n+m)] \end{aligned} \tag{9.1}$$

for $0 \leq m \leq N-1$ and $0 \leq r \leq R-1$.

The knowledge of the short-time autocorrelation is sufficient for all the guarantees provided in this paper. Note that the rth column of $\mathbf{Z}_w$ and the rth column of $\mathbf{a}_w$ are Fourier pairs. Hence, for a particular r, if $Z_w[m,r]$ for $0 \leq m \leq N-1$ is available, then $a_w[m,r]$ for $0 \leq m \leq N-1$ can be calculated by taking an inverse Fourier transform. The following lemma shows that $2W$ phaseless measurements per short-time section are sufficient to infer the short-time autocorrelation.

Lemma 9.1.1. *$Z_w[m,r]$ for $1 \leq m \leq 2W-1$ is sufficient to calculate $a_w[m,r]$ for $0 \leq m \leq N-1$.*

Proof. If the window length is W, then $\mathbf{a}_w$ has nonzero values only in the interval $0 \leq m \leq W-1$ and $N-W+1 \leq m \leq N-1$. Let $\mathbf{b}_w$ be the signal obtained by circularly shifting $\mathbf{a}_w$ by $W-1$ rows, so that $\mathbf{b}_w$ has nonzero values only in the interval $0 \leq m \leq 2W-2$. Since the submatrix of the N point DFT matrix obtained by considering the first $2W-1$ columns and any $2W-1$ rows is invertible (the Vandermonde structure is retained), $Z_w[m,r]$ for $1 \leq m \leq 2W-1$ and $b_w[m,r]$ for $0 \leq m \leq 2W-2$ are related by an invertible matrix. Note that $a_w[m,r]$ for $0 \leq m \leq N-1$ can be trivially calculated from $b_w[m,r]$ for $0 \leq m \leq 2W-2$. □

Consequently, if the N point DFT is used and $2W \leq M \leq N$ is satisfied, the affine

constraints in (5.5) can be rewritten in terms of $\mathbf{a}_w$ and $\mathbf{X}$ as:

$$a_w[m,r] = \sum_{n=0}^{N-1-m} X[n,n+m]w[rL-n]w^\star[rL-(n+m)].$$

9.2 Proof of Theorem 5.3.1

The symbol $\equiv$ is used to denote equality up to a global phase and time-shift[4]. We say that two signals $\mathbf{x}_1$ and $\mathbf{x}_2$ are distinct if $\mathbf{x}_1 \not\equiv \mathbf{x}_2$, and equivalent if $\mathbf{x}_1 \equiv \mathbf{x}_2$.

Let $\mathcal{P}$ denote the set of all distinct non-vanishing complex signals of length N. $\mathcal{P}$ is a manifold of dimension $2N-1$, i.e., $\mathcal{P}$ locally resembles a real $2N-1$ dimensional space. This can be seen as follows: In order to discard the global phase of non-vanishing signals, we can assume that $x[n_0]$ is real and positive at one index n_0, without loss of generality. Hence, $x[n_0]$ can take any value in $\mathbb{R}_+$, and $x[n]$, for each $0 \le n \le N-1$ not equal to n_0, can take any value in $\mathbb{R}^2\backslash\{0,0\}$, due to the one-to-one correspondence between $\mathbb{C}$ and $\mathbb{R}^2$.

Let $\mathcal{P}_c \subset \mathcal{P}$ be the set of distinct non-vanishing complex signals which cannot be uniquely identified from their STFT magnitude if $\mathbf{w}$ is chosen such that it is non-vanishing and $W \le \frac{N}{2}$. We show that $\mathcal{P}_c$ has measure zero in $\mathcal{P}$. In order to do so, our strategy is as follows:

We first characterize $\mathcal{P}_c$ using Lemma 9.2.1. In particular, we show that $\mathcal{P}_c$ is a finite union of images of continuously differentiable maps from $\mathbb{R}^{2N-2}$ to $\mathcal{P}$. Since $\mathcal{P}$ is a manifold of dimension $2N-1$, the following result completes the proof:

Theorem 9.2.1 ([OR70], Chapter 5)**.** *If $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_1}$ is a continuously differentiable map, then the image of f has measure zero in $\mathbb{R}^{N_1}$, provided $N_0 < N_1$.*

We use the following notation in this section: If $\mathbf{g}$ is a signal of length l_g, then $\mathbf{g} = (g[0], g[1], \ldots, g[l_g-1])^T$ such that $\{g[0], g[l_g-1]\} \neq 0$ and $g[n] = 0$ outside the interval $[0, l_g-1]$. The vector $\tilde{\mathbf{g}}$ denotes the conjugate-flipped version of $\mathbf{g}$, i.e., $\tilde{\mathbf{g}} = (g^\star[l_g-1], g^\star[l_g-2], \ldots, g^\star[0])^T$. Let u_r and v_r denote the smallest and largest index where the windowed signal $\mathbf{x} \circ \mathbf{w}_r$ has a nonzero value respectively.

Lemma 9.2.1. *Consider two signals $\mathbf{x}_1 \not\equiv \mathbf{x}_2$ of length N which have the same STFT magnitude. If the window $\mathbf{w}$ is chosen such that it is non-vanishing and $W \le \frac{N}{2}$,*

[4]For non-vanishing signals, there is no ambiguity due to time-shift.

then, for each r, there exists signals $\mathbf{g}_r$ and $\mathbf{h}_r$, of lengths l_{gr} and l_{hr} respectively, such that

(i) $\mathbf{x}_1 \circ \mathbf{w}_r \equiv \mathbf{g}_r \star \mathbf{h}_r, \;\; \mathbf{x}_2 \circ \mathbf{w}_r \equiv \mathbf{g}_r \star \tilde{\mathbf{h}}_r$

(ii) $l_{gr} + l_{hr} - 1 = v_r - u_r + 1$

(iii) $g_r[l_{gr} - 1] = 1$ *and* $\{g_r[0], h_r[0], h_r[l_{hr} - 1]\} \neq 0$

where $\star$ is the convolution operator. Further, there exists at least one r such that

(iv) $l_{hr} \geq 2$, $h_r[0]$ *is real and positive.*

Proof. In Lemma 7.1 of [JOH13b], it is shown that if two non-equivalent signals of length N have the same Fourier magnitude and if the DFT dimension is at least $2N$ (this would imply that they have the same autocorrelation), then there exists signals $\mathbf{g}$ and $\mathbf{h}$, of lengths l_g and l_h respectively, such that one signal can be decomposed as $\mathbf{g} \star \mathbf{h}$ and the other signal can be decomposed as $\mathbf{g} \star \tilde{\mathbf{h}}$. For each r, the rth column of the STFT magnitude is equivalent to the Fourier magnitude of the windowed signal $\mathbf{x} \circ \mathbf{w}_r$. The DFT dimension is N, and the windowed signal length is $v_r - u_r + 1$ (which is less than or equal to $\frac{N}{2}$). Since for every r, $\mathbf{x}_1 \circ \mathbf{w}_r$ and $\mathbf{x}_2 \circ \mathbf{w}_r$ have the same Fourier magnitude, the aforementioned result proves (i).

The conditions (ii) and (iii) are properties of convolution (see Lemma 7.1 of [JOH13b] for details), and therefore hold for every r.

