Phase retrieval in the high-dimensional regime

Milad Bakhshizadeh

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

Abstract

Phase retrieval in the high-dimensional regime

Milad Bakhshizadeh

The main focus of this thesis is on the phase retrieval problem. This problem has a broad range of applications in advanced imaging systems, such as X-ray crystallography, coherent diffraction imaging, and astrophotography. Thanks to its broad applications and its mathematical elegance and sophistication, phase retrieval has attracted researchers with diverse backgrounds.

Formally, phase retrieval is the problem of recovering a signal $\boldsymbol{x} \in \mathbb{C}^n$ from its phaseless linear measurements of the form $\left|\boldsymbol{a}_i^* \boldsymbol{x}\right| + \epsilon_i$, where sensing vectors $\boldsymbol{a}_i,\ i = 1, 2, \ldots, m$, are in the same vector space as $\boldsymbol{x}$ and ϵ_i denotes the measurement noise. Finding an effective recovery method in a practical setup, analyzing the required sample complexity and convergence rate of a solution, and discussing the optimality of a proposed solution are some of the major mathematical challenges that researchers have tried to address in the last few years.

In this thesis, our aim is to shed some light on some of these challenges and propose new ways to improve the imaging systems that have this problem at their core. Toward this goal, we focus on the high-dimensional setting where the ratio of the number of measurements to the ambient dimension of the signal remains bounded. This regime differs from the classical asymptotic regime in which the signal's dimension is fixed and the number of measurements is increasing. We obtain sharp results regarding the performance of the existing algorithms and the algorithms that are introduced in this thesis. To achieve this goal, we first develop a few sharp concentration inequalities. These inequalities enable us to obtain sharp bounds on the performance of our algorithms. We believe such results can be useful for researchers who work in other research areas as well. Second, we study the spectrum of some of the random matrices that play important roles in the phase retrieval problem, and use our tools to study the performance of some of the popular phase retrieval recovery schemes. Finally, we revisit the problem of structured signal recovery from phaseless measurements. We propose an iterative recovery method that can take advantage of any prior knowledge about the signal that is given as a compression code to efficiently solve the problem. We rigorously analyze the performance of our proposed method and provide extensive simulations to demonstrate its state-of-the-art performance.

Table of Contents

Acknowledgements

During my Ph.D. program, I had the privilege of meeting and interacting with several influential individuals without whose support I wouldn't have exceeded this challenging path. Their company has not only made every moment pleasant, but has also taught me many valuable lessons. I would like to take this chance to thank all those who assisted me in finding this amazing path and supported me throughout.

My first and foremost thanks go to my adviser, Professor Arian Maleki. The advice he gave me and the plans he suggested helped me develop valuable research skills. I have enjoyed and learned a lot from my discussions with Arian. I admire Arian's ability during discussions not only to think deeply but also to summarize it from different perspectives, and once he has done so I think I have a deeper understanding of the situation. Therefore I was always able to agree with Arian's suggestions. Furthermore, Arian has considerable insight into subject areas he has thought about and can express his views clearly in a discussion. This has made our conversations even more insightful for me. I would like to express my sincere gratitude for such wise conversations that Arian has generously offered me.

I would like to thank my co-advisor Professor Victor H. de la Pena who always has kindly supported me. I also would like to thank the committee members of my thesis, Professors Daniel Hsu, Predrag R. Jelenkovic, Cynthia Rush, and Ming Yuan, who accepted my invitation and gave me insightful comments to improve this thesis. My special thanks go to Cindy for chairing my defense session.

My sincere thanks go to Rishabh Dudeja who started as my officemate, and has since become a great friend and a trusted colleague. As a result of his nice and supportive attitude, he has become an invaluable friend over the years. Rishabh, thank you for all the help you offered me, particularly during the first challenging year of our Ph.D. program, and for all the enjoyment we shared together.

I would like to extend my warm thanks to all members of Columbia University with whom I had informative discussions and joyous times. I thank Professor Sumit Mukherjee for our discussions in probability theory. I am grateful to Professor John Wright who was a committee member of my oral examination and gave me valuable comments. I thank Jonathan Auerbach for his sarcastic conversations in all topics that came up, Florian Stebegg who has a very bright and clear mind and inspired me by his wise arguments, Promit Ghosal who is the most knowledgeable graduate student I have met, Wenda Zhou whose rapid speed of thinking and speaking makes it hard for me to follow him, but whenever I could follow I was fascinated by his intelligence, and Yixin Wang for her thoughtful recommendations in my first year, and Charles C. Margossian who always was making the loudest noise in the department and made the atmosphere cheerful. My deep gratitude goes to the administrative staffs of the Statistics Department, in particular to Dood Kalicharan and to Anthony Cruz who has always been welcoming with a warm smile. I appreciate all other members of the Statistics Departments who made my experience of the program more colorful. Through my interactions with them, I not only have developed a broader perspective, but also have made my Ph.D. journey joyful.

I thank my dear friends who made my times in New York city memorable. I appreciate Shervin, Sowmya, Amirhosein, Hasan, Amin, Zohreh, Sima and Hossein, Maryam and Ali, Jeanne and Kevin, Mahdieh and Hossein, Mohammad, Khadijeh, Aria, Bahar and Sadra, Sohrab, Fatemeh and Mohammad, Soroush, Alireza and all other friends with who I had happy times.

I deeply appreciate everyone who has inspired me to explore science. My genuine gratitude goes to Alireza Shaabani who has shown me the beauty of science when I was only a teenager. Undoubtedly, Alireza has been one of the best teachers I have ever had. My academic career would have been impossible had he not planted the seed of scientific reasoning in my mind. He was not only a teacher to me, but also has been a supportive friend and a wise adviser since we met. I would like to express my great thanks to Nasir Karimi who has helped me to develop a long-lasting interest in mathematics and problem solving and has taught me priceless life lessons.

Finally, my most profound gratitude goes to Mina, my beloved wife, whose unconditional love and support are the most precious gift I have ever had. My heartfelt thanks also go to my lovely mother and dearest siblings Mahta and Mohammad, whose love has always made my heart warm and kept me inspired in my most difficult moments. If it was not for my family's love and support I would not be here today. Thanks for everything, my dearests. I love you from the depths of my heart.

Dedication

This thesis is dedicated to two great women who are the most important and beloved ones in my life. First, I devote this dissertation to my beloved mother. She is the strongest person I have ever known, and I was blessed with her love and support since I was born. Obtaining high-level education is the manifestation of her dreams for me. I was so fortunate to always have her unconditional support and encouragement, in particular for education. I have consistently felt valued by her and she has sacrificed a lot to pave the way for my growth. I can never ever thank my mother enough for what she did for me. This dedication is only a humble effort to express my sincere respect and gratitude to her. Second, I would like to dedicate this work to my darling wife who has unconditionally loved and supported me since we met. If it was not for her support, I could never overcome the challenging years of the Ph.D. program. She is the greatest blessing of my life that has given me a reason to live, to fight, and to hope for a bright future. My dearest Mina, I dedicate this thesis to you with love from the bottom of my heart. Thanks for being there for me in all of my ups and downs and for your selfless support.

Notational Conventions

Notation	Description
$\mathbb{R}^n$	n-dimensional Euclidean space
$\mathbb{C}^n$	n-dimensional complex vector space
$\mathbb{O}_m$	Orthogonal group, set of all $m \times m$ orthogonal matrices
$\mathbb{U}_m$	Unitary group, set of all $m \times m$ unitary matrices
$\overline{A}$	closure of set A
$\boldsymbol{x}$	the signal of interest
$\boldsymbol{a}_i$	i^{th} sensing vectors
$\boldsymbol{A}$	$\begin{pmatrix} \boldsymbol{a}_1^* \\ \boldsymbol{a}_2^* \\ \ddots \\ \boldsymbol{a}_m^* \end{pmatrix}$, the sensing matrix
$\boldsymbol{a}^*($ or $\boldsymbol{A}^*)$	conjugate transposed of the vector $\boldsymbol{a}$ (or matrix $\boldsymbol{A}$)
$\boldsymbol{a}^T($ or $\boldsymbol{A}^T)$	transposed of the vector $\boldsymbol{a}$ (or matrix $\boldsymbol{A}$)
$\boldsymbol{A}_i$	i^{th} column of matrix $\boldsymbol{A}$
$A_{i,j}$	$(i, j)^{\text{th}}$ element of matrix $\boldsymbol{A}$
y_i	i^{th} observation, i^{th} element of vector $\boldsymbol{y}$
n	signal's dimension
m	number of observations
k	sparsity level
$\boldsymbol{I}($ or $\boldsymbol{I}_n)$	identity matrix (of size n)
$\boldsymbol{F}_n$	$n \times n$ DFT matrix
$\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle$	inner product, $\sum_{i=1}^n v_{1i} v_{2i}$
$\mathcal{T}$	trimming function

Notation	Description
X, Y, Z	random variables
$\mathbb{E}\left[\cdot\right]$	expectation of a random variable
$\mathbb{P}\left(\cdot\right)$	probability (of an event)
$\mathrm{Var}(\cdot)$	variance of a random variable
$\triangleq$	define
$\cdot \overset{d}{=} \cdot$	have the same distribution
$\xrightarrow{a.s.}$	almost sure convergence
$\xrightarrow{\mathrm{P}}$	convergence in probability
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and standard deviation σ
$\mathcal{CN}(0, 1)$	$\mathcal{N}(0, \frac{1}{2}) + \sqrt{-1}\mathcal{N}(0, \frac{1}{2})$
$\mathrm{Unif}(A)$	uniform distribution over set A
e	2.71828..., Euler's number
log	natural logarithm in base e
$\log_2$	logarithm in base 2
$\exp(\cdot)$	exponential function
$\mathrm{Re}(z)$	real part of complex number z
$\mathrm{Im}(z)$	imaginary part of complex number z
$\mathbb{I}(\cdot)$	indicator function
$\Gamma(t)$	$\int_0^\infty x^{t-1}\mathrm{e}^{-x}dx$
$f(t) \ll g(t)$	$\lim\limits_{t\to\infty} \frac{f(t)}{g(t)} = 0$
$\Phi(t)$	$\int_{-\infty}^{t} \frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx$
$\overline{\Phi}(t)$	$1 - \Phi(t) = \int_t^\infty \frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}dx$
$\mathrm{Diag}(d_1, d_2, \ldots, d_n)$	diagonal matrix with d_i as i^{th} diagonal element
$\mathsf{Tr}(\cdot)$	trace of a matrix
$\mathcal{E}_r$	encoder at rate r
$\mathcal{D}_r$	decoder at rate r
$\mathcal{C}_r$	code-words at rate r
$\mathcal{P}_{\mathcal{C}_r}$	projection on the set $\mathcal{C}_r$

Chapter 1: Introduction and Background

1.1 Phase Retrieval

1.1.1 Overview

Phase retrieval is an inverse problem that aims to recover a signal $\boldsymbol{x} \in \mathbb{C}^n$ from its phaseless measurements $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}| + \boldsymbol{\epsilon}$, where $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{\epsilon} \in \mathbb{R}^n$ denote the measurements and noise respectively. This problem appears in a variety of imaging systems, such as astrophotography [1], crystallography [2], and coherent diffraction imaging [3]. Applications of phase retrieval will be discussed in more details in Section 1.1.2. The study of phase retrieval problem dates back to the mid 20th century [4] in studying optical imaging systems [5]. Since then, an extensive literature has grown around this problem. An interested reader may refer to the surveys [4, 6, 7, 8] that review this large body of work.

Note that the lack of phase information in the measurements makes this problem more challenging than the classical linear model that appears in other imaging systems, such as magnetic resonance imaging. Hence, researchers have pursued two directions to obtain accurate estimates of the signal: (i) oversampling, and (ii) assuming or learning a certain structure about the signal of interest and incorporate that structure in the recovery algorithm. In this thesis, we study both directions.

One of the primary characteristics of the phase retrieval problem is having multiple (infinitely many) solutions. It is straightforward to confirm that if $\boldsymbol{x}$ satisfies $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|$, then so does $\mathrm{e}^{i\theta}\boldsymbol{x}$ for all $\theta \in \mathbb{R}$. This ambiguity makes it challenging for iterative methods that aim to converge to the solution by minimizing a cost function. When there are multiple solutions which form a non-convex set, no exact formulation of the problem would be convex. Moreover, in the region between multiple solutions some other stationary points may emerge which can slow down or disturb the convergence. To overcome this difficulty, two main techniques have been developed. The first one is to convexify the problem. It has been shown that under some conditions on the signal or the number of measurements the solution of the convex and the non-convex formulations coincide. However, convex methods require some technical adaption, such as lifting to much higher dimensional space, which make them unfeasible in most practical cases. The second method is to solve the original non-convex problem in a small vicinity of a solution which can exclude lots of irrelevant stationary points. The main challenge of the latter technique is to find such small neighborhood. Both of these techniques will be discussed in more details in the following sections of the current chapter.

1.1.2 Applications

Several well-known applications of phase retrieval are crystallography, coherent diffraction imaging, astrophtpgraphy, optics, and quantum mechanics. In this section, we briefly explain some of these applications.

X-ray crystallography is an imaging technique to obtain an image of a crystal structure. This technique was developed in early 1900s by discovering that the diffraction pattern of the X-ray beam is unique for each crystal [9, 10]. However, light detectors used in X-ray crystallography can only measure the intensity of the light ray. Hence, the problem of obtaining the phases from magnitude-only measurements emerged which is now called the phase retrieval problem.

The extension of X-ray crystallography for obtaining images of non-crystal structures was an important development in imaging systems which was achieved in the late 20th century [11]. This novel method of imaging is known as Coherent Diffraction Imaging (CDI). For the same reason as in X-ray crystallography, the phase retrieval problem is at the heart of CDI systems too. A similar technique can be used in electron microscopy which obtains images of higher resolution than what a microscope working with visible light can achieve [12].

Inability of detection devices, such as CCD cameras and photosensitive films, to record the phase of the measurements makes the phase retrieval problem a major component of many optical imaging systems. We refer to [7, 5] and the references therein for a detailed review of such applications.

In astrophotography, the resolution of images is limited by two main factors: First, the earth's atmosphere interferes with the light telescopes receive from far away objects. Second, the resolution of the obtained images is limited by the diameter of the telescope used for imaging. This diameter can become prohibitively large for a desired resolution. To increase the resolution of astronomical images a family of techniques, called interferometric imaging, are employed [1]. Similar to the previous cases, it is difficult in many cases to capture the phases of the measurements. We refer to [1] for the details on how the solution of phase retrieval can be beneficial in such imaging systems.

As our final example, we would like to note that the phase retrieval problem is closely related to Pauli's problem which is raised in the context of quantum physics [13]. Wolfgang Pauli asked initially whether a function can be uniquely identified by the modulus of itself and its Fourier transform. This problem and its extensions have attracted the attention of many researchers working in both theoretical and experimental physics. For further details on the connection of phase retrieval and Pauli's problem we refer the reader to the following articles and the references therein [13, 14, 15].

1.2 Convex Solutions

In this section, we briefly review some of the proposed solutions for the phase retrieval problem by taking advantage of the convex optimization tools.

A popular method to solve inverse problems is to define a loss function $d(\hat{\boldsymbol{x}}, y_1, ..., y_m)$ which measures the closeness of a candidate vector $\hat{\boldsymbol{x}} \in \mathbb{C}^n$ from the desired solution based on measurements $y_1, y_2, ..., y_m \in \mathbb{R}$. As should be expected, this loss function usually depends on the measurement matrix $\boldsymbol{A}$ as well. For making this dependency explicit we may denote it by $d_{\boldsymbol{A}}$.

Several relaxation techniques have been introduced that allow us to solve an easier computational problem at the expense of either the required number of measurements or the accuracy of the estimate. For instance, the authors of [16] lift the signal $\boldsymbol{x}$ to the matrix $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^*$ and solve a semidefinite programming (SDP) in this higher dimensional space. There is a convex relaxation in the formulation of this method which replaces the rank 1 constraint by trace minimization. Nevertheless, [16] proves given $m = \Omega(n \log n)$, the relaxed problem finds the exact solution when sensing vectors are uniformly sampled from the unit ball. Later, [17] improves this result by reducing the sample complexity to $m = \Omega(n)$. The authors of [18] focus on sparse real signals, and give a relaxed convex formulation for phase retrieval similar to [16]. This work proves that if the measurement matrix satisfies Restricted Isometry Property (RIP) [19], then the sparsest solution of the original problem and the relaxed problem are the same. Another SDP-based approach is proposed in [20], called PhaseCut. This work explicitly splits the amplitude and the phase of the signal and formulates an optimization problem which seeks for the most optimal phase. While many similarities between [16] and PhaseCut are discussed in [20], the main goal of this work is to offer a simpler expression for the constraints of the optimization problem they aim to solve. This simpler structure enables authors to employ a block coordinate descent algorithm whose iterations are less demanding than the update rules of [16]. The SDP-based methods suffer from a common disadvantage which is the requirement of lifting the signal to much higher dimension than its natural dimension. This avoids one to scale the proposed techniques for high-dimensional vectors and make them impractical for some realistic settings. To address this issue, [21, 22] appeared almost at the same time, and suggested a convex relaxation of the phase retrieval that does not require lifting, called PhaseMax. Both works offer recovery guarantees with different setups for the signal, sensing vectors, noise distribution, and sample complexity. This approach is computationally more affordable than SDP-based methods. However, the success of PhaseMax heavily depends on having access to an initial point which has strong correlation with the signal of interest. Finding such a good initial point is usually the most challenging part in utilizing PhaseMax.