Furthermore, if $l_{hr} = 1$ for all $0 \leq r \leq R - 1$, then $\mathbf{x}_1 \equiv \mathbf{x}_2$. Hence, $l_{hr} \geq 2$ for at least one r. For this r, since $e^{i\phi_1}\mathbf{x}_1$ and $e^{i\phi_2}\mathbf{x}_2$ have the same STFT magnitude, $h_r[0]$ can be assumed to be real and positive without loss of generality. Hence, (iv) holds for at least one r. □

Consequently, for each $\mathbf{x} \in \mathcal{P}_c$, condition (iv) of Lemma 9.2.1 holds for at least one r. Let $\mathcal{P}_c^{r l_r l_{r+1}} \subset \mathcal{P}_c$ denote the set of signals for which $l_{hr} = l_r \geq 2$ and $l_{h,r+1} = l_{r+1}$[2]. It suffices to show that for each r, l_r and l_{r+1}, there exists a set $\mathcal{Q}_c^{r l_r l_{r+1}} \supseteq \mathcal{P}_c^{r l_r l_{r+1}}$, which is the image of a continuously differentiable map from $\mathbb{R}^{2N-2}$ to $\mathcal{P}$.

[2]When $r = R$, we consider the short-time section $r - 1$ instead of $r + 1$. We show the detailed calculations for the case when short-time section $r + 1$ is considered, the arguments are symmetric for $r - 1$.

We first show the arguments for the $L = W - 1$ case as the expressions are simple and provide intuition for the technique. Then, we show the arguments for the $L < W - 1$ case.

(i) $L = W - 1$:

The set $Q_c^{r l_r l_{r+1}}$ is constructed as follows: Consider the variables $\{\mathbf{g}_r, \mathbf{h}_r, \mathbf{g}_{r+1}, \mathbf{h}_{r+1}\}$ satisfying $l_{hr} = l_r \geq 2$ and $l_{h,r+1} = l_{r+1}$, and $x[n]$ for $n \in [0, u_r) \cup (v_{r+1}, N-1]$. The map $f = (f_0, f_1, \ldots, f_{N-1})^T$ from these variables to $\mathcal{P}$ is the following:

$$f_n = \begin{cases} x[n] \quad \text{for} \quad n \in [0, u_r) \cup (v_{r+1}, N-1] \\ \sum_{m=0}^{n-u_r} g_r[m] h_r[n - u_r - m] \\ \qquad \text{for} \quad n \in [u_r, v_r] \\ \sum_{m=0}^{n-u_{r+1}} g_{r+1}[m] h_{r+1}[n - u_{r+1} - m] \\ \qquad \text{for} \quad n \in [u_{r+1}, v_{r+1}]. \end{cases} \tag{9.2}$$

Observe that, for $n = u_{r+1} = v_r$, f_n has two definitions. The variables can admit only those values for which the two definitions have the same value. In the following, we show that there is a one-to-one correspondence between the set of admissible values of the variables and a subset of $\mathbb{R}^{2N-2}$.

Each $x[n]$, for $n \in [0, u_r) \cup (v_{r+1}, N-1]$, can be chosen from $\subset \mathbb{R}^2$. The set of $\{\mathbf{g}_{r+1}, \mathbf{h}_{r+1}\}$ is a subset of $\mathbb{R}^{2(v_{r+1} - u_{r+1} + 1)}$, which can be seen as follows: $g_{r+1}[l_{g,r+1} - 1] = 1$ is fixed (see Lemma 9.2.1), there are $v_{r+1} - u_{r+1} + 1$ other terms and each can be chosen from $\subseteq \mathbb{R}^2$.

For each choice of $\{\mathbf{g}_{r+1}, \mathbf{h}_{r+1}\}$, consider the set of $\{\mathbf{g}_r, \mathbf{h}_r\}$ excluding the terms $h_r[0]$ and $h_r[l_{hr} - 1]$: $g_r[l_{gr} - 1] = 1$ is fixed, there are $v_r - u_r - 1$ other terms and each can be chosen from $\subseteq \mathbb{R}^2$. Hence, this set is a subset of $\mathbb{R}^{2(v_r - u_r - 1)}$.

Since the short-time sections r and $r+1$ overlap in the index v_r, $\mathbf{g}_r \star \mathbf{h}_r$ and $\mathbf{g}_{r+1} \star \mathbf{h}_{r+1}$ must be consistent in this index, i.e., $\{\mathbf{g}_r, \mathbf{h}_r\}$ must satisfy:

$$\frac{1}{w[0]} h_r[l_{hr} - 1] = \frac{1}{w[W-1]} g_{r+1}[0] h_{r+1}[0]. \tag{9.3}$$

Due to Lemma 9.2.1, $\mathbf{g}_r \star \tilde{\mathbf{h}}_r$ and $\mathbf{g}_{r+1} \star \tilde{\mathbf{h}}_{r+1}$ must also be consistent in this index up to a phase, i.e., $\{\mathbf{g}_r, \mathbf{h}_r\}$ must also satisfy:

$$h_r[0] \equiv \frac{w[0]}{w[W-1]} g_{r+1}[0] h_{r+1}^{\star}[l_{h,r+1} - 1]. \tag{9.4}$$

Observe that $\equiv$ is used in (9.4), due to the fact that the equality is only up to a phase. However, $h_r[0]$ is real and positive (see Lemma 9.2.1), due to which (9.4) fixes $h_r[0]$.

Consequently, the set of admissible values of the variables, excluding $h_r[0]$ and $h_r[l_{hr}-1]$, is a subset of $\mathbb{R}^{2N-2}$, as $2(N-v_{r+1}+u_r-1+v_{r+1}-u_{r+1}+1+v_r-u_r-1) = 2N-2$. For each point in this set, $h_r[0]$ and $h_r[l_{hr}-1]$ are uniquely determined. It is straightforward to check that the map f from this set to $\mathcal{P}$ is continuously differentiable. Consequently, $Q_c^{rl_rl_{r+1}}$ is the image of a continuously differentiable map from $\mathbb{R}^{2N-2}$ to $\mathcal{P}$.

(ii) $L < W-1$:

Consider the setup for which $2L \geq W$. The set of $\{\mathbf{g}_{r+1}, \mathbf{h}_{r+1}\}$, as earlier, is a subset of $\mathbb{R}^{2(v_{r+1}-u_{r+1}+1)}$.

The short-time sections r and $r+1$ overlap in the interval $[u_{r+1}, v_r]$. Let $v_r-u_{r+1}+1 = T$ (the number of indices in the overlapping interval). Due to $2L \geq W$, we have $T = W-L \leq \lfloor \frac{W}{2} \rfloor$. Hence, for each choice of $\{\mathbf{g}_{r+1}, \mathbf{h}_{r+1}\}$, $\{\mathbf{g}_r, \mathbf{h}_r\}$ must satisfy:

$$\sum_{m=0}^{n+u_{r+1}-u_r} \frac{1}{w_r[u_r+m]} g_r[m]h_r[n+u_{r+1}-u_r-m] \tag{9.5}$$
$$= \sum_{m=0}^{n} \frac{1}{w_{r+1}[u_{r+1}+m]} g_{r+1}[m]h_{r+1}[n-m]$$

for $0 \leq n \leq T-1$. In addition, $\{\mathbf{g}_r, \mathbf{h}_r\}$ must also satisfy:

$$\frac{1}{w[0]} h_r[0] \equiv \sum_{m=0}^{T-1} \frac{1}{w_{r+1}[u_{r+1}+m]} g_{r+1}[m]h_{r+1}[T-1-m]. \tag{9.6}$$

If $l_{hr} \geq \lfloor \frac{W}{2} \rfloor + 1$, then the T bilinear equations (9.5) can be written as $\mathbf{G}\mathbf{h}_r = \mathbf{c}$, where $\mathbf{G}$ has upper triangular structure with unit diagonal entries, due to which $\text{rank}(\mathbf{G}) = T$. The set of $\mathbf{g}_r$ is a subset of $\mathbb{R}^{2(l_{gr}-1)}$. For each choice of $\mathbf{g}_r$, the terms $\{h_r[l_{hr}-T], \ldots, h_r[l_{hr}-1]\}$ are fixed by $\mathbf{G}\mathbf{h}_r = \mathbf{c}$. The constraint (9.6) fixes the value of $h_r[0]$, as earlier. Each of the remaining $(l_{hr}-1-T)$ terms of $\mathbf{h}_r$ may be chosen from $\subseteq \mathbb{R}^2$.

Hence, the set of admissible values of the variables, excluding $\{h_r[l_{hr}-T], h_r[l_{hr}-T+1], \ldots, h_r[l_{hr}-1]\}$ and $h_r[0]$, is a subset of $\mathbb{R}^{2N-2}$, due to the fact that $2(N-v_{r+1}+u_r-1+v_{r+1}-u_{r+1}+1+v_r-u_r-T) = 2N-2$ (as $l_{gr}+l_{hr}-1 =$

$v_r - u_r + 1$). For each point in this set, $h_r[0]$ and $\{h_r[l_{hr} - T], \ldots, h_r[l_{hr} - 1]\}$ are uniquely determined. The rest of the arguments are identical to those of $L = W - 1$.

If $l_{gr} \geq \lfloor \frac{W}{2} \rfloor + 1$ instead, then the bilinear equations (9.5) can be equivalently written as $\mathbf{H}\mathbf{g}_r = \mathbf{c}$, the same arguments may be applied to draw the same conclusion. For the setup with $2L > W$, the same arguments hold for the short-time sections r and $r+t$, where t is the largest integer such that the short-time sections r and $r+t$ overlap (this ensures $T \leq \lfloor \frac{W}{2} \rfloor$).

9.3 Proof of Corollary 5.3.1

We now extend Theorem 5.3.1 to incorporate sparse signals. Let $\mathcal{P}^S$ denote the set of all distinct complex signals of length N with a support S. Here, S is a binary vector of length N, such that $x[n] \neq 0$ whenever $S[n] = 1$ and $x[n] = 0$ whenever $S[n] = 0$. Further, S has less than $\min\{L, W - L\}$ consecutive zeros.

Let $\mathcal{P}_c^S \subset \mathcal{P}^S$ denote the set of signals which cannot be uniquely identified from their STFT magnitude if $\mathbf{w}$ is chosen such that it is non-vanishing and $W \leq \frac{N}{2}$. We show that $\mathcal{P}_c^S$ has measure zero in $\mathcal{P}^S$.

In the proof of Theorem 5.3.1, in order to show dimension reduction, we used the fact that for sufficient pairs of adjacent short-time sections r and $r+1$, the following holds:

(i) There is at least one index in the non-overlapping indices $[u_r, u_{r+1} - 1]$ or $[v_r + 1, v_{r+1}]$ where the signals $\mathbf{x}_1$ and $\mathbf{x}_2$ have a nonzero value. This ensures that $h_r[0]$ is not constrained by $\{\mathbf{g}_{r+1}, \mathbf{h}_{r+1}\}$ in general. This condition can be ensured by imposing the constraint that the sparse signal cannot have L consecutive zeros.

(ii) There is at least one index in the overlapping indices $[u_{r+1}, v_r]$ where the signals $\mathbf{x}_1$ and $\mathbf{x}_2$ have a nonzero value. This ensures that $h_r[0]$ is constrained by $\{\mathbf{g}_{r+1}, \mathbf{h}_{r+1}\}$ (9.4) for signals which cannot be uniquely identified by their STFT magnitude. This condition can be ensured by imposing the constraint that the sparse signal cannot have $W - L$ consecutive zeros.

The only difference in the proof is the following: Unlike in the case of non-vanishing signals, there is time-shift ambiguity. Hence, the constraint (9.6) is replaced by

$$\frac{1}{w[0]} h_r[0] \equiv \sum_{m=0}^{n} \frac{1}{w_{r+1}[u_{r+1} + m]} g_{r+1}[m] h_{r+1}[n - m] \tag{9.7}$$

for *some* $0 \leq n \leq T - 1$. This fixes the value of $h_r[0]$ to one of at most T values, due to which there is a dimension reduction.

9.4 Proof of Theorem 5.4.1

We first show the arguments for the case $2W \leq M \leq N$ (short-time autocorrelation known) as the expressions are simple and provide intuition. Then, we show the arguments for the case $4L \leq M < 2W$ (super-resolution).

(i) $2W \leq M \leq N$:

The affine constraints in (5.5) can be rewritten as (see Section 9.1):

$$a_w[m, r] = \sum_{n=0}^{N-1-m} X[n, n+m]w[rL-n]w^\star[rL-(n+m)].$$

The proof strategy is as follows: We begin by focusing our attention on short-time section $r = 1$. We show that the prior information available, along with the affine autocorrelation measurements corresponding to $r = 1$ and the positive semidefinite constraint, will ensure that every feasible matrix of (5.5) satisfies $X[n, m] = x_0[n]x_0^\star[m]$ for $0 \leq n, m \leq L$. We then apply this argument incrementally, i.e., we show that the affine measurements corresponding to short-time section r, along with the entries of $\mathbf{X}$ uniquely determined and the positive semidefinite constraint, will ensure that $X[n, m] = x_0[n]x_0^\star[m]$ for $u_r \leq n, m \leq v_r$, where u_r and v_r denote the smallest and largest index where $\mathbf{w}_r$ has a nonzero value respectively. Consequently, the entries along the diagonal and the first $W - L$ off-diagonals of every feasible matrix of (5.5) match the entries along the diagonal and the first $W - L$ off-diagonals of the matrix $\mathbf{x}_0\mathbf{x}_0^\star$. Since the entries are sampled from a rank one matrix with nonzero diagonal entries (i.e., $\mathbf{x}_0\mathbf{x}_0^\star$), there is exactly one positive semidefinite completion, which is the rank one completion $\mathbf{x}_0\mathbf{x}_0^\star$ [HJ12].

Let $\mathbf{s}_0 = (x_0[0], x_0[1], \ldots, x_0[L])^T$ be a length $L + 1$ subsignal of $\mathbf{x}_0$, and $\mathbf{S}$ be the $(L+1) \times (L+1)$ submatrix of $\mathbf{X}$ corresponding to the first $L + 1$ rows and columns. We now show that $\mathbf{S} = \mathbf{s}_0\mathbf{s}_0^\star$ is the only feasible matrix under the constraints of (5.5).