1.3 Non-Convex Solutions

Despite the major developments of the convex formulation for phase retrieval, the non-convex algorithms have remained the most popular algorithms in applications. This is partly due to their simplicity and partly due to their superior performances. Gerchberg-Saxton algorithm [23] is one the earliest solutions proposed to solve the phase retrieval problem which was improved later by Fienup [24]. The main idea of these algorithms is to alternatively project a candidate vector on the image of the sensing matrix and the set of vectors whose magnitude matches with the given measurements. Since the latter set on which the projection is made is non-convex, we have a non-convex approach in this setup. While Gerchberg-Saxton algorithm is easily scalable and offers a great performance in practice, no theoretical guarantee for its convergence was shown. To overcome this issue, [25] proposes a non-convex algorithm, called Wirtinger Flow which provably converges to the right solution. In this work, authors suggest using gradient descent for minimizing a quadratic function of measurements which yields an estimate for the signal. It is shown in [25] that for a Gaussian sensing matrix and for sufficiently large number of measurements, Wirtinger Flow converges to the signal of interest with an overwhelming probability. In order to improve the performance of the recovery algorithm, [26] proposes a similar optimization problem based on an objective function which is slightly different from the one used in [25], called Truncated Amplitude Flow (TAF). The authors of [26] prove theoretical guarantee for TAF and compare its performance with many preceding methods thorough extensive simulations. There are several other works that adapted the ideas of the above works to take the prior knowledge about the signal into account and improve the performance of the recovery method [27]. We will discus methods for structured signals in section 1.4.

1.3.1 Initialization

When the objective function is non-convex, it is possible to have multiple local minimas and saddle points which are far from the desired signal. Hence, the point from which an iterative algorithm starts has a major impact on the final solution. Usually, there is a vicinity of the true signal in which no other stationary point exits. Finding an initial point in that vicinity, however, is a challenging task.

The authors of [28, 25] considered the leading eigenvector of the following matrix as an initial estimate:

$$\boldsymbol{M} = \frac{1}{m}\sum_{i=1}^{m}\mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^*.$$

Here, $y_i, \boldsymbol{a}_i$ are the i^{th} observation and sensing vector, respectively, and $\mathcal{T}$ is a trimming function which is chosen in a way to yield the best performance. Such an initialization is called the spectral method. In [25], the authors have shown that when $\boldsymbol{a}_i$ is sampled from a Gaussian distribution and $\mathcal{T} = id$ is the identity function we have

$$\mathbb{E}\left[\boldsymbol{M}\right] = \boldsymbol{I} + 2\boldsymbol{x}\boldsymbol{x}^*,$$

where $\boldsymbol{x}$ is the signal of interest, $\boldsymbol{I}$ denotes the identity matrix, and * denotes the Hermitian operator. Note that $\boldsymbol{x}$ is the leading eigenvector of $\mathbb{E}\left[\boldsymbol{M}\right]$ with the corresponding eigenvalue being equal to 3. Hence, if $\boldsymbol{M}$ concentrates around its mean, it makes sense to utilize the spectral method for initialization. It has been shown in [25] that with $m = \Omega(n \log n)$, the output of the spectral method is good enough to yield the right signal when it is passed to an iterative algorithm called Wirtinger Flow. Later, [29, 30, 31, 32, 33] studied the performance of the spectral method in more general settings and determined the optimal trimming function in some setups.

In Chapters 3 and 4 we provide very sharp analysis of spectral methods for several types of measurement matrices that are very popular in application.

1.3.2 Iterative Convergence

An iterative method starts from an initial point $\boldsymbol{x}_0$ and updates each iteration based on a rule $\boldsymbol{x}_n = g(\boldsymbol{x}_0, ..., \boldsymbol{x}_{n-1})$ in a way that it will eventually either converge to the true signal $\boldsymbol{x}$ or a point in its close proximity. The pioneer Gerchberg-Saxton algorithm [23], the Wirtinger Flow [25] (WF), and the Amplitude Flow [26] (AF) have proposed different iterations to reach this goal. While Gerchberg-Saxton algorithm utilizes an alternating projection idea, Wirtinger Flow and Amplitude Flow use the 'gradient descent' for the following objective functions:

$$d_{WF}(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^{m}\left(\left|\boldsymbol{a}_i^*\boldsymbol{u}\right|^2 - \left|\boldsymbol{a}_i^*\boldsymbol{x}\right|^2\right)^2, \quad d_{AF}(\boldsymbol{u}) = \frac{1}{m}\sum_{i=1}^{m}\left(\left|\boldsymbol{a}_i^*\boldsymbol{u}\right| - \left|\boldsymbol{a}_i^*\boldsymbol{x}\right|\right)^2.$$

As discussed in Section 1.3.1, starting from a sufficiently close initial point, each of the above iterative methods are guaranteed to converge to the right signal up to a global phase. Finding new iterative methods for solving the phase retrieval problem is still an active area of research [8]. Furthermore, some theoretical aspects of this problem are yet to be understood. We will discuss a few of such problems that are more related to the discussions of this thesis in Chapter 6.

1.4 Structured Data

Considering some structure on the signal (based on our prior knowledge) to improve the performance, has became a major component of many imaging systems. Employing such information may speed up the convergence rate, reduce the required sample complexity, and increase the accuracy of the estimate. Sparsity structure has been studied extensively in this context. A signal $\boldsymbol{x}$ is called k-sparse if it has at most k non-zero components, i.e. $\|\boldsymbol{x}\|_{\ell_0} \leq k$. Several articles have considered the phase retrieval problem for sparse signals [34, 35, 36, 37, 38, 39]. They include a variety of regularizers in the objective function, such as ℓ_1 penalty [39] and total variation [40], or include a truncation [34, 26] or projection step [24, 41] in the iterations. While it has been shown that $O(k \log n)$ measurements are sufficient to converge to a $k-$sparse signal if we are in its vicinity [42], all proposed iterative algorithms need $O(k^2 \log n)$ measurements to find a proper initialization. In a recent line of study, the sparse signal is replaced by the image of a generative model whose input lies on a k-dimensional space [43]. In this setup, a linear sample complexity in k, i.e. $O(k \log n)$ seems to be sufficient for a guaranteed recovery. Nevertheless, training a generative model requires thousands of samples from the distribution of interest which may not be available in some cases, in particular for cutting-edge imaging systems. Moreover, minimizing the value of k is not a priority of generative models since it may lead to less capacity of the model to learn the desired distribution. Therefore, such an approach might not be suitable to achieve the optimal sample-complexity in practice.

Chapter 5 of our thesis is dedicated to a general formulation of the structure based on compression codes. We refer the reader to this Chapter and the references therein for a more detailed review of the literature in phase retrieval for structured signals.

1.5 Organization of the following chapters

In this thesis, we study the phase retrieval problem in the high-dimensional settings. We consider the high-dimensional regime where the ratio of the number of measurements m to the ambient dimension of the signal n remains bounded. In this regime, many classical asymptotic results that are based on $\frac{n}{m} \to 0$ fail. Hence, new tools and analysis strategies are required for studying this problem. Therefore, we create new theoretical and practical tools and platforms to develop recovery methods and analyze phase retrieval algorithms. Below, we provide more details.

Chapter 2: In this chapter we obtain concentration and large deviation for the sums of independent and identically distributed random variables with heavy-tailed distributions. Our concentration results are concerned with random variables whose distributions satisfy $\mathbb{P}(X > t) \leq \mathrm{e}^{-I(t)}$, where $I : \mathbb{R} \to \mathbb{R}$ is an increasing function and $I(t)/t \to \alpha \in [0, \infty)$ as $t \to \infty$. Our main theorem can not only recover some of the existing results, such as the concentration of the sum of subWeibull random variables, but it can also produce new results for the sum of random variables with heavier tails. We show that the concentration inequalities we obtain are sharp enough to offer large deviation results for the sums of independent random variables as well. Our analyses which are based on standard truncation arguments simplify, unify and generalize the existing results on the concentration and large deviation of heavy-tailed random variables. This chapter is based on results published in [44].

Chapter 3: The success of iterative local search algorithms in phase retrieval depends heavily on their starting points. The most widely used initialization scheme is the spectral initialization, in which the eigenvector corresponding to the largest eigenvalue of a data-dependent matrix is used as a starting point. Recently, the performance of the spectral initialization was characterized accurately for measurement matrices with independent and identically distributed entries. This chapter aims to obtain the same level of knowledge for the partial orthogonal matrices, which are substantially better models for practical phase retrieval systems. Towards this goal, we consider the asymptotic setting in which the number of measurements m, and the dimension of the signal, n, diverge to infinity while $m/n \to \delta$, and obtain simple expression for the overlap between the initialization and the true signal. Results of this chapter have been published in [32].

Chapter 4: In the phase retrieval problem one seeks to recover an unknown n dimensional signal vector $\boldsymbol{x}$ from m measurements of the form $y_i = |(\boldsymbol{Ax})_i|$, where $\boldsymbol{A}$ denotes the sensing matrix. A popular class of algorithms for this problem are based on approximate message passing. For these algorithms, it is known that if the sensing matrix $\boldsymbol{A}$ is generated by sub-sampling n columns of a uniformly random (i.e. Haar distributed) orthogonal matrix, in the high dimensional asymptotic regime ($m, n \to \infty, n/m \to \kappa$), the dynamics of the algorithm are given by a deterministic recursion known as the state evolution. For a special class of linearized message passing algorithms, we show that the state evolution is universal: it continues to hold even when $\boldsymbol{A}$ is generated by randomly sub-sampling columns of certain deterministic orthogonal matrices such as the Hadamard-Walsh matrix, provided the signal is drawn from a Gaussian prior. This chapter is based on [33].

Chapter 5: Compressive phase retrieval refers to the problem of recovering a structured n-dimensional complex-valued vector from its phase-less under-determined linear measurements. The non-linearity of the measurement process makes designing theoretically-analyzable efficient phase retrieval algorithms challenging. As a result, to a great extent, existing recovery algorithms only take advantage of simple structures such as sparsity and its convex generalizations. The goal of this chapter is to move beyond simple models through employing compression codes. Such codes are typically developed to take advantage of complex signal models to represent the signals as efficiently as possible. In this chapter, it is shown how an existing compression code can be treated as a black box and integrated into an efficient solution for phase retrieval. First, COmpressive PhasE Retrieval (COPER) optimization, a computationally-intensive compression-based phase retrieval method, is proposed. COPER provides a theoretical framework for studying compression-based phase retrieval. The number of measurements required by COPER is connected to k, the α-dimension (closely related to the rate-distortion dimension) of a given family of compression codes. To find the solution of COPER, an efficient iterative algorithm called gradient descent for COPER (GD-COPER) is proposed. It is proven that under some mild conditions on the initialization and the compression code, if the number of measurements is larger than $Ck^2 \log^2 n$, where C is a constant, GD-COPER obtains an accurate estimate of the input vector in polynomial time. In the simulation results, JPEG2000 is integrated in GD-COPER to confirm the state-of-the-art performance of the resulting algorithm on real-world images. These results have been published in [41].

Chapter 6: Finally, Chapter 6 is devoted to some open problems.

Chapter 2: Concentration of Heavy tailed distribution

2.1 Introduction

The concentration of measure inequalities have recently received substantial attention in high-dimensional statistics and machine learning [45]. While concentration inequalities are well-understood for subGaussian and subexponential random variables, in many application areas, such as signal processing [46], machine learning [47] and optimization [48] we need concentration results for sums of random variables with heavier tails. The standard technique, i.e. finding upper bounds for the moment generating function (MGF), clearly fails for heavy-tailed distributions whose moment generating functions do not exist. Furthermore, other techniques, such as Chebyshev's inequality, are incapable of obtaining sharp results. The goal of this chapter is to show that under quite general conditions on the tail, a simple truncation argument can not only help us use the standard MGF argument for heavy-tailed random variables, but is also capable of obtaining sharp concentration results.

The problem of finding sharp concentration inequalities dates back to 1970s [49, 50, 51, 52, 53, 54]. For instance, [49] discusses several inequalities for finite sums of independent random variables with variety of tail decays [49]. The proof techniques of the present chapter have a similar flavor to what is used in [49]; we also use the truncation of random variables and bound the MGF of the truncated random variables. The generality of the inequalities presented in [49] in terms of the truncation levels, the moments of random variables, etc., makes the results difficult to use and interpret. In particular, obtaining the optimal choice of the parameters that appear in different upper bounds and simplifying the expressions for a given set of parameters is a time-consuming and cumbersome task. Compared to [49], we only consider the sum of random variables with bounded variances. For this class of distributions we are able to find the optimal truncation level. Using this right truncation level, we have been able to reduce the problem of obtaining sharp concentration results to that of finding an upper bound for the expectation of a smooth function of the distribution of individual random variables. We have also offered several insights on the quantity that is involved in our upper bound. As a result, our concentration results, while less general than [49], are much more interpretable and the calculations that are involved in them can be easily carried out. Despite the simpler form of our results, as we show through large deviation, they are still sharp. We should also emphasize that there have been more follow-up researches [50, 51, 52, 53, 54] that have appeared after [49]. These works suffer from similar issues as the ones we discussed about [49]. For instance, the bounds in [50] are written in terms of the solutions of some optimization problems which are not easily solvable for most distributions of interest.

Recently, a few papers have considered specific classes of distributions with heavier tails than exponential and obtained extensions of the classical concentration results. In [55], the authors consider the class of subWeibull variables, and prove a concentration inequality for the sum of independent random variables from this class by leveraging a novel Orlicz's norm. The first advantage of the approach proposed in this chapter compared to [55] is its generality; our approach is not tailored to the form of the tail. As a result, we have been able to consider much more generic tail decays, compared to [55], and study the effect of the tail-decay on the concentration. Furthermore, the same approach gives sharp large deviation bounds that shows the sharpness of the exponents that we obtain in our concentration results.

Although the result is novel and very useful for the extended class of distributions, it looks a bit overwhelming by lots of technical terms due the complexity of the norm they have used in the proof, hence in addition to the main result a ready to use inequality is also offered in this article. [47] also considers the class of subWeibull variables and shows how they naturally appear in the context of neural networks. This article offers some relations on the algebra of variables in this class between the algebraic operators and the tail. However, it does not study the concentrations with so much technical depth. These recent results usually lack the lower bound inequalities to check the sharpness of the offered tool.

To explore the accuracy of our concentration approach, we use our technique to obtain large deviation results. Not surprisingly, the tools we offer for our concentration results are also able to obtain the large deviation results that are consistent with the existing literature on the large deviation behavior of sums of independent, heavy-tailed random variables [56, 57, 58, 59, 60, 61, 62].

2.2 Our main contributions

2.2.1 Concentration

First, we discuss our concentration results for heavy-tailed distributions. Let us start with the following definition.

Definition 1. *Let $I : \mathbb{R} \to \mathbb{R}$ denote an increasing function. We say I captures the right tail of random variable X if*

$$\mathbb{P}\left(X > t\right) \leq \exp\left(-I(t)\right), \quad \forall t > 0. \tag{2.1}$$

Note that for the moment $I(t)$ can be a generic function. However, as we will see later, in our theorems we will impose some constraints on $I(t)$. Clearly, $I_{br}(t) = -\log \mathbb{P}\left(X > t\right)$ captures the right tail of X for any random variable X. We call $I_{br}(t)$ the basic rate capturing function. One can use $I_{br}(t)$ in our concentration results. However, as will be discussed later, it is often more convenient to approximate this basic tail capturing function.