Since $x_0[n]$ for $0 \leq n \leq \left\lfloor \frac{L}{2} \right\rfloor$ is known a priori, we have $S[n, m] = x_0[n]x_0^\star[m]$ for $0 \leq n, m \leq \left\lfloor \frac{L}{2} \right\rfloor$. Let $\mathcal{A}(\mathbf{S}) = \mathbf{c}$ denote these constraints due to prior information, along with the affine constraints corresponding to $r = 1$. In particular, $\mathcal{A}(\mathbf{S}) = \mathbf{c}$ denotes the following set of constraints:

$$S[n, m] = x_0[n]x_0^\star[m] \quad \text{for} \quad 0 \leq n, m \leq \left\lfloor \frac{L}{2} \right\rfloor,$$

$$a_w[m, 1] = \sum_{n=0}^{L-m} S[n, n+m]w[L-n]w^\star[L-(n+m)].$$

For each feasible matrix $\mathbf{S}$, these set of measurements fix (i) the $\left\lfloor \frac{L}{2} + 1 \right\rfloor \times \left\lfloor \frac{L}{2} + 1 \right\rfloor$ submatrix, corresponding to the first $\left\lfloor \frac{L}{2} + 1 \right\rfloor$ rows and columns, of $\mathbf{S}$ (ii) the appropriately weighted sum along the diagonal and each off-diagonal of $\mathbf{S}$ ($2L \leq W$ is implicitly used here).

Lemma 9.4.1. *If $\mathbf{S}_0 = \mathbf{s}_0\mathbf{s}_0^\star$ satisfies $\mathcal{A}(\mathbf{S}) = \mathbf{c}$, then it is the only positive semidefinite matrix which satisfies $\mathcal{A}(\mathbf{S}) = \mathbf{c}$.*

Proof. Let T be the set of Hermitian matrices of the form

$$T = \{\mathbf{S} = \mathbf{s}_0\mathbf{v}^\star + \mathbf{v}\mathbf{s}_0^\star : \mathbf{v} \in \mathbb{C}^n\}$$

and $T^\perp$ be its orthogonal complement. The set T may be interpreted as the tangent space at $\mathbf{s}_0\mathbf{s}_0^\star$ to the manifold of Hermitian matrices of rank one. Influenced by [CSV13], we use $\mathbf{S}_T$ and $\mathbf{S}_{T^\perp}$ to denote the projection of a matrix $\mathbf{S}$ onto the subspaces T and $T^\perp$ respectively.

Standard duality arguments in semidefinite programming show that the following is sufficient for $\mathbf{S}_0 = \mathbf{s}_0\mathbf{s}_0^\star$ to be the unique optimizer of (5.5):

(i) *Condition 1*: $\mathbf{S} \in T \quad$ and $\quad \mathcal{A}(\mathbf{S}) = 0 \Rightarrow \mathbf{S} = 0$.

(ii) *Condition 2*: There exists a *dual certificate* $\mathbf{M}$ in the range space of $\mathcal{A}^\star$ obeying:

- $\mathbf{M}\mathbf{s}_0 = 0$
- $\operatorname{rank}(\mathbf{M}) = L$
- $\mathbf{M} \succcurlyeq 0$.

The proof of this result is based on KKT conditions, and can be found in any standard reference on semidefinite programming (for example, see [VB96]).

We first show that *Condition 1* is satisfied. The set of constraints in $\mathcal{A}(\mathbf{S}) = 0$ due to prior information fix the entries of the first $\left\lfloor \frac{L}{2} + 1 \right\rfloor$ rows and columns of $\mathbf{S}$ to 0. Since $\mathbf{S} = \mathbf{s}_0\mathbf{v}^\star + \mathbf{v}\mathbf{s}_0^\star$ for some $\mathbf{v} = (v[0], v[1], \ldots, v[L])^T$ (due to $\mathbf{S} \in T$), we infer that $v[n] = icx_0[n]$ for $0 \leq n \leq \left\lfloor \frac{L}{2} \right\rfloor$, for some real constant c. Indeed, the equations of the form $s_0[n]v^\star[n] + v[n]s_0^\star[n] = 0$ imply $v[n] = ic_n x_0[n]$, for some real constant c_n. The equations $s_0[n]v^\star[m] + v[n]s_0^\star[m] = 0$ imply $c_n = c_m$.

The set of constraints in $\mathcal{A}(\mathbf{S}) = 0$ due to the measurements corresponding to $r = 1$, along with $v[n] = icx_0[n]$ for $0 \leq n \leq \left\lfloor \frac{L}{2} \right\rfloor$, imply $v[n] = icx_0[n]$ for $\left\lfloor \frac{L}{2} + 1 \right\rfloor \leq n \leq L$. Hence, for $\mathbf{S} \in T$, $\mathcal{A}(\mathbf{S}) = 0$ implies $\mathbf{v} = ic\mathbf{s}_0$, which in turn implies $\mathbf{S} = -ic\mathbf{s}_0\mathbf{s}_0^\star + ic\mathbf{s}_0\mathbf{s}_0^\star = 0$.

We next establish *Condition 2*. For simplicity of notation, we consider the case where $w[n] = 1$ for $0 \leq n \leq W - 1$. For a general non-vanishing $\mathbf{w}$, the same arguments hold (the Toeplitz matrix considered is appropriately redefined with weights).

The range space of $\mathcal{A}^\star$ is the set of all $L + 1 \times L + 1$ matrices which are a sum of the following two matrices: The first matrix can have any value in the $\left\lfloor \frac{L}{2} + 1 \right\rfloor \times \left\lfloor \frac{L}{2} + 1 \right\rfloor$ submatrix corresponding to the first $\left\lfloor \frac{L}{2} + 1 \right\rfloor$ rows and columns, and has a value zero outside this submatrix (dual of the set of constraints due to prior information). The second matrix has a Toeplitz structure (dual of the measurements corresponding to $r = 1$).

Suppose $\mathbf{s}_1$ is the vector containing the first $\left\lfloor \frac{L}{2} + 1 \right\rfloor$ entries of $\mathbf{s}_0$ and $\mathbf{s}_2$ is the vector containing the remaining entries of $\mathbf{s}_0$. Here, $\mathbf{s}_1$ corresponds to the locations where we have knowledge of the entries and $\mathbf{s}_2$ corresponds to the locations where the entries are not determined. Let $\mathbf{L}$ be a lower triangular $\left\lceil \frac{L}{2} \right\rceil \times \left\lfloor \frac{L}{2} + 1 \right\rfloor$ Toeplitz matrix satisfying $\mathbf{L}\mathbf{s}_1 + \mathbf{s}_2 = 0$. Such an $\mathbf{L}$ always exists if $s_1[0]$ is nonzero and the length of $\mathbf{s}_1$ is greater than or equal to the length of $\mathbf{s}_2$. Let Λ be any $\left\lfloor \frac{L}{2} + 1 \right\rfloor \times \left\lfloor \frac{L}{2} + 1 \right\rfloor$ positive semidefinite matrix with rank $\left\lfloor \frac{L}{2} \right\rfloor$ satisfying $\Lambda\mathbf{s}_1 = 0$. Again, such a Λ always exists (any positive semidefinite matrix with eigenvectors perpendicular to $\mathbf{s}_1$). Consider the following dual certificate:

$$\mathbf{M} = \begin{bmatrix} \mathbf{L}^\star\mathbf{L} + \Lambda & \mathbf{L}^\star \\ \mathbf{L} & \mathbf{I}_{\left\lceil \frac{L}{2} \right\rceil} \end{bmatrix}. \tag{9.8}$$