Given a sequence of independent and identically distributed random variables $X_1, X_2, \dots, X_m$ with $\mathbb{E}\left[X_i\right] < \infty$, the goal of this chapter is to study

$$\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right),$$

where $S_m = \sum\limits_{i=1}^{m} X_i$. Based on the definition of the rate capturing function we state our concentration result. In the rest of the chapter, we use the notation X^L to denote the truncated version of the random variable X, i.e.,

$$X^L = X\mathbb{I}(X \leq L).$$

Theorem 1 (General Concentration). *Suppose $X_1, ..., X_m \overset{d}{=} X$ are independent and identically distributed random variables whose right tails are captured by an increasing and continuous function $I : \mathbb{R} \to \mathbb{R}^{\geq 0}$ with the property $I(t) = O(t)$ as $t \to \infty$. Define $Z^L \triangleq X^L - \mathbb{E}[X]$. Moreover, for $\beta \in (0, 1]$, $L > 0$, and $\lambda = \beta \frac{I(L)}{L}$, define*

$$c_{L,\beta} \triangleq \mathbb{E}\left[\left(Z^L\right)^2 \mathbb{I}\left(Z^L \leq 0\right) + \left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right) \mathbb{I}\left(Z^L > 0\right)\right]. \tag{2.2}$$

Finally, define $t_{\max}(\beta) \triangleq \sup\left\{t \geq 0 : \ t \leq \beta c_{mt,\beta} \frac{I(mt)}{mt}\right\}$.[1] *Then,*

$$\mathbb{P}\left(S_m - \mathbb{E}[S_m] > mt\right) \leq \begin{cases} \exp\left(-c_t \beta I(mt)\right) + m \exp\left(-I(mt)\right), & t \geq t_{\max}(\beta), \\ \exp\left(-\dfrac{mt^2}{2c_{mt_{\max},\beta}}\right) + m \exp\left(-\dfrac{mt_{\max}(\beta)^2}{\beta c_{mt_{\max},\beta}}\right), & 0 \leq t < t_{\max}(\beta), \end{cases} \tag{2.3}$$

where c_t is a constant between $\frac{1}{2}$ and 1. More precisely, $c_t = 1 - \frac{1}{2}\frac{\beta c_{mt,\beta}}{t}\frac{I(mt)}{mt}$.

The proof of this theorem can be found in Section 2.4. Note that the concentration result we obtain is similar to the concentration results that exist for sub-exponential random variables; there is a region for t in which the distribution of the sum looks like a Gaussian, and a second region in which the sum has heavier tail than a Gaussian. We will apply our theorem to some popular examples, including the subexponential distributions later. Before that, let us discuss some of the main features of this theorem.

Remark 1. *As is clear from the proof of Theorem 1 one can replace $c_{mt,\beta}$ with an upper bound. In other words, if $c_{mt,\beta} \leq c$, then Theorem 1 remains valid by replacing $c_{mt,\beta}$ with c in the definition of $t_{\max}$ and the coefficients appeared in* (2.3).

Obtaining an accurate upper bound for $c_{L,\beta}$ is a key to using Theorem 1 for different applications. Since, we are often interested in the behavior of $c_{mt,\beta}$ for large values of mt, it is usually instructive to understand the behavior of $c_{L,\beta}$ for large values of L. Suppose that there exists a function $g(X)$

[1] We set $t_{\max} = 0$, when the set is empty.

such that

$$\left|\left(Z^L\right)^2 \mathbb{I}\left(Z^L \leq 0\right) + \left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right) \mathbb{I}\left(Z^L > 0\right)\right| < g(X),$$

and that $\mathbb{E}\left[g(X)\right] < \infty$. Further, assume that $I(L) = o(L)$. Then, from the dominated convergence theorem we have

$$\limsup_{L\to\infty} \mathbb{E}\left[\left(Z^L\right)^2 \mathbb{I}\left(Z^L \leq 0\right) + \left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right) \mathbb{I}\left(Z^L > 0\right)\right] = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right].$$

Hence, if the random variables have bounded variances, then we expect $c_{L,\beta} < \infty$ for all values of L. If we replace $c_{mt,\beta}$ in Theorem 1 with a fixed number, then the statement of the theorem becomes simpler. Note that this argument is based on an asymptotic argument and is not particularly useful when we want to derive concentration bounds. Hence, our next few lemmas obtain simpler integral forms for $\mathbb{E}\left[\left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right) \mathbb{I}\left(Z^L > 0\right)\right]$.

Lemma 1. *Let $Z^L = X^L - \mathbb{E}\left[X\right]$, and $I_{br}(t) = -\log \mathbb{P}\left(X > t\right)$ denote the basic tail capturing function. Then,*

$$\mathbb{E}\left[\left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right) \mathbb{I}\left(Z^L > 0\right)\right] = \int_0^{L-\mathbb{E}[X]} \exp\left(\lambda t - I_{br}(t + \mathbb{E}\left[X\right])\right)\left(2t + \lambda t^2\right) dt.$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}\left[\left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right) \mathbb{I}\left(Z^L > 0\right)\right] \\
&= \int_0^{\infty} \mathbb{P}\left(\left(Z^L\right)^2 \exp\left(\lambda Z^L\right) > u, Z^L > 0\right) du \\
&\overset{(*)}{=} \int_0^{L-\mathbb{E}[X]} \mathbb{P}\left(X > t + \mathbb{E}\left[X\right]\right) du, \\
&= \int_0^{L-\mathbb{E}[X]} \exp\left(-I_{br}(t + \mathbb{E}\left[X\right])\right)\left(2t + \lambda t^2\right) \exp\left(\lambda t\right) dt \\
&= \int_0^{L-\mathbb{E}[X]} \exp\left(\lambda t - I_{br}(t + \mathbb{E}\left[X\right])\right)\left(2t + \lambda t^2\right) dt.
\end{aligned}$$

In in the equation tagged by $(*)$, we applied the following change of variable: $t^2 \exp\left(\lambda t\right) = u$.

□

We can use the integral expression we derived in Lemma 1, and the specific properties of the rate function that we have, to obtain simpler upper bounds for $\mathbb{E}\left[\left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right)\mathbb{I}\left(Z^L > 0\right)\right]$. The following simple lemma is an upper bound we will use in our examples.

Lemma 2. *Suppose that $\frac{I(t)}{t}$ is a nonincreasing function, and let $\lambda = \frac{\beta I(L)}{L}$ Then,*

$$
\begin{aligned}
&\mathbb{E}\left[\left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right)\mathbb{I}\left(Z^L > 0\right)\right] \\
\leq\ & \exp\left(-\beta\mathbb{E}\left[X\right]\frac{I(L)}{L}\right)\int_0^{L-\mathbb{E}[X]} \exp\left(-(1-\beta)I(t+\mathbb{E}\left[X\right])\right)\left(2t+\beta\frac{I(L)}{L}t^2\right)dt \\
\leq\ & \exp\left(-\beta\mathbb{E}\left[X\right]\frac{I(L)}{L}\right)\int_0^{L-\mathbb{E}[X]} \exp\left(-(1-\beta)I(t+\mathbb{E}\left[X\right])\right)\left(2t+\beta t I(t)\right)dt.
\end{aligned}
$$

Proof. Similar to the proof of Lemma 1, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right)\mathbb{I}\left(Z^L > 0\right)\right] \\
&\leq \int_0^{L-\mathbb{E}[X]} \exp\left(-I(t+\mathbb{E}\left[X\right])\right)\left(2t+\lambda t^2\right)\exp\left(\beta\frac{I(L)}{L}t\right)dt \\
&\leq \exp\left(-\beta\mathbb{E}\left[X\right]\frac{I(L)}{L}\right)\int_0^{L-\mathbb{E}[X]} \exp\left(-(1-\beta)I(t+\mathbb{E}\left[X\right])\right)\left(2t+\beta\frac{I(L)}{L}t^2\right)dt \\
&\leq \exp\left(-\beta\mathbb{E}\left[X\right]\frac{I(L)}{L}\right)\int_0^{L-\mathbb{E}[X]} \exp\left(-(1-\beta)I(t+\mathbb{E}\left[X\right])\right)\left(2t+\beta t I(t)\right)dt,
\end{aligned}
$$

where to obtain the last two inequalities we used the fact that $\frac{I(t)}{t}$ is a nonincreasing function. □

We will later show how combining Theorem 1 and Lemma 2 leads to sharp concentration results for some well-known tail capturing functions. It is straightforward to see that as long as $(1-\beta)I(t+\mathbb{E}\left[X\right]) > 2a\log t$ for some $a > 1$, the upper bound given by Lemma 2 remains bounded even when $L \to \infty$. Hence, we can use these upper bounds for a broad range of tail decays. We will discuss this in more details at the end of this section.

Remark 2. *Note that Theorem 1 considers the case where $I(L) = O(L)$. The other cases, i.e.*

$I(L) = \Omega(L)$, can be studied using standard arguments based on the moment generating function and hence, are not explored in this chapter. We will later emphasize that this condition is not enough for the usefulness of Theorem 1. For instance, if the tail is too heavy then $c_{L,\beta}$ will be infinite. We will discuss this issue in more details later.

Let us now show how Theorem 1 can be used in a few concrete examples which are popular in application areas. Our first example considers the well-studied class of subexponential distributions.

Corollary 1. *Let $I(t) = kt$ for some fixed coefficient k. Then, for all $\beta \in (0,1)$ and $L > \mathbb{E}[X]$ we have*

$$c_{L,\beta} \leq \mathbb{E}\left[(X - \mathbb{E}[X])^2 \mathbb{I}(X \leq \mathbb{E}[X])\right] + \frac{1}{(1-\beta)^3} \frac{2}{k^2 \exp\left(k\mathbb{E}[X]\right)} = c_\beta. \tag{2.4}$$

Hence for $m > \frac{\mathbb{E}[X]}{\beta c_\beta k}$,

$$\mathbb{P}\left(S_m - \mathbb{E}[S_m] > mt\right) \leq \begin{cases} \exp\left(-c_t \beta k m t\right) + m \exp\left(-kmt\right), & t \geq \beta c_\beta k, \\ \exp\left(-\frac{1}{2c_\beta} m t^2\right) + m \exp\left(-\beta c_\beta k^2 m\right), & 0 \leq t < \beta c_\beta k, \end{cases} \tag{2.5}$$

where $c_t = 1 - \frac{1}{2}\frac{\beta c_\beta k}{t}$.

Proof. We would like to use Theorem 1 for proving the concentration. Toward this goal, we use Lemma 2 to obtain an upper bound for $c_{L,\beta}$. First note that, $\lambda = \beta \frac{I(L)}{L} = \beta k$. Hence, according to

Lemma 1 we have

$$\begin{aligned}
&\mathbb{E}\left[\left(Z^L\right)^2 \exp\left(\lambda\left(Z^L\right)\right)\mathbb{I}\left(Z^L > 0\right)\right] \\
\leq\ & \exp\left(-\beta k \mathbb{E}\left[X\right]\right)\int_0^{L-\mathbb{E}[X]} \exp\left(-(1-\beta)k(t+\mathbb{E}\left[X\right])\right)\left(2t+\beta k t^2\right)dt \\
\leq\ & \exp\left(-\beta k \mathbb{E}\left[X\right]\right)\int_0^{\infty} \exp\left(-(1-\beta)k(t+\mathbb{E}\left[X\right])\right)\left(2t+\beta k t^2\right)dt \\
=\ & \exp\left(-k \mathbb{E}\left[X\right]\right)\int_0^{\infty} \exp\left(-(1-\beta)kt\right)\left(\beta k t^2 + 2t\right)dt \\
=\ & \exp\left(-k \mathbb{E}\left[X\right]\right)\left(\frac{\beta k}{(1-\beta)^3 k^3}\Gamma(3) + \frac{2}{(1-\beta)^2 k^2}\Gamma(2)\right) \\
=\ & \exp\left(-k \mathbb{E}\left[X\right]\right)\frac{2}{k^2}\frac{1}{(1-\beta)^3}.
\end{aligned}$$

We also have that if $L > \mathbb{E}\left[X\right]$, then

$$\mathbb{E}\left[\left(Z^L\right)^2 \mathbb{I}\left(Z^L \leq 0\right)\right] = \mathbb{E}\left[\left(X^L - \mathbb{E}\left[X\right]\right)^2 \mathbb{I}\left(X^L \leq \mathbb{E}\left[X\right]\right)\right] = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2 \mathbb{I}\left(X \leq \mathbb{E}\left[X\right]\right)\right].$$

□

Our next example considers subWeibull distributions.

Corollary 2. *Let X be a centered random variable, i.e. $\mathbb{E}\left[X\right] = 0$, whose tail is captured by $c_\alpha \sqrt[\alpha]{t}$ for some $\alpha \geq 1$. Moreover, assume $\mathbb{E}\left[X^2\mathbb{I}(X \leq 0)\right] = \sigma_-^2 < \infty$. Then, we have*

$$c_{L,\beta} \leq \sigma_-^2 + \frac{\Gamma(2\alpha+1)}{\left((1-\beta)c_\alpha\right)^{2\alpha}} + L^{\frac{1}{\alpha}-1}\frac{\beta c_\alpha \Gamma(3\alpha+1)}{3\left((1-\beta)c_\alpha\right)^{3\alpha}}.$$

Hence, one can apply Theorem 1 with the above bound. In this case, two regions of the concentration are separated by $t_{\max}(\beta) = \left(\beta c_{mt,\beta} c_\alpha\right)^{\frac{\alpha}{2\alpha-1}} m^{-\frac{\alpha-1}{2\alpha-1}}$.

Proof. Note that since $\alpha \geq 1$, $\frac{I(t)}{t}$ is indeed nonincreasing. We just need to apply Lemma 2 with

$I(t) = c_\alpha \sqrt[\alpha]{t}$ to obtain

$$\begin{aligned}
&\int_0^L \exp\left(-(1-\beta)c_\alpha \sqrt[\alpha]{t}\right)\left(2t + \beta c_\alpha L^{\frac{1}{\alpha}-1} t^2\right) dt \\
&\quad \leq \int_0^\infty \exp\left(-u\right)\left(\frac{2u^\alpha}{\left((1-\beta)c_\alpha\right)^\alpha} + \frac{\beta c_\alpha L^{\frac{1}{\alpha}-1} u^{2\alpha}}{\left((1-\beta)c_\alpha\right)^{2\alpha}}\right)\frac{\alpha u^{\alpha-1}}{\left((1-\beta)c_\alpha\right)^\alpha} du \\
&\quad = \frac{2\alpha}{\left((1-\beta)c_\alpha\right)^{2\alpha}}\Gamma(2\alpha) + \frac{\beta c_\alpha L^{\frac{1}{\alpha}-1}\alpha}{\left((1-\beta)c_\alpha\right)^{3\alpha}}\Gamma(3\alpha) \\
&\quad = \frac{\Gamma(2\alpha+1)}{\left((1-\beta)c_\alpha\right)^{2\alpha}} + L^{\frac{1}{\alpha}-1}\frac{\beta c_\alpha \Gamma(3\alpha+1)}{3\left((1-\beta)c_\alpha\right)^{3\alpha}}.
\end{aligned}$$

Finally, it is straightforward to note that

$$\mathbb{E}\left[(X^L)^2 \mathbb{I}(X^L \leq 0)\right] \leq \mathbb{E}\left[X^2 \mathbb{I}(X \leq 0)\right].$$

□

In our last example, we consider random variables with polynomially decaying tails.

Corollary 3. *Let X be a centered random variable, i.e. $\mathbb{E}\left[X\right] = 0$, whose tail is captured by $\gamma \log t$, where $\gamma > 2$. Moreover, assume $\mathbb{E}\left[X^2 \mathbb{I}(X \leq 0)\right] = \sigma_-^2 < \infty$. Then, we have*

$$c_{L,\beta} \leq \begin{cases} \sigma_-^2 + L^{\frac{\gamma\beta}{L}} + \dfrac{2 - \frac{\gamma\beta}{2-\gamma(1-\beta)}}{2-\gamma(1-\beta)}\left(L^{2-\gamma(1-\beta)} - 1\right) + \dfrac{\gamma\beta L^{2-(1-\beta)\gamma}\log L}{2-(1-\beta)\gamma}, & \beta \neq 1 - \dfrac{2}{\gamma}, \\ \sigma_-^2 + L^{\frac{\gamma-2}{L}} + 2\log L + \dfrac{\gamma-2}{2}\left(\log L\right)^2, & \beta = 1 - \dfrac{2}{\gamma}. \end{cases} \tag{2.6}$$

Proof. Note that

$$\begin{aligned}\mathbb{E}\left[\left(X^L\right)^2 \exp\left(\lambda X^L\right)\mathbb{I}(X^L>0)\right] &= \mathbb{E}\left[\left(X^L\right)^2 \exp\left(\lambda X^L\right)\mathbb{I}(0<X^L\le 1)\right]\\ &+ \ \mathbb{E}\left[\left(X^L\right)^2 \exp\left(\lambda X^L\right)\mathbb{I}(X^L>1)\right]\\ &\le \ \exp\left(\beta\gamma\frac{\log(L)}{L}\right) + \mathbb{E}\left[\left(X^L\right)^2 \exp\left(\lambda X^L\right)\mathbb{I}(X^L>1)\right]\\ &= \ L^{\frac{\gamma\beta}{L}} + \mathbb{E}\left[\left(X^L\right)^2 \exp\left(\lambda X^L\right)\mathbb{I}(X^L>1)\right].\end{aligned}$$

Thus, for $\beta \neq 1-\frac{2}{\gamma}$, using the upper bound given in Lemma 2, we just need to show

$$\int_1^L \exp\left(-(1-\beta)\gamma\log t\right)\left(2t+\beta\gamma t\log t\right)dt = \frac{2-\frac{\gamma\beta}{2-\gamma(1-\beta)}}{2-\gamma(1-\beta)}\left(L^{2-\gamma(1-\beta)}-1\right)+\frac{\gamma\beta L^{2-(1-\beta)\gamma}\log L}{2-(1-\beta)\gamma}. \tag{2.7}$$

Toward this goal, note that

$$\begin{aligned}\int_1^L \exp\left(-(1-\beta)\gamma\log t\right)\left(2t+\beta\gamma t\log t\right)dt &= \int_1^L t^{1-(1-\beta)\gamma}\left(2+\beta\gamma\log t\right)dt\\ &= \ \frac{t^{2-(1-\beta)\gamma}}{2-(1-\beta)\gamma}\left(2+\beta\gamma\left(-\frac{1}{2-(1-\beta)\gamma}+\log t\right)\right)\Bigg|_1^L\\ &= \ \frac{2-\frac{\gamma\beta}{2-\gamma(1-\beta)}}{2-\gamma(1-\beta)}\left(L^{2-\gamma(1-\beta)}-1\right)+\frac{\gamma\beta L^{2-(1-\beta)\gamma}\log L}{2-(1-\beta)\gamma}.\end{aligned}$$

In the above equality, we are using $\int t^k = \frac{1}{k+1}t^{k+1}$ and $\int t^k\log t = \left(-\frac{1}{(k+1)^2}+\frac{\log t}{k+1}\right)t^{k+1}$.

For $\beta = 1-\frac{2}{\gamma}$ we have $1-(1-\beta)\gamma = -1$ and $\gamma\beta = \gamma-2$. Hence

$$\begin{aligned}\int_1^L \exp\left(-(1-\beta)\gamma\log t\right)\left(2t+\beta\gamma t\log t\right)dt &= \int_1^L t^{-1}\left(2+(\gamma-2)\log t\right)dt\\ &= \ 2\log t+\frac{\gamma-2}{2}\left(\log t\right)^2\Bigg|_1^L = 2\log L+\frac{\gamma-2}{2}\left(\log L\right)^2,\end{aligned}$$

which concludes the proof. □

Remark 3. *Note that $\beta < 1 - \frac{2}{\gamma}$ is equivalent to $2 - (1-\beta)\gamma < 0$. Hence, the right hand side of* (2.6) *remains bounded when L grows to infinity. By letting β get closer to* 1 *we can cover any $\gamma > 2$. Hence, we can obtain a concentration inequality for the sum of independent and identical random variables with polynomially decaying tail as long as $P(X > t) < \frac{1}{t^\gamma}$ for some $\gamma > 2$.*

Let us try to find another bound for $c_{L,\beta}$ for the distributions we discussed in Corollaries 2 and 3. These bounds enable us to obtain another concentration result that is in some sense sharper than the one we derived above and shows the flexibility of our framework.

Lemma 3. *Suppose that* $\mathrm{var}(X) < \infty$ *and the right tails of random variables X is captured by $I(t)$. Suppose that $I(t)$ satisfies one of the following conditions:*

(a) $I(t) = I_\alpha(t) = c_\alpha \sqrt[\alpha]{t}$ for $\alpha > 1$ and $\beta < 1$,

(b) $I(t) = \gamma \log t$ for $\gamma > 2$ and $\beta < 1 - \frac{2}{\gamma}$.

Then, if we set $\lambda_{L,\beta} = \beta\frac{I(L)}{L}$ we have

$$\lim_{L\to\infty} \mathbb{E}\left[\left(X_L - \mathbb{E}\left[X\right]\right)^2 \left(\mathbb{I}\left(X^L \leq \mathbb{E}\left[X\right]\right) + \exp\left(\lambda_{L,\beta}\left(X_L - \mathbb{E}\left[X\right]\right)\right)\mathbb{I}\left(X^L > \mathbb{E}\left[X\right]\right)\right)\right] = \mathrm{Var}(X).$$

The proof of this lemma is presented in Section 2.4.2. This lemma implies that if L is large enough, then we should expect $c_{L,\beta}$ to be very close to $\mathrm{Var}(X)$. So, assuming mt is large enough we can obtain a more accurate concentration result.

Corollary 4. *Suppose that the right tails of independent and identically distributed random variables $X_1, X_2, \ldots, X_m$ are captured by $c_\alpha \sqrt[\alpha]{t}$ for $\alpha > 1$, and $Var(X_i) = \sigma^2$. Define $S_m = \sum\limits_{i \leq m} X_i$. Then, for any $0 < \beta < 1$ and $\epsilon > 0$, there is a constant C_ϵ such that for all $mt > C_\epsilon$*

$$\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right) \leq \begin{cases} \exp\left(-c_t\beta c_\alpha \sqrt[\alpha]{mt}\right) + m\exp\left(-c_\alpha \sqrt[\alpha]{mt}\right), & t > t_{\max}, \\ \exp\left(-\dfrac{mt^2}{2\left(\sigma^2+\epsilon\right)}\right) + m\exp\left(-\dfrac{mt_{\max}^2}{\beta(\sigma^2+\epsilon)}\right), & t \leq t_{\max}, \end{cases} \tag{2.8}$$

where $t_{\max} = \left(\beta(\sigma^2+\epsilon)c_\alpha\right)^{\frac{\alpha}{2\alpha-1}} m^{-\frac{\alpha-1}{2\alpha-1}}$ *and* $c_t = 1 - \frac{1}{2}\beta(\sigma^2+\epsilon)c_\alpha m^{\frac{1}{\alpha}-1} t^{\frac{1}{\alpha}-2}$ *varies between* $\frac{1}{2}$ *and* 1.

Proof. Note that, by Lemma 3, for any given $\epsilon > 0$ we can find a positive constant C_ϵ, such that

$$\mathbb{E}\left[\left(X_L - \mathbb{E}[X]\right)^2 \left(\mathbb{I}\left(X^L \leq \mathbb{E}[X]\right) + \exp\left(\lambda_{L,\beta}\left(X_L - \mathbb{E}[X]\right)\right)\mathbb{I}\left(X^L > \mathbb{E}[X]\right)\right)\right] \leq \sigma^2+\epsilon, \quad \forall L > C_\epsilon.$$

Hence, for all $mt > C_\epsilon$, Theorem 1 is applicable with $c_{L,\beta} = \sigma^2 + \epsilon$. The corollary follows by substituting this $c_{L,\beta}$ and $I(t) = c_\alpha \sqrt[\alpha]{t}$ in Theorem 1. □

Remark 4. *According to Corollary 4, if* $C_\epsilon < mt \leq mt_{\max}$*, then* $\mathbb{P}\left(S_m - \mathbb{E}[S_m] > mt\right)$ *is upper bounded by* $\exp\left(-\frac{mt^2}{2(\sigma^2+\epsilon)}\right) + m\exp\left(-\frac{mt_{\max}^2}{\beta(\sigma^2+\epsilon)}\right)$*. Note that* $\exp\left(-\frac{mt^2}{2(\sigma^2+\epsilon)}\right)$ *is very close to the term that appears in the central limit theorem. Furthermore, if* $mt > t_{\max}$*, then* $\mathbb{P}\left(S_m - \mathbb{E}[S_m] > mt\right)$ *is bounded from above by* $\exp\left(-c_t\beta c_\alpha \sqrt[\alpha]{mt}\right) + m\exp\left(-c_\alpha \sqrt[\alpha]{mt}\right)$*. Again we will show in the next section that this bound is sharp. Hence, an accurate bound for* $c_{L,\beta}$ *results in an accurate concentration result.*

Remark 5. *Using part (b) of Lemma 3, a corollary similar to Corollary 4 can be also written for* $I(t) = \gamma \log t$ *with* $\gamma > 2$*. For the sake of brevity, we do not repeat this corollary. Hence, Theorem 1 can be used to obtain concentration results as long as* $I(t) > \gamma \log t$ *with* $\gamma > 2$ *(for large enough values of* t*). Note that if* $I_{br}(t) = \gamma \log t$ *for* $\gamma < 2$*, then the variance of the random variable is unbounded. This is the region in which the sum of independent and identically distributed random variables does not converge to a Gaussian and it converges to other stable distributions (See Chapter 1 of [63]). We leave the study of the concentration of sums of such random variables to future research.*

2.2.2 Large deviation

In this section, as a simple byproduct of what we have proved for obtaining concentration bounds and also evaluating the sharpness of our results, we study the large deviation properties of the sums of independent and identically distributed random variables. Towards this goal, we consider the limiting version of Definition 1 in which the exact rate of decay of the tail is captured by $I(t)$.

Definition 2. *Let $I : \mathbb{R} \to \mathbb{R}$ denote an increasing function. We say I captures the right tail of random variable X in the limit if*

$$\lim_{t\to\infty} \frac{-\log\left(\mathbb{P}\left(X > t\right)\right)}{I(t)} = 1. \tag{2.9}$$

We say a random variable is super-exponential if its tail is captured in limit by a function I such that $I(t) = o(t)$ as $t \to \infty$.

Note that if the basic right tail capturing function satisfies $I_{br}(t) = o(t)$, then the moment generating function of the distribution is infinity for $\lambda \in (0, \infty)$. Hence, Cramer's theorem is not useful. Our next theorem offers a sharp large deviation result for superexponential random variables.

Theorem 2 (General Large Deviation)**.** *Suppose that $X_1, X_2, \ldots, X_m$ are super-exponential random variables with finite variance whose tails are captured in the limit by $I(t)$. Furthermore, suppose that I is an increasing function and $\lim_{t\to\infty} \frac{\log(t)}{I(t)} = 0$. Finally, let γ_m be an increasing sequence of real numbers that satisfy*

$$\log m \ll I(\gamma_m) \ll \frac{\gamma_m^2}{m}.^2 \tag{2.10}$$

If (2.2) *remains bounded for X_1 and for all $\beta < 1$, then*

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > \gamma_m\right)}{I(\gamma_m)} = 1. \tag{2.11}$$

The proof of this theorem is presented in Section 2.4.3. Again we use this theorem to obtain large deviation results for a few concrete examples.

[2] $f(t) \ll g(t)$ means that $f(t) = o(g(t))$ as $t \to \infty$.

Corollary 5. *Let the tail of independent and identically distributed random variables $X_1, X_2, \ldots, X_m$ be captured by $I(t) = c_\alpha \sqrt[\alpha]{t}$ in the limit, where $\alpha > 1$. Then, we have*

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right)}{\sqrt[\alpha]{mt}} = c_\alpha.$$

Proof. It suffices to choose $\gamma_m = mt$ and apply Theorem 2 with $I(t) = c_\alpha \sqrt[\alpha]{t}$. Note that

$$\log m \ll c_\alpha \sqrt[\alpha]{mt} \ll \frac{(mt)^2}{m} = mt^2, \tag{2.12}$$

for all $\alpha > 1$. □

Remark 6. *We should emphasize that the large deviation result for subWeibull distribution has been studied in the literature [64], [62]. Being able to answer this question for subWeibull distributions, although it is not novel, shows the strength of the results developed in this chapter. Note that even if t grows with m, as long as* (2.12) *is satisfied, i.e. $mt_m \gg m^{\frac{\alpha}{2\alpha-1}}$, we have*

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt_m\right)}{\sqrt[\alpha]{mt_m}} = c_\alpha.$$

On the other hand, it is known that if $mt_m \ll m^{\frac{\alpha}{2\alpha-1}}$, then the decay is characterized by $\overline{\Phi}\left(\frac{mt_m}{\sqrt{m\mathrm{Var}(X)}}\right)$, where $\overline{\Phi} = 1 - \Phi$, and Φ denotes the cumulative distribution function of a standard normal random variable [64]. According to Table 3.1 of [64] a similar result as the one presented in Corollary 5 has been known for $\gamma_m \gg m^{\frac{\alpha}{2\alpha-2}}$ when $0 \le \frac{1}{\alpha} \le \frac{1}{2}$. However as we discussed above, an extension of Corollary 5 fills the gap between $m^{\frac{\alpha}{2\alpha-2}}$ and $m^{\frac{\alpha}{2\alpha-1}}$, and shows that in this region still the tail of the sum behaves like the tail of the maximum.

Theorem 2 does not cover the polynomially-decaying tails. Hence, for the sake of completeness we discuss the polynomial example below.

Corollary 6. *Suppose X has zero mean and finite variance σ^2 and its right tail is captured by $I(t) = \alpha \log t$ for some $\alpha > 2$. For any sequence γ_m that satisfies any of the following conditions* *(i)* $\lim\limits_{m\to\infty} \frac{\log m}{\log \gamma_m} = k$ *for some* $k < 2$*,*

(ii) $\lim\limits_{m\to\infty} \frac{\log m}{\log \gamma_m} = 2$ *and* $\gamma_m \gg \sqrt{m \log m}$*,*

we have

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m > \gamma_m\right)}{I(\gamma_m) - \log m} = 1. \tag{2.13}$$

The proof can be found in Section 2.4.4.

Remark 7. *The result of Corollary 6 is known in the literature. For instance, the interested reader may refer to Proposition 3.1 in [64]). The main reason it is mentioned here is to show that this is also a simple byproduct of our main results in Section 2.2.1. Note that the conditions Corollary 6 imposes on the growth of* γ_m *cover all sequences that satisfy* $\gamma_m \gg \sqrt{m \log m}$ *(maybe after passing to a subsequence to make* $\lim \frac{\log m}{\log \gamma_m}$ *exist). For sequences that grow slower than* $\sqrt{m \log m}$ *the rate function for large deviations is not* $I(\gamma_m) - \log m$ *anymore [64].*

2.3 Discussion of the sharpness of Theorem 1

In this section, we would like to discuss that the bounds offered by Theorem 1 are sharp if compared with the limiting expressions obtained from the large deviation results. We clarify this point through the following two examples: Let $I(t)$ capture the right tail of a centered random variable X and also captures its right tail in the limit ($I_{br}(t)$ has this property). Assume that $X_1, ..., X_m$ are independent copies of X and $S_m = \sum\limits_{i\leq m} S_m$. Below we discuss the subWeibull distributions and the distributions with polynomial tail decays.

1. $I(t) = c_\alpha \sqrt[\alpha]{t}$: Theorem 1 yields

$$\mathbb{P}\left(S_m > \gamma_m\right) \leq \begin{cases} \exp\left(-c_{\frac{\gamma_m}{m}} \beta I(\gamma_m)\right) + m \exp\left(-I(\gamma_m)\right), & \gamma_m \gg m^{\frac{\alpha}{2\alpha-1}}, \\ \exp\left(-\dfrac{{\gamma_m}^2}{2m\left(\sigma^2+\epsilon\right)}\right) + m \exp\left(-\dfrac{m t_{\max}^2}{\beta(\sigma^2+\epsilon)}\right), & \gamma_m \ll m^{\frac{\alpha}{2\alpha-1}}. \end{cases} \tag{2.14}$$

Note that $1-\beta$ and ϵ can be chosen arbitrarily small. Moreover, $\lim_{\gamma_m \to \infty} c_{\frac{\gamma_m}{m}} = 1$. Hence, in the first case the right hand side behaves like its dominant term which is $\exp\left(-I(\gamma_m)\right)$. As proven in Theorem 2, $\mathbb{P}\left(S_m > \gamma_m\right) \sim \exp\left(-I(\gamma_m)\right)$ which proves the asymptotic sharpness of our first bound. Furthermore, when $\gamma_m \ll m^{\frac{\alpha}{2\alpha-1}}$ the right hand side of Inequality (2.14) behaves like $\exp\left(-\frac{\gamma_m^2}{2m\sigma^2}\right)$. It is known for γ_m growing at this speed we have

$$\lim_{m\to\infty} \frac{\mathbb{P}\left(S_m > \gamma_m\right)}{\overline{\Phi}\left(\frac{\gamma_m}{\sigma\sqrt{m}}\right)} = 1,$$

where $\overline{\Phi} = 1-\Phi(t)$ and Φ is the CDF of standard normal distribution [64]. Since $\overline{\Phi}\left(\frac{\gamma_m}{\sigma\sqrt{m}}\right) \sim \frac{\sqrt{m}\sigma}{\sqrt{2\pi}\gamma_m}\exp\left(-\frac{\gamma_m^2}{2m\sigma^2}\right)$ we have

$$\lim_{m\to\infty} \frac{-\log\overline{\Phi}\left(\frac{\gamma_m}{\sigma\sqrt{m}}\right)}{\frac{\gamma_m^2}{2m\sigma^2}} = 1.$$

Hence,

$$\lim_{m\to\infty} \frac{-\log\mathbb{P}\left(S_m > \gamma_m\right)}{\frac{\gamma_m^2}{2m\sigma^2}} = 1.$$

This proves the asymptotic sharpness of our second bound.