Clearly, $\mathbf{M}$ is in the range space of $\mathcal{A}^\star$. Also, $\mathbf{M}\mathbf{s}_0 = 0$ by construction. From the Schur complement, it is straightforward to see that $\text{rank}(\mathbf{M}) = L$ and $\mathbf{M} \succcurlyeq 0$. □

We have shown that $\mathbf{S}_0 = \mathbf{s}_0\mathbf{s}_0^\star$ is the only positive semidefinite matrix which satisfies the prior information and the measurements corresponding to $r = 1$. Redefine $\mathbf{s}_0$ and $\mathbf{S}$ such that $\mathbf{s}_0 = (x_0[0], x_0[1], \ldots, x_0[2L])^T$ is the $2L + 1$ length subsignal of $\mathbf{x}$ and $\mathbf{S}$ is the $(2L + 1) \times (2L + 1)$ submatrix of $\mathbf{X}$ corresponding to the first $2L + 1$ rows and columns.

We already have $S[n,m] = x_0[n]x_0^\star[m]$ for $0 \leq n, m \leq L$ from above. Let $\mathcal{A}(\mathbf{S}) = \mathbf{c}$ denote these constraints, along with the affine constraints corresponding to $r = 2$. Due to $2L \leq W$, Lemma 9.4.1 proves that $\mathbf{S}_0 = \mathbf{s}_0\mathbf{s}_0^\star$ is the only psd matrix which satisfies the prior information and the measurements corresponding to $r = 1, 2$. Applying this argument incrementally, the entries along the diagonal and the first $W - L$ off-diagonals of every feasible matrix of (5.5) match the entries along the diagonal and the first $W - L$ off-diagonals of the matrix $\mathbf{x}_0\mathbf{x}_0^\star$.

Sparse signals: The arguments can be seamlessly extended to incorporate sparse signals.

(i) The fact that there exists a unique positive semidefinite completion once the diagonal and the first $W - L$ off-diagonal entries are sampled from $\mathbf{x}_0\mathbf{x}_0^\star$ holds when $\mathbf{x}_0$ has less than $W - L$ consecutive zeros.

(ii) Note that the length of $\mathbf{s}_2$ is at most L, as it corresponds to the locations in the window where the entries are not determined. Since we know $x_0[n]$ for $i_0 \leq n < i_0 + L$ a priori, where i_0 is the smallest index such that $x_0[i_0] \neq 0$, the length of $\mathbf{s}_1$ is $W - L$. Redefine $\mathbf{s}_1$ so that it corresponds to the locations in the window where the entries are determined, starting from the smallest index which has a nonzero value in order to ensure $s_1[0] \neq 0$. If $\mathbf{x}_0$ has at most $W-2L$ consecutive zeros, then the length of $\mathbf{s}_1$ is at least $(W - L) - (W - 2L) = L$. Hence, a lower triangular Toeplitz matrix $\mathbf{L}$, satisfying $\mathbf{L}\mathbf{s}_1 + \mathbf{s}_2 = 0$, always exists.

(ii) $4L \leq M < 2W$:

The range space of the dual certificate is the set of all $L + 1 \times L + 1$ matrices which are a sum of the following two matrices: The first matrix can have any value in the $\left\lfloor \frac{L}{2} + 1 \right\rfloor \times \left\lfloor \frac{L}{2} + 1 \right\rfloor$ submatrix corresponding to the first $\left\lfloor \frac{L}{2} + 1 \right\rfloor$ rows and columns, and has a value zero outside this submatrix (dual of the set of constraints due to prior information). The second matrix has the form $\sum_{m=1}^{M} \alpha_m \mathbf{W}_r^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{W}_r$, where α_m is real-valued for each m (dual of the measurements corresponding to $r = 1$).

Let $\mathbf{l} = (l[0], l[1], \ldots, l[N-1])^T$ be a vector that satisfies:

(i) $l[0] = 1, l[n] = l[N - n] = 0 \quad \text{for} \quad 1 \leq n \leq \left\lceil \frac{L}{2} \right\rceil - 1$

(ii) $\sum_{n=0}^{m} x_0[n]l[m - n] = \sum_{n=0}^{m} x_0^\star[n]l[N - m + n] = 0 \quad \text{for} \quad \left\lfloor \frac{L}{2} + 1 \right\rfloor \leq m \leq L$

(iii) $\mathbf{f}_m^\star \mathbf{l} = 0 \quad \text{for} \quad M + 1 \leq m \leq N.$

These constraints together can be written as $\mathbf{Al} = \mathbf{b}$. When $M \geq 4\lceil \frac{L}{2} \rceil$, the matrix $\mathbf{A}$ is square or wide, and almost always (pseudo) invertible. This can be seen as follows: The determinant of $\mathbf{A}$ is a polynomial function of the entries of $\mathbf{x}_0$, due to which it is either always zero or almost surely nonzero. By substituting $x_0[0] = 1$ and $x_0[n] = 0$ for $n \neq 0$, it is straightforward to check that the determinant is nonzero. Hence, such an $\mathbf{l}$ almost always exists.

If the last row in (9.8) is chosen as $(l[L], l[L-1], \ldots, l[0])$, then we have: (i) The lower right block is an identity matrix. (ii) $\mathbf{Ls}_1 + \mathbf{s}_2 = 0$ is satisfied. (iii) Since $\mathbf{b}$ is a real vector, $\mathbf{l}$ satisfies $l[n] = l^\star[N-n]$. Therefore, $\mathbf{l}$ is in the range space of $\sum_{m=1}^{M} \alpha_m \mathbf{f}_m$ where α_m is real-valued, due to which the resulting second matrix is in the range space of $\sum_{m=1}^{M} \alpha_m \mathbf{W}_r^\star \mathbf{f}_m \mathbf{f}_m^\star \mathbf{W}_r$.

Therefore, $\mathbf{M}$ satisfies all the requirements. The arguments are applied incrementally as earlier, with $M \geq 4L$ for $r > 1$.

9.5 Alternative Proof of Theorem 5.4.2

The affine constraints in (5.5) can be rewritten as (see Section 9.1):

$$a_w[m, r] = \sum_{n=0}^{N-1-m} X[n, n+m] w[rL-n] w^\star[rL-(n+m)].$$

Due the constraint corresponding to $r = 0$, $\mathbf{X}$ has to satisfy:

$$|w[0]|^2 X[0,0] = a_w[0,0] = |w[0]|^2 |x_0[0]|^2,$$

because of which $X[0,0]$ is fixed to $|x_0[0]|^2$ as $w[0] \neq 0$. Due to the constraints corresponding to $r = 1$, $\mathbf{X}$ has to satisfy:

$$\begin{aligned} |w[0]|^2 X[1,1] + |w[1]|^2 X[0,0] &= a_w[0,1] \\ &= |w[0]|^2 |x_0[1]|^2 + |w[1]|^2 |x_0[0]|^2, \\ w^\star[0] w[1] X[0,1] &= a_w[1,1] = w^\star[0] w[1] x_0[0] x_0^\star[1]. \end{aligned}$$

Since $X[0,0] = |x_0[0]|^2$ and $w[0]w[1] \neq 0$, $X[1,1]$ and $X[0,1]$ are fixed to $|x_0[1]|^2$ and $x_0[0]x_0^\star[1]$ respectively.