2. $I(t) = \gamma\log t$ for $\gamma > 2$: Theorem 1 yields

$$\mathbb{P}\left(S_m > \gamma_m\right) \leq \begin{cases} \exp\left(-c_{\frac{\gamma_m}{m}}\beta I(\gamma_m)\right) + \exp\left(-\left(I(\gamma_m) - \log m\right)\right), & \gamma_m \gg \sqrt{m\log m}, \\ \exp\left(-\frac{{\gamma_m}^2}{2m\left(\sigma^2+\epsilon\right)}\right) + m\exp\left(-\frac{mt_{\max}^2}{\beta(\sigma^2+\epsilon)}\right), & \gamma_m \ll \sqrt{m\log m}, \end{cases}$$

for any $\beta < 1-\frac{2}{\gamma}$. This time, one can easily check $m\exp\left(-I(\gamma_m)\right) = m\gamma_m^{-\gamma}$ and $\exp\left(-\frac{{\gamma_m}^2}{2m(\sigma^2+\epsilon)}\right)$ will be dominant for the first and second cases, respectively. One more time, the rate function given by Corollary 1 in the first case, and the Gaussian CDF approximation in the second case [64] match the dominant terms offered by Theorem 1.

2.4 Proofs of our main results

In this section, we state and prove a key lemma about the truncated random variable. This lemma is important in the proof of our concentration and large deviation results.

2.4.1 Proof of Theorem 1

Lemma 4. *If $X^L \triangleq X\mathbf{1}_{X\leq L}$, then for all $\lambda > 0$ and $L > 0$ we have*

$$\log \mathbb{E}\left[\exp\left(\lambda(X^L - \mathbb{E}[X])\right)\right] \leq \frac{k_{L,\lambda}}{2}\lambda^2,$$

where

$$\begin{aligned} k_{L,\lambda} = {} & \mathbb{E}\left[\left(X^L - \mathbb{E}[X]\right)^2 \mathbb{I}\left(X^L \leq \mathbb{E}[X]\right)\right] \\ & + \mathbb{E}\left[\left(X^L - \mathbb{E}[X]\right)^2 \exp\left(\lambda\left(X^L - \mathbb{E}[X]\right)\right)\mathbb{I}\left(X^L > \mathbb{E}[X]\right)\right]. \end{aligned}$$

Proof. From the mean value theorem we have

$$\exp(\lambda X_L) = \exp\left(\mathbb{E}[\lambda X]\right) + \left(\lambda X_L - \mathbb{E}[\lambda X]\right)\exp\left(\lambda\mathbb{E}[X]\right) + \frac{1}{2}\left(\lambda X_L - \mathbb{E}[\lambda X]\right)^2 \exp(\lambda Y), \quad (2.15)$$

where Y is a random variable whose value is always between $\mathbb{E}[X]$ and X_L. Hence,

$$\begin{aligned} & \log \mathbb{E}\left[\exp\left(\lambda X \mathbf{1}_{X\leq L}\right)\right] = \lambda\mathbb{E}[X] \\ + \quad & \log\left(1 + \lambda\left(\mathbb{E}[X_L] - \mathbb{E}[X]\right) + \frac{1}{2}\lambda^2\mathbb{E}\left[\left(X_L - \mathbb{E}[X]\right)^2 \exp\left(\lambda Y - \mathbb{E}[\lambda X]\right)\right]\right). \quad (2.16) \end{aligned}$$

Note that $X_L - X \leq 0$ and $\lambda > 0$. Thus,

$$\begin{aligned}
&\log \mathbb{E}\left[\exp\left(\lambda\left(X\mathbf{1}_{X\le L} - \mathbb{E}\left[X\right]\right)\right)\right] \\
&\quad = \log\left(1 + \lambda\left(\mathbb{E}\left[X_L\right] - \mathbb{E}\left[X\right]\right) + \frac{1}{2}\lambda^2 \mathbb{E}\left[\left(X_L - \mathbb{E}\left[X\right]\right)^2 \exp\left(\lambda Y - \mathbb{E}\left[\lambda X\right]\right)\right]\right) \\
&\quad \le \log\left(1 + \frac{1}{2}\lambda^2 \mathbb{E}\left[\left(X_L - \mathbb{E}\left[X\right]\right)^2 \exp\left(\lambda Y - \mathbb{E}\left[\lambda X\right]\right)\right]\right) \\
&\quad \le \frac{1}{2}\lambda^2 \mathbb{E}\left[\left(X_L - \mathbb{E}\left[X\right]\right)^2 \exp\left(\lambda Y - \mathbb{E}\left[\lambda X\right]\right)\right]. \qquad (2.17)
\end{aligned}$$

Since Y falls between $\mathbb{E}\left[X\right]$ and X^L we have

$$Y \le \mathbb{E}\left[X\right] \mathbb{I}\left(X^L \le \mathbb{E}\left[X\right]\right) + X^L \mathbb{I}\left(X^L > \mathbb{E}\left[X\right]\right).$$

Hence the expectation in (2.17) is bounded by

$$\mathbb{E}\left[\left(X^L - \mathbb{E}\left[X\right]\right)^2 \mathbb{I}\left(X^L \le \mathbb{E}\left[X\right]\right)\right] + \mathbb{E}\left[\left(X^L - \mathbb{E}\left[X\right]\right)^2 \exp\left(\lambda\left(X^L - \mathbb{E}\left[X\right]\right)\right) \mathbb{I}\left(X^L > \mathbb{E}\left[X\right]\right)\right].$$

□

Proof of Theorem 1. Note that by Lemma 4 and (2.2) we have

$$\log \mathbb{E}\left[\exp\left(\lambda(X^L - \mathbb{E}\left[X\right])\right)\right] \le \frac{c_{L,\beta}}{2}\lambda^2.$$

Moreover,

$$\begin{aligned}
\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right) &\le \mathbb{P}\left(\sum X_i^L - \mathbb{E}\left[S_m\right] > mt\right) + \mathbb{P}\left(\exists i \quad X_i > L\right) \\
&\le \exp\left(-\lambda m t\right) \mathbb{E}\left[\exp\left(\lambda(X^L - \mathbb{E}\left[X\right])\right)\right]^m + m\mathbb{P}\left(X > L\right) \\
&\le \exp\left(m\left(-\lambda t + \frac{c_{L,\beta}}{2}\lambda^2\right)\right) + m\exp\left(-I(L)\right). \qquad (2.18)
\end{aligned}$$

The main remaining step is to find good choices for the free parameters L and λ. The goal is to choose the values of λ, L such that we get the best upper bound in (2.18). We consider two cases: (i) $t > t_{\max}$, and (ii) $t \leq t_{\max}$. In each case, we select these parameters accordingly.

- Case 1 ($t > t_{\max}$): In this case, we choose $L = mt$ and $\lambda = \beta\frac{I(mt)}{mt}$. We have

  $$\begin{aligned}\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right) &\leq \exp\left(-\beta\left(1 - \frac{\beta c_{L,\beta} I(mt)}{2mt^2}\right) I(mt)\right) + m\exp\left(-I(mt)\right)\\ &= \exp\left(-\beta c_t I(mt)\right) + m\exp\left(-I(mt)\right).\end{aligned}$$

  Note that since for all $t > t_{\max}$ we have $t > \beta c_{L,\beta}\frac{I(mt)}{mt}$, we can conclude $\frac{1}{2} \leq c_t < 1$.

- Case 2 ($t \leq t_{\max}$): In this case, we pick $L = mt_{\max}$ and $\lambda = \frac{t}{c_{L,\beta}} \leq \frac{t_{\max}}{c_{L,\beta}} = \beta\frac{I(L)}{L}$. Then, (2.18) implies

  $$\begin{aligned}\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right) &\leq \exp\left(-\frac{1}{2c_{L,\beta}}mt^2\right) + m\exp\left(-I(mt_{\max})\right)\\ &= \exp\left(-\frac{1}{2c_{L,\beta}}mt^2\right) + m\exp\left(-\frac{1}{\beta c_{L,\beta}}mt_{\max}^2\right).\end{aligned}$$

  Note that $c_{L,\beta}$ is increasing in β. Hence, choosing a smaller value for λ, as we did in this case, causes no problem.

□

2.4.2 Proof of Lemma 3

First, we prove the lemma under Assumption (a). Note that for $L > \mathbb{E}\left[X\right]$ we have $\left(X_L - \mathbb{E}\left[X\right]\right)^2 \leq \left(X - \mathbb{E}\left[X\right]\right)^2 \in \mathcal{L}^1$. Furthermore, $X_L \xrightarrow{a.s.} X$. Hence, by using the dominant convergence theorem we obtain

$$\mathbb{E}\left[\left(X^L - \mathbb{E}\left[X\right]\right)^2 \mathbb{I}\left(X \leq \mathbb{E}\left[X\right]\right)\right] \xrightarrow{L\to\infty} \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2 \mathbb{I}\left(X \leq \mathbb{E}\left[X\right]\right)\right]. \tag{2.19}$$

Furthermore, it is straightforward to show that

$$(X^L - \mathbb{E}\left[X\right])^2 \exp\left(\lambda_{L,\beta}(X^L - \mathbb{E}\left[X\right]\right) \xrightarrow{a.s.} (X - \mathbb{E}\left[X\right])^2. \tag{2.20}$$

Hence, if we find an $\mathcal{L}^1$ function that dominates $(X^L - \mathbb{E}\left[X\right])^2 \exp\left(\lambda_{L,\beta}(X^L - \mathbb{E}\left[X\right]\right)$, then we can use the dominant convergence theorem to complete the proof. Toward this goal, we consider

$$Y = \left(X - \mathbb{E}\left[X\right]\right)^2 \exp\left(\beta c_\alpha \sqrt[\alpha]{\max(X,0)} + 1\right) \mathbb{I}\left(X > \mathbb{E}\left[X\right]\right).$$

Note that for $X > \mathbb{E}\left[X\right]$, $L > 2\mathbb{E}\left[X\right]$ and $-\lambda_{L,\beta}\mathbb{E}\left[X\right] \leq 1$, we have (we remind the reader that $\lambda_{L,\beta} = \beta\frac{I(L)}{L} \to 0$ as $L \to \infty$)

$$\exp\left(\lambda_{L,\beta}(X^L - \mathbb{E}\left[X\right])\right) \leq \exp\left(\lambda_{L,\beta}X^L + 1\right) = \exp\left(\beta c_\alpha \frac{\sqrt[\alpha]{L}}{L} X^L + 1\right) \leq \exp\left(\beta c_\alpha \sqrt[\alpha]{\max(X,0)} + 1\right). \tag{2.21}$$

Thus, for L large enough we have

$$(X^L - \mathbb{E}\left[X\right])^2 \exp\left(\lambda_{L,\beta}(X^L - \mathbb{E}\left[X\right]\right) \mathbb{I}\left(X > \mathbb{E}\left[X\right]\right) \leq Y. \tag{2.22}$$

To prove the integrability of Y, note that

$$\begin{aligned}
&\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2 \exp\left(\beta c_\alpha \sqrt[\alpha]{\max(X,0)}\right) \mathbb{I}(X > \mathbb{E}\left[X\right])\right] \\
&\quad = \int_0^\infty \mathbb{P}\left(\left(X - \mathbb{E}\left[X\right]\right)^2 \exp\left(\beta c_\alpha \sqrt[\alpha]{\max(X,0)}\right) > u, X > \mathbb{E}\left[X\right]\right) du \\
&\quad \leq \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2 \mathbb{I}\left(\mathbb{E}\left[X\right] \leq X < 0\right)\right] + \int_0^\infty \mathbb{P}\left(X > t\right) du \qquad (t - \mathbb{E}\left[X\right])^2 \exp\left(\beta c_\alpha \sqrt[\alpha]{t}\right) = u \\
&\quad \leq Var(X) + \int_0^\infty \exp\left(-c_\alpha \sqrt[\alpha]{t}\right) du \\
&\quad \leq Var(X) + \int_0^\infty \exp\left(-c_\alpha \sqrt[\alpha]{t}\right) \left(2(t - \mathbb{E}\left[X\right]) + \frac{\beta c_\alpha}{\alpha} t^{\frac{1}{\alpha}-1}(t - \mathbb{E}\left[X\right])^2\right) \exp\left(\beta c_\alpha \sqrt[\alpha]{t}\right) dt \\
&\quad \leq Var(X) + \int_0^\infty \exp\left(-c_\alpha(1-\beta)\sqrt[\alpha]{t}\right) \text{Poly}\left(t^{\frac{1}{\alpha}-1}, t\right) dt < \infty.
\end{aligned}$$

Recall that $\beta < 1$ and $c_\alpha > 0$, hence the exponent of the last line is negative. Thus Y is integrable as it was desired.

The proof under assumption (b) is analogous to the proof of part (a). The only difference is to prove the dominant convergence theorem for the following variable:

$$(X^L - \mathbb{E}[X])^2 \exp\Big(\lambda_{L,\beta}(X^L - \mathbb{E}\left[X\right)]\Big) \mathbb{I}\left(X > \mathbb{E}[X]\right).$$

Toward this goal we use the dominant variable:

$$\begin{aligned} Y &= \left(X - \mathbb{E}[X]\right)^2 \exp\Big(\beta\gamma \log\left(X - \mathbb{E}[X]\right)\Big) \mathbb{I}\left(X > \mathbb{E}[X]\right) \\ &= \left(X - \mathbb{E}[X]\right)^{2+\beta\gamma} \mathbb{I}\left(X > \mathbb{E}[X]\right). \end{aligned}$$

The proof of the integrability of this variable is left to the readers.

2.4.3 Proof of Theorem 2

We start with a lemma that will be used in our proof later.

Lemma 5. *Let a_n, b_n and c_n be sequences of positive numbers such that*

$$\lim_{n\to\infty} \frac{\log a_n}{c_n} = a, \quad \lim_{n\to\infty} \frac{\log b_n}{c_n} = b, \quad \lim_{n\to\infty} c_n = \infty.$$

Then

$$\lim_{n\to\infty} \frac{\log(a_n + b_n)}{c_n} = \max\{a, b\}. \tag{2.23}$$

Proof. Without loss of generality assume $a \geq b$, hence $a_n \geq b_n$ for large enough n. Thus

$$a = \lim_{n\to\infty} \frac{\log a_n}{c_n} \leq \lim_{n\to\infty} \frac{\log(a_n + b_n)}{c_n} \leq \lim_{n\to\infty} \frac{\log 2a_n}{c_n} = \lim_{n\to\infty} \frac{\log 2}{c_n} + \lim_{n\to\infty} \frac{\log a_n}{c_n} = a.$$

Therefore

$$\lim_{n\to\infty} \frac{\log(a_n + b_n)}{c_n} = a.$$

□

First note that

$$\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > \gamma_m\right) \geq \mathbb{P}\left(X > \gamma_m\right)\mathbb{P}\left(S_{m-1} - \mathbb{E}\left[S_{m-1}\right] \geq \mathbb{E}\left[X\right]\right).$$

Since $\frac{S_{m-1}-\mathbb{E}[S_{m-1}]}{\sqrt{m-1}} \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(X))$ and $\frac{\mathbb{E}[X]}{\sqrt{m-1}} \to 0$ we have

$$\mathbb{P}\left(S_{m-1} - \mathbb{E}\left[S_{m-1}\right] \geq \mathbb{E}\left[X\right]\right) \geq C > 0,$$

for a positive constant C and large enough m. Therefore,

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > \gamma_m\right)}{I(\gamma_m)} \leq \lim_{m\to\infty} \frac{-\log \mathbb{P}\left(X > \gamma_m\right)}{I(\gamma_m)} + \frac{-\log C}{I(\gamma_m)} = 1. \tag{2.24}$$

To obtain the last equality we used the fact that since $\log m \ll I(\gamma_m)$ we have $I(\gamma_m) \to \infty$ as $m \to \infty$. Hence,

$$\lim_{m\to\infty} \frac{-\log C}{I(\gamma_m)} = 0.$$

On the other hand,

$$\begin{aligned}\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > \gamma_m\right) &\leq \exp\left(-\lambda\gamma_m\right)\mathbb{E}\left[\exp\left(\lambda(X^L - \mathbb{E}\left[X\right])\right)\right]^m + m\mathbb{P}\left(X > L\right) \\ &\leq \exp\left(-\lambda\gamma_m\right)\exp\left(\frac{k_{L,\lambda}}{2}\lambda^2 m\right) + m\mathbb{P}\left(X > L\right), \end{aligned} \tag{2.25}$$

where we used Lemma 4 to obtain the last inequality. Let $L = \gamma_m$ and $\lambda = \beta\frac{I(\gamma_m)}{\gamma_m}$. Moreover, assume c_β is the bound for $c_{L,\beta}$ when L is large enough. Then, (2.25) implies that

$$\mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > \gamma_m\right) \leq \exp\left(-\beta I(\gamma_m) + \frac{\beta^2 c_\beta}{2}\frac{mI(\gamma_m)^2}{\gamma_m^2}\right) + m\mathbb{P}\left(X > \gamma_m\right). \tag{2.26}$$

In order to find a lower bound for $\lim \frac{-\log \mathbb{P}\left(S_m - \mathbb{E}[S_m] > \gamma_m\right)}{I(\gamma_m)}$, we use Lemma 5. Hence, we need to bound each term of (2.26) separately.

$$\lim_{m\to\infty} \frac{\beta I(\gamma_m) - \frac{\beta^2 c_\beta}{2} \frac{m I(\gamma_m)^2}{\gamma_m^2}}{I(\gamma_m)} = \beta + \frac{\beta^2 c_\beta}{2} \lim_{m\to\infty} \frac{-m I(\gamma_m)}{\gamma_m^2} = \beta, \tag{2.27}$$

where we used $I(\gamma_m) = o(\frac{\gamma_m^2}{m})$ to obtain the last equality. Moreover,

$$\lim_{m\to\infty} \frac{-\log(m\mathbb{P}\left(X > \gamma_m\right))}{I(\gamma_m)} = 1. \tag{2.28}$$

The last equality holds because I captures the tail of X asymptotically and grows faster than $\log(m)$. Hence, using (2.26), (2.27) and (2.28) we obtain

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right)}{I(mt)} \geq \beta, \qquad \forall \beta < 1,$$

which implies

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right)}{I(mt)} \geq 1. \tag{2.29}$$

By using (2.24) and (2.29) we obtain

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m - \mathbb{E}\left[S_m\right] > mt\right)}{I(mt)} = 1,$$

which concludes the proof.

2.4.4 Proof of Corollary 6

First, assume γ_m satisfies (i). Let $\beta < 1 - \frac{k}{\alpha}$, hence $(1-\beta)\alpha = k' > k$. According to Corollary 3 for this β and $L = \gamma_m$ we have

$$c_{\gamma_m,\beta} \leq C\gamma_m^{2-(1-\beta)\alpha} \log \gamma_m = C\gamma_m^{2-k'} \log \gamma_m.$$

Therefore

$$\frac{\gamma_m}{m} \gg \gamma_m^{2-k'} \frac{\log m}{\gamma_m} \geq C'\beta c_{\gamma_m,\beta} \frac{I(\gamma_m)}{\gamma_m},$$

since we have $\lim \frac{\log m}{\log \gamma_m} = k < k'$. Thus, for large enough m, when applying Theorem 1 with $t = \frac{\gamma_m}{m}$ and the chosen β above we will be in the $t > t_{\max}$ regime.

For the second case that γ_m satisfies (ii), Lemma 3 implies that for any $\beta < 1 - \frac{2}{\alpha}$, $c_{\gamma_m,\beta}$ remains bounded. Hence we have

$$\frac{\gamma_m}{m} \gg \beta c_{\gamma_m,\beta} \frac{I(\gamma_m)}{\gamma_m} = O\left(\frac{\log m}{\gamma_m}\right),$$

which means we still are in the region $t > t_{\max}$. Hence,

$$\mathbb{P}\left(S_m > \gamma_m\right) \leq \exp\left(-c_{\frac{\gamma_m}{m}} \beta I(\gamma_m)\right) + m \exp\left(-I(\gamma_m)\right). \tag{2.30}$$

Note that $c_{\frac{\gamma_m}{m}} = 1 - \frac{1}{2} \frac{\beta c_{\gamma_m,\beta}}{\frac{\gamma_m}{m}} \frac{I(\gamma_m)}{\gamma_m} \xrightarrow{m\to\infty} 1$, so we obtain

$$\lim_{m\to\infty} \frac{c_{\frac{\gamma_m}{m}} \beta I(\gamma_m)}{I(\gamma_m) - \log m} = \lim_{m\to\infty} \frac{\beta}{1 - \frac{\log m}{I(\gamma_m)}} = \lim_{m\to\infty} \frac{\beta}{1 - \frac{\log m}{\alpha \log \gamma_m}} = \frac{\beta}{1 - \frac{k}{\alpha}} = \frac{\alpha - k'}{\alpha - k}, \qquad \forall k' > k. \tag{2.31}$$

Moreover,

$$\lim_{m\to\infty} \frac{I(\gamma_m) - \log m}{I(\gamma_m) - \log m} = 1. \tag{2.32}$$

By combining (2.30), (2.31) and (2.32) we obtain

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m > \gamma_m\right)}{I(\gamma_m) - \log m} \geq \frac{\alpha - k'}{\alpha - k}, \qquad \forall k' > k,$$

which implies

$$\lim_{m\to\infty} \frac{-\log \mathbb{P}\left(S_m > \gamma_m\right)}{I(\gamma_m) - \log m} \geq 1. \tag{2.33}$$

On the other hand,

$$\begin{aligned}
\mathbb{P}\left(S_m > \gamma_m\right) &\geq \sum_{j=1}^{m} \mathbb{P}\left(\sum_{i\neq j} X_i > -\epsilon\sqrt{m}, \quad \max_{i\neq j} X_i < \gamma_m\right)\mathbb{P}\left(X_j \geq \gamma_m + \epsilon\sqrt{m}\right) \\
&= m\mathbb{P}\left(\frac{S_{m-1}}{\sqrt{m}} > -\epsilon, \max_{i\leq m-1} < \gamma_m\right)\mathbb{P}\left(X_m \geq \gamma_m + \epsilon\sqrt{m}\right) \\
&\geq \left(\mathbb{P}\left(\frac{S_{m-1}}{\sqrt{m}} > -\epsilon\right) - \mathbb{P}\left(\exists i \leq m-1, X_i > \gamma_m\right)\right) m\mathbb{P}\left(X \geq \gamma_m + \epsilon\sqrt{m}\right) \\
&\geq \left(\mathbb{P}\left(\frac{S_{m-1}}{\sqrt{m}} > -\epsilon\right) - (m-1)\mathbb{P}\left(X > \gamma_m\right)\right) m\mathbb{P}\left(X \geq \gamma_m + \epsilon\sqrt{m}\right). \qquad (2.34)
\end{aligned}$$

Note that by the central limit theorem we have $\mathbb{P}\left(\frac{S_{m-1}}{\sqrt{m}} > -\epsilon\right) \geq \mathbb{P}\left(\frac{S_{m-1}}{\sqrt{m}} > 0\right) \xrightarrow{m\to\infty} \frac{1}{2}$. Furthermore,

$$\begin{aligned}
(m-1)\mathbb{P}\left(X > \gamma_m\right) &= \exp\left(\log(m-1) - I_{br}(\gamma_m)\right) \\
&\sim \exp\left(\log(m-1) - \alpha\log(\gamma_m)\right) \\
&\sim \exp\left((1 - \frac{\alpha}{k})\log m\right). \qquad (2.35)
\end{aligned}$$

Since $k \leq 2 < \alpha$, the right hand side of (2.35) goes to 0 as m grows. Hence for large enough m, we have

$$\left(\mathbb{P}\left(\frac{S_{m-1}}{\sqrt{m}} > -\epsilon\right) - (m-1)\mathbb{P}\left(X > \gamma_m\right)\right) \geq \frac{1}{3}.$$

Therefore, by (2.34) we obtain

$$\begin{aligned}\lim_{m\to\infty}\frac{-\log\mathbb{P}\left(S_m>\gamma_m\right)}{I(\gamma_m)-\log m} &\leq \lim_{m\to\infty}\frac{\log 3-\log\mathbb{P}\left(X>\gamma_m+\epsilon\sqrt{m}\right)-\log m}{\alpha\log\gamma_m-\log m}\\ =\ \lim_{m\to\infty}\frac{\alpha\log\left(\gamma_m+\epsilon\sqrt{m}\right)-\log m}{\alpha\log\gamma_m-\log m}=1\end{aligned}$$

To obtain the last equality we have used $\lim_{m\to\infty}\frac{\log\left(\gamma_m+\epsilon\sqrt{m}\right)}{\log\gamma_m}=1$ which can be easily proved by noting that $\log\gamma_m\leq\log\left(\gamma_m+\epsilon\sqrt{m}\right)\leq\log\gamma_m+\frac{\epsilon\sqrt{m}}{\gamma_m}$ and that $\sqrt{m}\ll\gamma_m$ since $k<2$.

2.5 Conclusion

We developed a framework to study the concentration of the sum of independent and identically distributed random variables with heavy tails. In particular, we considered distributions for which the moment generating function does not exist. Techniques that we offered in this chapter are pretty simple and yet effective for all distributions that have finite variances. The generality and simplicity of the tools not only enable us to recognize different deviation behaviors, but also help us to determine the boundary of such phase transitions precisely. Furthermore, we showed the tools that we developed for obtaining concentration inequalities are sharp enough to offer large deviation results as well. Note that there are plenty of results in the literature, such as Hanson-Wright inequality [65] and Gartner-Ellis Theorem [66], whose proof heavily relies on the moment generating function. We believe that the framework presented here can extend all such results to the class of distributions with finite variance.

Chapter 3: Spectral Method for Phase Retrieval

3.1 Introduction

Phase retrieval refers to the problem of recovering a signal $\boldsymbol{x} \in \mathbb{C}^n$ from a set of phaseless linear observations $\boldsymbol{y} \in \mathbb{R}^m$. Under the absence of the measurement noise, the acquisition process is modeled as

$$y_i = |(\boldsymbol{A}\boldsymbol{x})_i|,$$

where $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is a measurement matrix and $(\cdot)_i$ denotes the i^{th} element of a vector. The phase retrieval problem is intended to model practical imaging systems where it is difficult to measure the phase of the measurements [7]. A number of recent recovery algorithms pose Phase retrieval as a non-convex optimization problem, and employ a local search algorithm to find the minimizer [25, 67, 26, 68]. For instance, the well known Wirtinger Flow algorithm [25] solves the optimization problem:

$$\min_{\boldsymbol{z}} \quad \sum_{i=1}^{m} (y_i^2 - |\boldsymbol{a}_i^* \boldsymbol{z}|^2)^2, \tag{3.1}$$

using gradient descent.

Since the optimization problem (3.1) is non-convex, the initialization can have an impact on the success of local search algorithms. The most widely used initialization scheme, known as spectral initialization [28, 67, 26, 29, 69, 30], uses the leading eigenvector of the following data-dependent matrix:

$$\boldsymbol{M} \overset{\Delta}{=} \boldsymbol{A}^* \boldsymbol{T} \boldsymbol{A} \tag{3.2}$$

as the starting point for local search algorithms. In the above equation,

$$\boldsymbol{T} = \text{Diag}(\mathcal{T}(y_1), \mathcal{T}(y_2), \ldots, \mathcal{T}(y_m)),$$

and $\mathcal{T}(\cdot)$ denotes a suitable trimming function. Let $\hat{\boldsymbol{x}}$ denote the leading eigenvector of $\boldsymbol{M}$ normalized to have unit Euclidean (ℓ_2) norm. That is,

$$\hat{\boldsymbol{x}} \overset{\Delta}{=} \max_{\|z\|=1} \boldsymbol{z}^* \boldsymbol{M} \boldsymbol{z}. \tag{3.3}$$

The earliest analysis [28, 25] of the spectral estimator showed that if number of measurements m is large enough (for a fixed n), then the leading eigenvector of $\boldsymbol{M}$ is a consistent estimator of the true signal vector. However these analyses had two drawbacks: (i) They only provide information about the order of measurements required for a successful initialization and not a sharp requirement on the sampling ratio m/n, (ii) These analyses fail to capture the difference in the performance of various trimming functions. Recently, Lu and Li [29] have analyzed the spectral estimator for measurement matrices that are composed of independent and identically distributed (i.i.d.) standard normal entries in the high dimensional asymptotic regime. More specifically, Lu and Li considered the asymptotic setting in which $m, n \to \infty$, $m/n = \delta$, and obtained a sharp characterization for the overlap between the leading eigenvector and the true signal. In follow up work by Mondelli and Montanari [69] and Luo, Alghamdi and Lu [30] this characterization was leveraged to design optimal trimming functions. For the optimal trimming function, the overlap $|\hat{\boldsymbol{x}}^* \boldsymbol{x}|^2 / \|\boldsymbol{x}\|^2$ converges to zero when $\delta < 1$, and converges to a strictly positive value otherwise.

A major assumption in the analysis of [29, 69, 30] is that the measurement matrix $\boldsymbol{A}$ contains i.i.d. Gaussian entries. However, it is well-known that many important applications of phase retrieval are concerned with Fourier-type matrices [70]. This leads to the following natural questions: (i) Are the conclusions of [29, 69, 30] correct for other matrices that are employed in practice? (ii) Is the optimal choice of trimming that was derived in [29, 69, 30] for Gaussian measurement matrices optimal for other matrices employed in practice? In response to these questions, Ma *et al.* [71] considered a popular class of matrices that can be used in phase retrieval systems, known as coded diffraction pattern (CDP) [72]. Through an extensive numerical study, the authors showed that the performance of the spectral initialization for such matrices closely approximates the performance of the spectral estimator for partial orthogonal matrices. The authors then designed an Expectation Propagation (EP) [73, 74] algorithm for the eigenvalue problem given in (3.3). EP algorithms had previously been proposed for partial orthogonal matrices in [75, 76] and their State Evolution (SE) had been analyzed in [77, 78]. Ma *et al.* used the SE of derived EP algorithm for the eigenvalue problem to derive a (conjectured) formula for the asymptotic overlap $|\hat{\boldsymbol{x}}^* \boldsymbol{x}|^2 / \|\boldsymbol{x}\|^2$ between the true signal vector and the spectral initialization. However, while it is believed that EP algorithm indeed solves the eigenvalue problem (this has also been observed in simulations), this has not been shown rigorously. As a result of such studies, the authors conjectured that for partial orthogonal matrices if the trimming function is chosen optimally, then for $\delta > 2$, $|\hat{\boldsymbol{x}}^* \boldsymbol{x}|^2 / \|\boldsymbol{x}\|^2 > 0$, and for $\delta < 2$, $|\hat{\boldsymbol{x}}^* \boldsymbol{x}|^2 / \|\boldsymbol{x}\|^2 = 0$, in the asymptotic setting where $n, m = \delta n \to \infty$. As mentioned previously, the simulations in [71] suggest that these conjectures are also likely to hold for CDP matrices.

In this paper, we prove the conjectures presented in [71] for partial orthogonal matrices using tools from the free probability theory [79]. We believe this is the first theoretical justification that the expectation propagation framework can correctly predict the statistical properties of the solutions to non-convex optimization problems. The main technical step in our proof is the identification of the location of the largest eigenvalue using a subordination function [79]. Interestingly, this subordination function appears naturally in the expectation propagation (EP) algorithm of [71].

3.2 Main result

3.2.1 Notation

For Linear Algebraic Aspects

For a matrix $\boldsymbol{A}$, $\boldsymbol{A}^*$ refers to the conjugate transpose of $\boldsymbol{A}$. For a matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$, with real eigenvalues, we use $\lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \cdots \geq \lambda_n(\boldsymbol{A})$ to denote the eigenvalues arranged in descending order. We use $\sigma(\boldsymbol{A})$ to refer to the spectrum of $\boldsymbol{A}$ which is simply the set of eigenvalues

$\{\lambda_1(\boldsymbol{A}), \lambda_2(\boldsymbol{A}) \dots \lambda_n(\boldsymbol{A})\}$. Finally we define the spectral measure of $\boldsymbol{A}$, denoted by $\mu_{\boldsymbol{A}}$ as,

$$\mu_{\boldsymbol{A}} \overset{\Delta}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\boldsymbol{A})}.$$

For $m, n \in \mathbb{N}$, we denote the $m \times m$ identity matrix by $\boldsymbol{I}_m$ and a $m \times n$ matrix of all zero entries by $\boldsymbol{0}_{m,n}$. For $m \geq n$, We also define the special matrix $\boldsymbol{S}_{m,n}$ as:

$$\boldsymbol{S}_{m,n} \overset{\Delta}{=} \begin{bmatrix} \boldsymbol{I}_n \\ \boldsymbol{0}_{m-n,n} \end{bmatrix}. \tag{3.4}$$

For Complex Analytic Aspects

For a complex number $z \in \mathbb{C}$, $\mathrm{Re}(z), \mathrm{Im}(z), \mathsf{Arg}(z), |z|, \overline{z}$ refer to the real part, imaginary part, argument, modulus and conjugate of z. We denote the complex upper half plane and lower half planes by

$$\mathbb{C}^+ \overset{\Delta}{=} \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\} \text{ and } \mathbb{C}^- \overset{\Delta}{=} \{z \in \mathbb{C} : \mathrm{Im}(z) < 0\}.$$

For Probabilistic Aspects

We use $\mathcal{CN}(0,1)$ to denote the standard, circularly symmetric, complex Gaussian distribution. $\mathrm{Unif}(\mathbb{U}_m)$ denotes the Haar measure on the unitary group. We denote almost sure convergence, convergence in probability and convergence in distribution by $\xrightarrow{\text{a.s.}}$, $\xrightarrow{\text{P}}$ and $\xrightarrow{\text{d}}$ respectively. Two random variables X, Y are equal in distribution, denoted by $X \overset{\text{d}}{=} Y$ if they have the same distribution. Throughout this paper, the random variables Z, T refer to the pair of random variables with the joint distribution given by $Z \sim \mathcal{CN}(0,1)\,, T = \mathcal{T}(|Z|/\sqrt{\delta})$. For a borel probability measure μ, we use $\mathrm{Supp}(\mu)$ to denote the support of μ.