Applying this argument incrementally, the measurements corresponding to short-time section r, with the help of the entries of $\mathbf{X}$ uniquely determined, fix the value of $X[r,r]$ and $X[r-1,r]$ to $|x_0[r]|^2$ and $x_0[r-1]x_0^\star[r]$ respectively. Hence, the diagonal and the first off-diagonal entries of every feasible matrix of (5.5) match the diagonal and the first off-diagonal entries of the matrix $\mathbf{x}_0\mathbf{x}_0^\star$. Since the entries

are sampled from a rank one matrix with nonzero diagonal entries (i.e., $\mathbf{x}_0\mathbf{x}_0^\star$), there is exactly one positive semidefinite completion, which is the rank one completion $\mathbf{x}_0\mathbf{x}_0^\star$ [HJ12].

In particular, due to the aforementioned determined entries of $\mathbf{X}$ and the positive semidefinite constraint $\mathbf{X} \succcurlyeq 0$, the convex program (5.5) has only one feasible matrix, given by $\mathbf{x}_0\mathbf{x}_0^\star$. The underlying signal $\mathbf{x}_0$ can be recovered (up to a global phase) by a simple decomposition.

Chapter 10

SUPPLEMENTARY MATERIALS FOR CHAPTER VI

10.1 Proof of Theorem 6.5.1

In the proof of Theorem 6.4.1, we showed that the measurements corresponding to the 0th mask determine the diagonal entries of $\mathbf{W} = \mathbf{F}_M \mathbf{X} \mathbf{F}_M^\star$, and the measurements corresponding to the $0, \pm r$th mask determine the entries along the rth off-diagonal of $\mathbf{W}$.

In order to prove this theorem, we will use a proof technique commonly known as *dual certificate method*. Let $\mathbf{w}_0 = (y_0[0], y_0[1], \ldots, y_0[M-1])^T$, T be the set of symmetric matrices of the form

$$T = \{\mathbf{W} = \mathbf{w}_0 \mathbf{v}^\star + \mathbf{v}\mathbf{w}_0^\star : \mathbf{v} \in \mathbb{C}^M\}$$

and $T^\perp$ be its orthogonal complement. T may be interpreted as the tangent space at $\mathbf{w}_0\mathbf{w}_0^\star$ to the manifold of symmetric matrices of rank one. Influenced by [CSV13], we use $\mathbf{W}_T$ and $\mathbf{W}_{T^\perp}$ to denote the projection of a matrix $\mathbf{W}$ onto the subspaces T and $T^\perp$ respectively.

Standard duality arguments in semidefinite programming show that sufficient conditions for $\mathbf{w}_0\mathbf{w}_0^\star$ to be the unique optimizer of

$$\begin{aligned} &\text{minimize} && 0 \\ &\text{subject to} && \mathcal{A}(\mathbf{W}) = \mathbf{c} \\ & && \mathbf{W} \succcurlyeq 0, \end{aligned} \tag{10.1}$$

where $\mathcal{A}(\mathbf{W}) = \mathbf{c}$ denotes the set of constraints corresponding to the knowledge of the diagonal and the first R off-diagonal entries of $\mathbf{W}$, i.e., $W[n,m] = w_0[n]w_0^\star[m]$ for $|n-m| \leq R$, is:

(i) *Condition 1*: $\mathbf{W} \in T \quad \& \quad \mathcal{A}(\mathbf{W}) = 0 \Rightarrow \mathbf{W} = 0$.

(ii) *Condition 2*: There exists a *dual certificate* $\mathbf{M}$ in the range space of $\mathcal{A}^\star$ obeying:

- $\mathbf{M}\mathbf{w}_0 = 0$

- $\text{rank}(\mathbf{M}) = M - 1$
- $\mathbf{M} \succcurlyeq 0$.

The range space of $\mathcal{A}^\star$ is the set of all matrices which have nonzero entries only along the diagonal or the first R off-diagonals (dual of the measurement constraints). For the setup with $R = 1$, consider the following *dual certificate*:

$$\mathbf{M} = \sum_{m=0}^{M-2} \mathbf{u}_{m,m+1}\mathbf{u}_{m,m+1}^\star, \tag{10.2}$$

where, for each m, $\mathbf{u}_{m,m+1}$ is an $M \times 1$ vector such that $u_{m,m+1}[m] = \frac{w_0^\star[m+1]}{\|\mathbf{w}_0\|_2}$, $u_{m,m+1}[m+1] = -\frac{w_0^\star[m]}{\|\mathbf{w}_0\|_2}$, and zero everywhere else.

We first show that *Condition 1* is satisfied. Since the diagonal entries of $\mathbf{W}$ are fixed by the measurements, we have the identity $w_0[m]v^\star[m] + v[m]w_0^\star[m] = 0$ for each m. Consequently, $v[m] = ic_m w_0[m]$ holds for some real constant c_m. Since the first off-diagonal entries of $\mathbf{W}$ are determined by the measurements too, we have the identity $w_0[m]v^\star[m+1] + v[m]w_0^\star[m+1] = 0$ for each m. We infer $c_m = c_{m+1}$, due to which $\mathbf{v} = ic\mathbf{w}_0$ for some real constant c. Hence, for $\mathbf{W} \in T$, $\mathcal{A}(\mathbf{W}) = 0$ implies $\mathbf{v} = ic\mathbf{w}_0$, which in turn implies $\mathbf{W} = -ic\mathbf{w}_0\mathbf{w}_0^\star + ic\mathbf{w}_0\mathbf{w}_0^\star = 0$.

We next establish *Condition 2*. By construction, $\mathbf{u}_{m,m+1}^\star\mathbf{w}_0 = 0$ holds for all m, due to which we have $\mathbf{M}\mathbf{w}_0 = 0$. Also, since $\sum_{m=0}^{M-2} \alpha_{m,m+1}\mathbf{u}_{m,m+1} = 0$ implies $\alpha_{m,m+1} = 0$ for all m (linear independence of $\mathbf{u}_{m,m+1}$), we have $\text{rank}(\mathbf{M}) = M - 1$. The matrix $\mathbf{M}$ is positive semidefinite by construction.

For the setup with $R > 1$, we now construct a dual certificate (10.3), which is similar in nature to (10.2).

Let $p_1[m]$ and $p_2[m]$ be integers between 0 and $M - 1$, for $0 \le m \le M - 2$, chosen as follows: We set $p_1[0]$ to be the index in $\mathbf{w}_0$ with the largest absolute value, and $p_2[m] = p_1[0] + 1 + m$ where m ranges from 0 until $p_2[m] = M - 1$ (say, this equality happens at $m = m_0 - 1$). For these choices of m, we choose $p_1[m]$ to be the index in the range $\{p_2[m] - R, p_2[m] - 1\}$ with the largest absolute value in $\mathbf{w}_0$. Then, we set $p_2[m_0 + m] = p_1[0] - 1 - m$ where m ranges from 0 until $p_2[m_0 + m] = 0$. For these choices of m, we choose $p_1[m_0 + m]$ to be the index in the range $\{p_2[m + m_0] + 1, p_2[m + m_0] + R\}$ with the largest absolute value in $\mathbf{w}_0$. The intuition is the following: Instead of setting the phase of $w_0[0]$ to zero and decoding the phases of $w_0[m]$ in the increasing order of m using relative phase

information of $w_0[m-1]$ and $w_0[m]$, we can instead initialize this process at the index corresponding to the largest absolute value in $\mathbf{w}_0$, and decode such that, at the mth step, we decode the phase of $w_0[p_2[m]]$ using the relative phase information of $w_0[p_1[m]]$ and $w_0[p_2[m]]$. By doing so, we ensure $|w_0[p_1[m]]|^2 \geq \gamma_{min}$ for all m, which mitigates the impact of indices with low absolute values in the decoding process.

In particular, we consider the dual certificate

$$\mathbf{M} = \sum_{m=0}^{M-2} \mathbf{u}_{p_1[m],p_2[m]} \mathbf{u}^{\star}_{p_1[m],p_2[m]}, \tag{10.3}$$

where, for each m, $\mathbf{u}_{p_1[m],p_2[m]}$ is an $M \times 1$ vector such that $u_{p_1[m],p_2[m]}[p_1[m]] = \frac{w_0^{\star}[p_2[m]]}{\|\mathbf{w}_0\|_2}$, $u_{p_1[m],p_2[m]}[p_2[m]] = -\frac{w_0^{\star}[p_1[m]]}{\|\mathbf{w}_0\|_2}$, and zero everywhere else. It is straightforward to check that all the conditions are satisfied, using the same arguments as earlier. We will now use this dual certificate (10.3) to bound the reconstruction error, for a given bound on the ℓ_1 norm of the noise vector η.

Let $\hat{\mathbf{W}} = \mathbf{w}_0\mathbf{w}_0^{\star} + \mathbf{H}$ be the optimizer of (10.1) in the noisy setting, i.e., with $\|\mathcal{A}(\mathbf{W}) - \mathbf{c}\|_1 \leq \eta$. Since both $\mathbf{w}_0\mathbf{w}_0^{\star}$ and $\hat{\mathbf{W}}$ are feasible, we infer $\sum_{r=-R}^{R} \sum_m |H[m, m+r]| \leq 2\eta$.

Upper bound of $\|\mathbf{H}_{T^{\perp}}\|_2$:

We have $\langle \mathbf{M}, \mathbf{H} \rangle = \text{trace}(\mathbf{M}^{\star}\mathbf{H}) \leq \|\mathbf{M}\|_{\infty} \sum_{r=-R}^{R} \sum_m |H[m, m+r]| \leq 2R \frac{\gamma_{max}}{\|\mathbf{w}_0\|_2^2} \times 2\eta$, where we use the fact that $\frac{\gamma_{max}}{\|\mathbf{w}_0\|_2^2}$ bounds the infinity norm of each term of the summation in (10.3), and each location can be nonzero in at most $2R$ of the summands. Also, note that $\langle \mathbf{M}, \mathbf{H} \rangle = \langle \mathbf{M}, \mathbf{H}_T \rangle + \langle \mathbf{M}, \mathbf{H}_{T^{\perp}} \rangle \geq \|\mathbf{H}_{T^{\perp}}\|_2 \sigma(\mathbf{M})$, where $\sigma(\mathbf{M})$ is the smallest nonzero singular value of $\mathbf{M}$. Here, we use $\langle \mathbf{M}, \mathbf{H}_T \rangle = 0$ (due to $\mathbf{M}\mathbf{w}_0 = 0$) and $\mathbf{H}_{T^{\perp}} \succcurlyeq 0$. Consequently, $\|\mathbf{H}_{T^{\perp}}\|_2 \leq \frac{4R\gamma_{max}}{\|\mathbf{w}_0\|_2^2 \sigma(\mathbf{M})} \eta = \theta(\mathbf{w}_0)\eta$, for convenience of notation.

Calculation of $\sigma(\mathbf{M})$:

Suppose $\min_{\mathbf{h}^{\star}\mathbf{w}_0 = 0, \|\mathbf{h}\|_2 = 1} \mathbf{h}^{\star}\mathbf{M}\mathbf{h} \leq \delta^2$. Then, for each m, we have $|\mathbf{u}^{\star}_{p_1[m],p_2[m]}\mathbf{h}| \leq \delta$, or in other words, $\left| \frac{h[p_1[m]]}{w_0[p_1[m]]} - \frac{h[p_2[m]]}{w_0[p_2[m]]} \right| \leq \frac{\delta \|\mathbf{w}_0\|_2}{|w_0[p_1[m]] w_0[p_2[m]]|}$. Consequently, $\left| \frac{h[p_2[m]]}{w_0[p_2[m]]} \right| \geq \left| \frac{h[p_1[m]]}{w_0[p_1[m]]} \right| - \frac{\delta \|\mathbf{w}_0\|_2}{|w_0[p_1[m]] w_0[p_2[m]]|}$.

By iteration, we have $\left| \frac{h[p_2[m]]}{w_0[p_2[m]]} \right| \geq \left| \frac{h[p_1[0]]}{w_0[p_1[0]]} \right| - \frac{\delta \|\mathbf{w}_0\|_2}{|w_0[p_1[m]] w_0[p_2[m]]|} - \frac{M\delta \|\mathbf{w}_0\|_2}{\gamma_{min}}$. For convenience of notation, let $\alpha = \frac{h[p_1[0]]}{w_0[p_1[0]]} \|\mathbf{w}_0\|_2 \geq 0$ without loss of generality

(global phase factor ambiguity). If $\delta < \frac{\gamma_{min}}{4M\|\mathbf{w}_0\|_2^2}$, then $\mathbf{h}^T\mathbf{w}_0 = \sum_m \frac{h[m]}{w_0[m]}|w_0[m]|^2 > (\alpha - 1/4)\|\mathbf{w}_0\|_2 - \frac{\|\mathbf{w}_0\|_1^2}{4M\|\mathbf{w}_0\|_2}$, due to the fact that $\sqrt{\gamma_{min}} \leq |w[p_1[m]]| \leq \|\mathbf{w}_0\|_1$ for all m. If we show that $\alpha \geq \frac{1}{2}$, then we are through as $\mathbf{h}^\star\mathbf{w}_0 > 0$ is a contradiction. Note that $\mathbf{h}^\star\mathbf{h} = 1 = \sum_m \left|\frac{h[m]}{w_0[m]}\right|^2 |w_0[m]|^2 \leq \sum_m \left(\frac{\alpha+1/4}{\|\mathbf{w}_0\|_2} + \frac{\|\mathbf{w}_0\|_1}{4M|w_0[m]|\|\mathbf{w}_0\|_2}\right)^2 |w_0[m]|^2 \leq (\alpha + 1/4)^2 + 2(\alpha + 1/4)1/4 + 1/16 = (\alpha + 1/2)^2$, which completes the proof. Therefore, $\sigma(\mathbf{M}) \geq \frac{\gamma_{min}^2}{16M^2\|\mathbf{w}_0\|_2^4}$.