Miscellaneous:

Let A be a subset of $\mathbb{R}$ or $\mathbb{C}$. $\overline{A}$ denotes the closure of A. The distance from a point $x \in \mathbb{R}$ to A is defined by $\mathrm{dist}(x, A) = \inf_{y \in A} |x - y|$. We define the ϵ neighborhood of A, denoted by A_ϵ as

$$A_\epsilon \overset{\Delta}{=} \{x : \mathrm{dist}(x, A) < \epsilon\}.$$

The symbol $\emptyset$ is used to denote the empty set.

3.2.2 Measurement Model and Spectral Estimator

In the phase retrieval problem we are given m observations $\boldsymbol{y} \in \mathbb{R}^m$ generated as:

$$\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|$$

where $\boldsymbol{x} \in \mathbb{C}^n$ is the unknown signal vector and $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is the sensing matrix. We assume that $\|\boldsymbol{x}\| = \sqrt{n}$ and that the matrix $\boldsymbol{A}$ is generated according to the following process: Sample $\boldsymbol{H}_m \in \mathbb{U}(m)$ from the Haar measure on the unitary group $\mathbb{U}(m)$ and set $\boldsymbol{A}$ to be the matrix formed by picking the first n columns of $\boldsymbol{H}_m$. More formally,

$$\boldsymbol{A} = \boldsymbol{H}\boldsymbol{S}_{m,n}, \ \boldsymbol{H} \sim \mathrm{Unif}(\mathbb{U}(m)),$$

and $\boldsymbol{S}$ is defined in (3.4). An important parameter for our analysis will be *the sampling ratio*, denoted by $\delta \overset{\Delta}{=} m/n$. Let $\mathcal{T} : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}$ be a trimming function. We study spectral estimators $\hat{\boldsymbol{x}}$ constructed as the leading eigenvector of the matrix $\boldsymbol{M}$, defined below:

$$\hat{\boldsymbol{x}} = \arg\max_{\|\boldsymbol{u}\|=1} \boldsymbol{u}^*\boldsymbol{M}\boldsymbol{u},$$

where $\boldsymbol{M} = \boldsymbol{A}^*\boldsymbol{T}\boldsymbol{A}$ and $\boldsymbol{T} = \mathrm{Diag}(\mathcal{T}(y_1), \mathcal{T}(y_2) \ldots \mathcal{T}(y_m))$.

3.2.3 Assumptions & Asymptotic Framework

We analyze the performance of the spectral estimator in an asymptotic setup where $n, m \to \infty, m/n = \delta > 1$. In particular, we consider a sequence of independent phase retrieval problems realized on the same probability space with increasing n, m. We assume some regularity assumptions on the trimming function $\mathcal{T}$ which are stated below.

Assumption 1. *The trimming function $\mathcal{T}$ satisfies the following conditions:*

1. *$\mathcal{T}$ is Lipschitz continuous.*
2. $\sup_{y \geq 0} \mathcal{T}(y) = 1,\ \inf_{y \geq 0} \mathcal{T}(y) = 0.$
3. *The random variable T, defined by $Z \sim \mathcal{CN}(0,1)$ and $T = \mathcal{T}(|Z|/\sqrt{\delta})$ has a density with respect to the Lebesgue measure on $\mathbb{R}$.*

In the following remarks, we discuss why each of these assumptions are required and whether they can be relaxed.

Remark 8. *We need the trimming function $\mathcal{T}$ to be Lipschitz continuous so that the trimmed measurements $\mathcal{T}(y_i)$ can be approximated in distribution by $\mathcal{T}(|Z|/\sqrt{\delta}), Z \sim \mathcal{CN}(0,1)$. We expect this approximation to hold under weaker smoothness hypothesis on $\mathcal{T}$ than Lipschitz continuity.*

Remark 9. *The assumptions:*

$$\sup_{y \geq 0} \mathcal{T}(y) = 1,\ \inf_{y \geq 0} \mathcal{T}(y) = 0$$

are no stronger than the assumption that $\mathcal{T}$ is a bounded trimming function. In fact, given any arbitary bounded trimming function with $\inf_{y \geq 0} \mathcal{T}(y) = a$ and $\sup_{y \geq 0} \mathcal{T}(y) = b$, the spectral estimator constructed using $\mathcal{T}$ has the same performance as the spectral measure constructed using

$$\tilde{\mathcal{T}}(y) \overset{\Delta}{=} (\mathcal{T}(y) - a)/(b - a).$$

This is because,

$$\begin{aligned}\widetilde{\boldsymbol{M}} \triangleq \boldsymbol{A}^*\widetilde{\boldsymbol{T}}\boldsymbol{A} &= \frac{1}{b-a}\boldsymbol{A}^*\boldsymbol{T}\boldsymbol{A} - \frac{a}{b-a}\boldsymbol{I}_n \\ &= \frac{1}{b-a}\boldsymbol{M} - \frac{a}{b-a}\boldsymbol{I}_n.\end{aligned}$$

In particular $\boldsymbol{M}$ and $\widetilde{\boldsymbol{M}}$ have the same leading eigenvector. We require the assumption that the trimming function is bounded since a number of results in free probability theory that we rely on assume this.

Remark 10. *We need (3) in Assumption 1 to ensure that the limiting spectral measure of the matrix $\boldsymbol{M}$ has no discrete component. We expect that this assumption can be completely removed by a careful analysis since the location of point masses in the limiting spectral measure of $\boldsymbol{M}$ is well understood.*

3.2.4 Main Result

In order to state our main result about the performance of the spectral estimator, we need to introduce the following four functions:

$$\begin{aligned}\Lambda(\tau) &\triangleq \tau - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\tau-T}\right]}, \ \psi_1(\tau) \triangleq \frac{\mathbb{E}\left[\frac{|Z|^2}{\tau-T}\right]}{\mathbb{E}\left[\frac{1}{\tau-T}\right]}, \\ \psi_2(\tau) &\triangleq \frac{\mathbb{E}\left[\frac{1}{(\tau-T)^2}\right]}{\left(\mathbb{E}\left[\frac{1}{\tau-T}\right]\right)^2}, \ \psi_3^2(\tau) \triangleq \frac{\mathbb{E}\left[\frac{|Z|^2}{(\tau-T)^2}\right]}{\left(\mathbb{E}\left[\frac{1}{\tau-T}\right]\right)^2}.\end{aligned} \tag{3.5}$$

In the above display, the random variables Z, T have the joint distribution given by $Z \sim \mathcal{CN}(0,1)$, $T = \mathcal{T}(|Z|/\sqrt{\delta})$. The functions Λ, ψ_1 are defined on $[1,\infty)$ and the functions ψ_2, ψ_3 are defined on $(1,\infty)$.

Remark 11. *Under Assumption 1, the support of the random variable T is the interval $[0,1]$. Hence the definition of these functions at $\tau = 1$ needs some clarification. First, note that the random* *variable* $(1-T)^{-1} \geq 0$. *Hence, the* $\mathbb{E}[(1-T)^{-1}]$ *is well-defined, but maybe* ∞. *If it is finite, each of the above functions are well-defined at* $\tau = 1$. *If* $\mathbb{E}[(1-T)^{-1}] = \infty$, *we define,* $\Lambda(1) = 1, \psi_1(1) = 1$. *This corresponds to interpreting* $1/\infty = 0$ *and* $\infty/\infty = 1$ *in the definition of these functions.*

Theorem 3. *Define* $\tau_r \triangleq \arg\min_{\tau\in[1,\infty)} \Lambda(\tau)$. *Also, let* $\theta_\star$ *denote the unique value of* $\theta > \tau_r$ *that satisfies* $\psi_1(\theta) = \frac{\delta}{\delta-1}$. *Then, under Assumption 1, we have*

$$\lambda_1(\boldsymbol{M}) \overset{a.s.}{\longrightarrow} \begin{cases} \Lambda(\tau_r), & \psi_1(\tau_r) \leq \frac{\delta}{\delta-1}, \\ \Lambda(\theta_\star), & \psi_1(\tau_r) > \frac{\delta}{\delta-1}. \end{cases}$$

Furthermore,

$$\frac{|\boldsymbol{x}^*\hat{\boldsymbol{x}}|^2}{n} \overset{a.s.}{\longrightarrow} \begin{cases} 0, & \psi_1(\tau_r) < \frac{\delta}{\delta-1}, \\ \frac{\left(\frac{\delta}{\delta-1}\right)^2 - \frac{\delta}{\delta-1}\cdot\psi_2(\theta_\star)}{\psi_3(\theta_\star)^2 - \frac{\delta}{\delta-1}\cdot\psi_2(\theta_\star)}, & \psi_1(\tau_r) > \frac{\delta}{\delta-1}. \end{cases}$$

Remark 12. *The proof of Theorem 3 shows that if* $\psi_1(\tau_r) > \delta/(\delta-1)$, *there exists* exactly *one solution to the equation* $\psi_1(\theta) = \delta/(\delta-1),\ \theta \in (\tau_r, \infty)$. *Hence,* $\theta_\star$ *is well-defined.*

The proof of this result is postponed until Section 3.4. Before we proceed to the proof of this theorem, let us clarify some of its interesting features. First, note that similar to the Gaussian sensing matrices, even in the case of partial orthogonal matrices, the maximum eigenvector exhibits a phase transition behavior. For certain values of $\delta > 1$, the inequality $\psi_1(\tau_r) < \frac{\delta}{\delta-1}$ holds, and hence the maximum eigenvector does not carry information about $\boldsymbol{x}$. For other values of δ, the inequality $\psi_1(\tau_\star) > \frac{\delta}{\delta-1}$ holds and hence, the direction of the maximum eigenvector starts to offer information about the direction of $\boldsymbol{x}$. For typical choices of the trimming function $\mathcal{T}$, there exists a critical value of δ, denoted by $\delta_{\mathcal{T}}$ such that, when $\delta < \delta_{\mathcal{T}}$, the spectral estimator is asymptotically orthogonal to the signal vector. When $\delta > \delta_{\mathcal{T}}$, the spectral estimator makes a non-trivial angle with the signal vector. This phase transition phenomena is illustrated in Figure 3.1 for 3 different choices of $\mathcal{T}$.

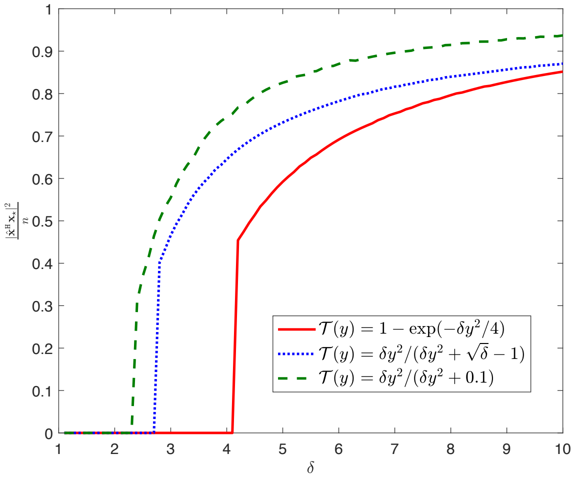


Figure 3.1: Plot of the asymptotic cosine similarity between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$.

Remark 13 (Choice of Trimming function)**.** *The trimming function in Figure 3.1 are supported on* $[0, 1]$.

1. $\mathcal{T}(y) = \delta y^2/(\delta y^2 + \sqrt{\delta} - 1)$ *is a translated and re-scaled version of the trimming function proposed by [69].*

2. $\mathcal{T}(y) = \delta y^2/(\delta y^2 + 0.1)$ *is a regularized version of the trimming function proposed by [30].*

Remark 14 (Extensions to generalized linear measurements)**.** *While we focus on the phase retrieval problem in this paper, our results extend straightforwardly to the generalized linear estimation, where the measurements* y_i *are generated as follows:*

$$y_i \sim f(\cdot|(\boldsymbol{A}\boldsymbol{x})_i),$$

where $f(\cdot|\cdot)$ *denotes a conditional distribution modelling a possibly randomized output channel. Under suitable regularity assumptions on* f*, Theorem 1 holds with the change that the joint distribution of the random variables* T, Z *is now given by:*

$$Z \sim \mathcal{CN}(0, 1)\,,\; Y \sim f\left(\cdot\,\middle|\,\frac{Z}{\sqrt{\delta}}\right),\; T = \mathcal{T}(Y).$$

3.3 Optimal Trimming Functions

Theorem 3 can used to design the trimming function $\mathcal{T}$ optimally in order to obtain the best possible value of $|\boldsymbol{x}_\star^* \hat{\boldsymbol{x}}|^2$. Most of the work towards this goal was already done in [71] where the result in Theorem 3 was stated as a conjecture and was used to design the optimal trimming function. In particular, [71] showed the following impossibility result.

Proposition 1 ([71])**.** *Let $\mathcal{T}$ be any trimming function for which Theorem 3 holds. Then,*

$$\limsup_{\substack{m,n\to\infty\\ m=n\delta}} \frac{|\boldsymbol{x}^* \hat{\boldsymbol{x}}|^2}{n} \overset{a.s.}{\leq} \rho^2_{\mathsf{opt}}(\delta),$$

where,

$$\rho^2_{\mathsf{opt}}(\delta) \triangleq \begin{cases} 0, & \delta \leq 2 \\ \frac{\theta_\star^{\mathsf{opt}} - 1}{\theta_\star^{\mathsf{opt}} - \frac{1}{\delta}}, & \delta > 2 \end{cases},$$

where $\theta_\star^{\mathsf{opt}}$ is the solution to the equation (in τ):

$$\psi_1^{\mathsf{opt}}(\tau) = \frac{\delta}{\delta - 1},\ \psi_1^{\mathsf{opt}}(\tau) \triangleq \frac{\mathbb{E}\left[\frac{|Z|^2}{\tau - T_{\mathsf{opt}}}\right]}{\mathbb{E}\left[\frac{1}{\tau - T_{\mathsf{opt}}}\right]},\ \tau \in (1, \infty),$$

which exists uniquely when $\delta > 2$ and, the random variable T_{opt} is distributed as:

$$Z \sim \mathcal{CN}(0, 1),\ T_{\mathsf{opt}} = 1 - \frac{1}{|Z|^2}.$$

The work [71] also provided a candidate for the optimal trimming function:

$$\mathcal{T}_{\mathsf{opt}}(y) = 1 - \frac{1}{\delta y^2}.$$

They showed that if the characterization given in Theorem 3 holds for $\mathcal{T}_{\mathsf{opt}}$, then it achieves the asymptotic squared correlation $\rho^2_{\mathsf{opt}}(\delta)$. Unfortunately, since $\mathcal{T}_{\mathsf{opt}}$ is unbounded, Theorem 3 does not apply to it. Extending Theorem 3 to unbounded trimming functions would likely require extending previously known results in free probability to unbounded measures, and we don't pursue this approach in our work. Instead, we suitably modify the arguments of [71] to show that the family of bounded trimming functions:

$$\mathcal{T}_{\mathsf{opt},\epsilon}(y) = 1 - \frac{1}{\delta y^2 + \epsilon}, \ \epsilon > 0,$$

attains an asymptotic squared correlation that can be made arbitrarily close to $\rho^2(\delta)$ as $\epsilon \downarrow 0$.

Proposition 2. *Let $\hat{\boldsymbol{x}}_\epsilon$ denote the spectral estimator for $\boldsymbol{x}$ obtained by using $\mathcal{T}_{\mathsf{opt},\epsilon}$ as the trimming function. We have, almost surely,*

$$\lim_{\epsilon \downarrow 0} \lim_{\substack{m,n\to\infty \\ m=n\delta}} \frac{|\boldsymbol{x}^* \hat{\boldsymbol{x}}_\epsilon|^2}{n} = \rho^2_{\mathsf{opt}}(\delta).$$

We provide a proof of this result in Appendix 3.6.