Upper bound of $\|\mathbf{H}_T\|_2$:

We have $\sum_{r=-R}^{R} \sum_m |H_T[m, m+r] + H_{T^\perp}[m, m+r]| \leq 2\eta$, from which we can infer $\sum_{r=-R}^{R} \sum_m \left|w_0[m]w_0^\star[m+r]\right| \left|\frac{v^\star[m+r]}{w_0^\star[m+r]} + \frac{v[m]}{w_0[m]}\right| \leq \sum_r \sum_m |H_{T^\perp}[m, m+r]| + 2\eta \leq (\theta(\mathbf{w}_0)\sqrt{M}+2)\eta$. When $r > R$, $\left|\frac{v^\star[m+r]}{w_0^\star[m+r]} + \frac{v[m]}{w_0[m]}\right|$ can be expressed using a summation of terms with differences $\leq R$, i.e., $\left|\frac{v^\star[m+r]}{w_0^\star[m+r]} + \ldots - \frac{v^\star[m+r_1]}{w_0^\star[m+r_1]} + \frac{v^\star[m+r_1]}{w_0^\star[m+r_1]} + \frac{v[m]}{w_0[m]}\right|$ with $r_1 \leq R$, $r_2 - r_1 \leq R$ and so on, while ensuring $|w_0[m + r_i]|^2 \geq \gamma_{min}$. By using triangle inequality, we get $\sum_r \left|w_0[m]w_0^\star[m + r]\right| \left|\frac{v^\star[m+r]}{w_0^\star[m+r]} + \frac{v[m]}{w_0[m]}\right| \leq \frac{\gamma_{max}}{\gamma_{min}} M \times (\theta(\mathbf{w}_0)\sqrt{M} + 2)\eta$, by using the fact that $\left|w_0[m]w_0^\star[m + r]\right| = \left|w_0[m]w_0^\star[m + r_i]\right| \left|\frac{w_0^\star[m+r]}{w_0^\star[m+r_i]}\right| \leq \left|w_0[m]w_0^\star[m + r_i]\right| \frac{\gamma_{max}}{\gamma_{min}}$. The same quantity also bounds the sum of the absolute values along each column, and the spectral norm of $\mathbf{H}_T$ (due to Holder's inequality). Consequently, $\|\mathbf{H}_T\|_2 \leq 2\|\mathbf{H}_T\| \leq 2\frac{\gamma_{max}}{\gamma_{min}} M(\theta(\mathbf{w}_0)\sqrt{M} + 2)\eta$.

Combining the expressions, we infer $\|\mathbf{H}\|_2 = \|\mathbf{H}_T\|_2 + \|\mathbf{H}_{T^\perp}\|_2 \leq C_0 \frac{\|\mathbf{w}_0\|_2^2\gamma_{max}^2}{\gamma_{min}^3} RM^{3.5}\eta$. In other words, we have $\|\hat{\mathbf{W}} - \mathbf{W}_0\|_1 \leq C_0 \frac{\|\mathbf{w}_0\|_2^2\gamma_{max}^2}{\gamma_{min}^3} RM^4\eta$.

The next step is to bound $\|\hat{\mathbf{X}} - \mathbf{X}_0\|_1$, where $\hat{\mathbf{X}}$ is the solution of the convex program (6.8) in the noisy setting. This can be done by extending the arguments in the proof of Theorem 1.5 in [CF14].

Let $\mathbf{P} = \hat{\mathbf{X}} - \mathbf{X}_0$, and $\mathbf{P}_{S,.}$, $\mathbf{P}_{.,S}$ and $\mathbf{P}_{S,S}$ denote the projections onto the linear space of matrices supported on rows indexed by S, columns indexed by S, and rows and columns indexed by S respectively. Also, let $\mathcal{P}_M = \mathbf{F}_M^\star\mathbf{F}_M$ and $\mathcal{P}_M^\perp = \mathbf{I} - \mathbf{F}_M^\star\mathbf{F}_M$.

A straightforward extension of Lemma 3.1 in [CF14] results in the bound $\|\mathbf{P}_{S,S}\|_1 \leq \rho\|\mathbf{P}_{S^c,S} + \mathbf{P}_{S,S^c} + \mathbf{P}_{S^c,S^c}\|_1$, for any $\mathbf{P}$ satisfying $\mathbf{F}_M\mathbf{P}\mathbf{F}_M^\star = 0$, where $\rho = 1 - \frac{\alpha}{SRF^4}$. Here, SRF is the super-resolution factor $\frac{N}{M}$ and α is a positive constant.

We have $\|\mathbf{X}_0\|_1 \geq \|\mathbf{X}_0 + \mathbf{P}\|_1 \geq \|\mathbf{X}_0 + \mathcal{P}_M^\perp\mathbf{P}\mathcal{P}_M + \mathcal{P}_M\mathbf{P}\mathcal{P}_M^\perp + \mathcal{P}_M^\perp\mathbf{P}\mathcal{P}_M^\perp\|_1 - \|\mathcal{P}_M\mathbf{P}\mathcal{P}_M\|_1$, which leads to $\|\mathbf{X}_0\|_1 \geq \|\mathbf{X}_0\|_1 + \|(\mathcal{P}_M^\perp\mathbf{P}\mathcal{P}_M + \mathcal{P}_M\mathbf{P}\mathcal{P}_M^\perp + \mathcal{P}_M^\perp\mathbf{P}\mathcal{P}_M^\perp)_{S^c,S^c} + (\mathcal{P}_M^\perp\mathbf{P}\mathcal{P}_M +$

$\mathcal{P}_M\mathbf{P}\mathcal{P}_M^{\perp} + \mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M^{\perp})_{S^c,S} + (\mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M + \mathcal{P}_M\mathbf{P}\mathcal{P}_M^{\perp} + \mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M^{\perp})_{S,S^c}\|_1 - \|(\mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M + \mathcal{P}_M\mathbf{P}\mathcal{P}_M^{\perp} + \mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M^{\perp})_{S,S}\|_1 - \|\mathcal{P}_M\mathbf{P}\mathcal{P}_M\|_1$.

Consequently, we infer $\|(\mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M + \mathcal{P}_M\mathbf{P}\mathcal{P}_M^{\perp} + \mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M^{\perp})_{\{S,S^c\}\cup\{S^c,S\}\cup\{S^c,S^c\}}\|_1 \leq \frac{1}{1-\rho}\|\mathcal{P}_M\mathbf{P}\mathcal{P}_M\|_1$. Combining the bounds, we get

$\|\mathbf{P}\|_1 \leq \|\mathcal{P}_M\mathbf{P}\mathcal{P}_M\|_1 + \|\mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M + \mathcal{P}_M\mathbf{P}\mathcal{P}_M^{\perp} + \mathcal{P}_M^{\perp}\mathbf{P}\mathcal{P}_M^{\perp}\|_1 \leq \frac{2}{1-\rho}\|\mathcal{P}_M\mathbf{P}\mathcal{P}_M\|_1 \leq \frac{4}{1-\rho}\|\hat{\mathbf{W}} - \mathbf{W}_0\|_1$.

Since $1-\rho$ is of the form $\frac{\alpha}{SRF^4}$, the upper bound is of the form $C\frac{\|\mathbf{w}_0\|_2^2\gamma_{max}^2}{\gamma_{min}^3}RM^4SRF^4\eta$.