The regularized trimming functions $\mathcal{T}_{\mathsf{opt},\epsilon}$ are not only useful from a theoretical point of view to prove an achievability result, but also from a computational stand point: In simulations we have observed that the power iterations are slow to converge when $\mathcal{T}_{\mathsf{opt}}$ is used as the trimming function due to presence of large negative eigenvalues and this problem is mitigated by using $\mathcal{T}_{\mathsf{opt},\epsilon}$ with a small value of ϵ (such as 0.1 or 0.01) with a negligible degradation in performance.

3.4 Proof of Theorem 3

3.4.1 Roadmap

Our proof follows the general strategy taken by [29]. In this subsection, we state several key lemmas and show how they fit together in the proof of Theorem 3.2. First we note that without loss of generality, for the purpose of analysis of the spectral estimator, we can assume $\boldsymbol{x} = \sqrt{n}\boldsymbol{e}_1$. The following lemma supports this claim.

Lemma 6. *The distribution of the cosine similarity, $\rho^2 = |\boldsymbol{x}^*\hat{\boldsymbol{x}}|^2/n$ is independent of $\boldsymbol{x}$.*

Proof. Let $\boldsymbol{x}$ be an arbitrary signal vector with $\|\boldsymbol{x}\| = \sqrt{n}$. Let $\boldsymbol{y}, \boldsymbol{T}, \hat{\boldsymbol{x}}$ denote the measurements, trimmed measurements and spectral estimate generated when the sensing matrix was $\boldsymbol{A}$ and the signal vector was $\boldsymbol{x}$. Note that the cosine similarity ρ^2 is a (deterministic) function of $\boldsymbol{A}, \boldsymbol{x}$ and hence we use the notation $\rho^2(\boldsymbol{A}, \boldsymbol{x})$ to denote the cosine similarity when the sensing matrix is $\boldsymbol{A}$ and the signal vector is $\boldsymbol{x}$.

Let $\boldsymbol{\Gamma} \in \mathbb{U}(n)$ be such that $\sqrt{n}\boldsymbol{\Gamma}\boldsymbol{e}_1 = \boldsymbol{x}$. We have $\boldsymbol{x}^*\hat{\boldsymbol{x}} = \sqrt{n}\boldsymbol{e}_1^*\boldsymbol{\Gamma}^*\hat{\boldsymbol{x}}$. Next we note that $\hat{\boldsymbol{x}}' \overset{\Delta}{=} \boldsymbol{\Gamma}^*\hat{\boldsymbol{x}}$ is the leading eigenvector of the matrix $\boldsymbol{M}' \overset{\Delta}{=} \boldsymbol{\Gamma}^*\boldsymbol{M}\boldsymbol{\Gamma} = (\boldsymbol{A}\boldsymbol{\Gamma})^*\boldsymbol{T}\boldsymbol{A}\boldsymbol{\Gamma} = \boldsymbol{A}'^*\boldsymbol{T}\boldsymbol{A}'$, where we defined $\boldsymbol{A}' \overset{\Delta}{=} \boldsymbol{A}\boldsymbol{\Gamma}$. Noting that $\boldsymbol{T}$ is a diagonal matrix consisting of the trimmed observations $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}| = \sqrt{n}|\boldsymbol{A}'\boldsymbol{e}_1|$, we conclude that $\hat{\boldsymbol{x}}'$ is the spectral estimate generated when the sensing matrix was $\boldsymbol{A}'$ and the signal vector was $\sqrt{n}\boldsymbol{e}_1$. Hence, we have concluded that

$$\rho^2(\boldsymbol{A}, \boldsymbol{x}) = \rho^2(\boldsymbol{A}', \sqrt{n}\boldsymbol{e}_1).$$

Next we note that $\boldsymbol{A}$ was generated from the sub-sampled Haar model, that is $\boldsymbol{A} = \boldsymbol{H}_m\boldsymbol{S}_{m,n}$ where $\boldsymbol{H}_m \sim \mathrm{Unif}(\mathbb{U}(m))$. Since the Haar measure on $\mathbb{U}(n)$ is invariant to right multiplication by unitary matrices, we have

$$\boldsymbol{H}_m \overset{\mathrm{d}}{=} \boldsymbol{H}_m \cdot \begin{bmatrix} \boldsymbol{\Gamma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{m-n} \end{bmatrix},$$

where the notation $\overset{\mathrm{d}}{=}$ means that two random vectors have the same distributions. Consequently $\boldsymbol{A} = \boldsymbol{H}_m\boldsymbol{S}_{m,n} \overset{\mathrm{d}}{=} \boldsymbol{A}\boldsymbol{\Gamma} = \boldsymbol{A}'$. Therefore, $\rho^2(\boldsymbol{A}, \boldsymbol{x}) = \rho^2(\boldsymbol{A}', \sqrt{n}\boldsymbol{e}_1) \overset{\mathrm{d}}{=} \rho^2(\boldsymbol{A}, \sqrt{n}\boldsymbol{e}_1)$, and the distribution of ρ^2 is independent of $\boldsymbol{x}$. □

In the light of the above lemma, in the rest of the paper, we will assume $\boldsymbol{x} = \sqrt{n}\boldsymbol{e}_1$. Next, we partition $\boldsymbol{A}$ by separating the first column

$$\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_{-1}],$$

where $\boldsymbol{A}_{-1}$ denotes all the remaining columns of $\boldsymbol{A}$ (except $\boldsymbol{A}_1$). Hence we can partition $\boldsymbol{A}^*\boldsymbol{T}\boldsymbol{A}$ in the following way:

$$\boldsymbol{A}^*\boldsymbol{T}\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1^*\boldsymbol{T}\boldsymbol{A}_1 & \boldsymbol{A}_1^*\boldsymbol{T}\boldsymbol{A}_{-1} \\ \boldsymbol{A}_{-1}^*\boldsymbol{T}\boldsymbol{A}_1 & \boldsymbol{A}_{-1}^*\boldsymbol{T}\boldsymbol{A}_{-1} \end{bmatrix}. \tag{3.6}$$

Our strategy will be to reduce questions about the spectrum of the matrix $\boldsymbol{M}$ to questions about the spectrum of a matrix of the form $\boldsymbol{X} = \boldsymbol{E}\boldsymbol{U}\boldsymbol{F}\boldsymbol{U}^*$, where $\boldsymbol{U}$ is a uniformly random unitary matrix, $\boldsymbol{E}$ is a random matrix independent of $\boldsymbol{U}$ and $\boldsymbol{F}$ is deterministic. This matrix model has been well studied in Free Probability [79]. The starting point of our reduction is Proposition 2 from [29], stated below.

Proposition 3 ([29])**.** *Let $\boldsymbol{D}$ be an arbitrary deterministic symmetric matrix partitioned as:*

$$\boldsymbol{D} = \begin{bmatrix} a & \boldsymbol{q}^* \\ \boldsymbol{q} & \boldsymbol{P} \end{bmatrix}.$$

Then, we have

$$\lambda_1(D) = L(\vartheta_\star),$$

where $L(\vartheta) = \lambda_1(\boldsymbol{P} + \vartheta\boldsymbol{q}\boldsymbol{q}^*)$, and $\vartheta_\star > 0$ is the unique solution to the fixed point equation $L(\vartheta) = \frac{1}{\vartheta} + a$. Furthermore, let $\boldsymbol{v}_1$ be the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{D}$. Then,

$$|\boldsymbol{e}_1^*\boldsymbol{v}_1|^2 \in \left[\frac{\partial_- L(\vartheta_\star)}{\partial_- L(\vartheta_\star) + (1/\vartheta_\star)^2}, \frac{\partial_+ L(\vartheta_\star)}{\partial_+ L(\vartheta_\star) + (1/\vartheta_\star)^2}\right],$$

where ∂_- and ∂_+ denote the left and right derivatives respectively. In particular, if $L(\vartheta)$ is differentiable at $\vartheta_\star$, then

$$|\boldsymbol{e}_1^*\boldsymbol{v}_1|^2 = \frac{L'(\vartheta_\star)}{L'(\vartheta_\star) + (1/\vartheta_\star)^2}.$$

A straightforward corollary of the above proposition to our problem is given below. Define the function

$$L_m(\vartheta) \triangleq \lambda_1\left(\boldsymbol{A}_{-1}^*(\boldsymbol{T}+\vartheta \boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{A}_{-1}\right).$$

Corollary 7. *Let $\vartheta_m > 0$ be the unique solution of $L_m(\vartheta) = 1/\vartheta + \boldsymbol{A}_1^*\boldsymbol{T}\boldsymbol{A}_1$. Then, $\lambda_1(\boldsymbol{A}^*\boldsymbol{T}\boldsymbol{A}) = L_m(\vartheta_m)$ and*

$$|\boldsymbol{e}_1^*\hat{\boldsymbol{x}}|^2 \in \left[\frac{\partial_- L_m(\vartheta_m)}{\partial_- L_m(\vartheta_m) + (1/\vartheta_m)^2}, \frac{\partial_+ L_m(\vartheta_m)}{\partial_+ L(\vartheta_m) + (1/\vartheta_m)^2}\right].$$

In particular, if $L_m(\vartheta)$ is differentiable at ϑ_m, then

$$|\boldsymbol{e}_1^*\hat{\boldsymbol{x}}|^2 = \frac{L_m'(\vartheta_m)}{L_m'(\vartheta_m) + (1/\vartheta_m)^2}.$$

Hence, we shift our focus to characterizing the function $L_m(\vartheta)$. Recall the decomposition of the matrix $\boldsymbol{M}$ given in (3.6). Recall that since $\boldsymbol{x} = \sqrt{n}\boldsymbol{e}_1$, the diagonal matrix $\boldsymbol{T}$ is a deterministic function of $\boldsymbol{A}_1$. If the sensing matrix $\boldsymbol{A}$ consisted of independent Gaussian entries, then $\boldsymbol{T}, \boldsymbol{A}_1$ would have been independent of $\boldsymbol{A}_{-1}$. This is no longer true when $\boldsymbol{A}$ is a partial unitary matrix. In order to take care of this, the following lemma leverages a conditioning trick to get rid of the dependence. The following lemma also establishes the link between the function $L_m(\vartheta)$ and the study of the spectrum of a matrix of the form $\boldsymbol{X} = \boldsymbol{E}\boldsymbol{U}\boldsymbol{F}\boldsymbol{U}^*$, where $\boldsymbol{U}$ is a uniformly random unitary matrix, $\boldsymbol{E}$ is a random matrix independent of $\boldsymbol{U}$ and $\boldsymbol{F}$ is deterministic.

Lemma 7. *We have*

$$L_m(\vartheta) = \lambda_1\left(\boldsymbol{B}^*(\boldsymbol{T}+\vartheta \boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{B}\boldsymbol{H}_{m-1}\boldsymbol{R}\boldsymbol{H}_{m-1}^*\right), \tag{3.7}$$

where

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I}_{n-1} & \boldsymbol{0}_{n-1,m-n} \\ \boldsymbol{0}_{m-n,n-1} & \boldsymbol{0}_{m-n,m-n} \end{bmatrix},$$

$\boldsymbol{B} \in \mathbb{C}^{m\times m-1}$ *is an arbitrary basis matrix for* $\boldsymbol{A}_1^{\perp}$*, which denotes the subspace orthogonal to* $\boldsymbol{A}_1$*, and* $\boldsymbol{H}_{m-1} \sim \mathrm{Unif}(\mathbb{U}(m-1))$ *is independent of* $\boldsymbol{A}_1$.

Proof. We condition on $\boldsymbol{A}_1$. Conditioned on $\boldsymbol{A}_1$, we can realize $\boldsymbol{A}_{-1}$ as:

$$\boldsymbol{A}_{-1} = \boldsymbol{B}\boldsymbol{H}_{m-1}\boldsymbol{S}_{m-1,n-1}.$$

In the above equation, $\boldsymbol{B} \in \mathbb{C}^{m\times m-1}$ is matrix whose columns form an orthonormal basis of the orthogonal complement of $\boldsymbol{A}_1$ and $\boldsymbol{H}_{m-1}$ is a Haar Unitary of size $m-1$ independent of $\boldsymbol{A}_1$. Hence, we obtain

$$\begin{aligned} L_m(\vartheta) &= \lambda_1\left(\boldsymbol{A}_{-1}^*(\boldsymbol{T} + \vartheta\boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{A}_{-1}\right) \\ &\overset{a}{=} \lambda_1\left(\boldsymbol{B}^*(\boldsymbol{T} + \vartheta\boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{B}\cdot\boldsymbol{H}_{m-1}\boldsymbol{R}\boldsymbol{H}_{m-1}^*\right). \end{aligned}$$

In the step marked (a), We used the fact that for any two matrices $\boldsymbol{\Lambda}, \boldsymbol{\Gamma}$ (of appropriate dimensions), $\boldsymbol{\Lambda}\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}\boldsymbol{\Lambda}$ have the same non-zero eigenvalues. In particular, we used this fact with:

$$\begin{aligned} \boldsymbol{\Lambda} &= \boldsymbol{S}_{m-1,n-1}^*\boldsymbol{H}_{m-1}^* \\ \boldsymbol{\Gamma} &= \boldsymbol{B}^*(\boldsymbol{T} + \vartheta\boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{B}\boldsymbol{H}_{m-1}\boldsymbol{S}_{m-1,n-1}. \end{aligned}$$

□

Define the matrix,

$$\boldsymbol{E}(\vartheta) \overset{\Delta}{=} \boldsymbol{B}^*(\boldsymbol{T} + \vartheta\boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{B}. \tag{3.8}$$

The following lemma characterizes the asymptotic limit of the function $L_m(\vartheta)$. Define $\Lambda_+(\tau)$ as

$$\Lambda_+(\tau) = \begin{cases} \tau - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\tau-T}\right]} & \text{if } \tau > \tau_r, \\ \min_{\tau \geq 1}\left(\tau - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\tau-T}\right]}\right) & \text{if } \tau \leq \tau_r, \end{cases}$$

where $T = \mathcal{T}(|Z|/\sqrt{\delta})$ and $Z \sim \mathcal{CN}(0,1)$, and

$$\tau_r \triangleq \arg\min_{\tau \geq 1}\left(\tau - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\tau-T}\right]}\right).$$

Lemma 8. *Let* $\vartheta_c \triangleq \left(1 - \left(\mathbb{E}\left[\frac{|Z|^2}{1-T}\right]\right)^{-1} - \mathbb{E}[|Z|^2T]\right)^{-1}$. *Define the function* $\theta(\vartheta)$ *as:*

- *When* $\vartheta > \vartheta_c$*: Let* $\theta(\vartheta)$ *be the unique value of* λ *that satisfies the equation:*

$$\lambda - \mathbb{E}[|Z|^2T] - 1/\vartheta = \left(\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]\right)^{-1},$$

  in the interval:

$$\lambda \in \Big(\max(1, \mathbb{E}[|Z|^2T] + 1/\vartheta), \infty\Big).$$

- *When* $\vartheta \leq \vartheta_c$*:* $\theta(\vartheta) \triangleq 1.$

Then, we have $L_m(\vartheta) \xrightarrow{a.s.} \Lambda_+(\theta(\vartheta))$*, where* $L_m(\vartheta)$ *is defined in* (3.7).

The proof of Lemma 8 can be found in Section 3.4.5.

From Corollary 7, we know that $\lambda_1(\boldsymbol{M})$ solves the fixed point equation (in ϑ): $L_m(\vartheta) = 1/\vartheta + \boldsymbol{A}_1^*\boldsymbol{T}\boldsymbol{A}_1$. Simple concentration arguments (see Lemma 12, Section 3.4.3) show that asymptotically:

$$\boldsymbol{A}_1^*\boldsymbol{T}\boldsymbol{A}_1 \approx \mathbb{E}|Z|^2T.$$

Combining this with Lemma 8 suggests that asymptotically $\lambda_1(\boldsymbol{M})$ behaves like the solution to the following fixed point equation (in ϑ):

$$\Lambda_+(\theta(\vartheta)) = 1/\vartheta + \mathbb{E}|Z|^2 T.$$

The following lemma analyzes the behavior of this asymptotic fixed point equation. The proof of this lemma can be found in Section 3.4.5.

Lemma 9. *The following hold for the equation:*

$$\Lambda_+(\theta(\vartheta)) = 1/\vartheta + \mathbb{E}[|Z|^2 T], \ \vartheta > 0.$$

1. *This equation has a unique solution.*

2. *Let $\vartheta_\star$ denote the solution of the above equation. Then:*

Case 1 *If $\psi_1(\tau_r) \le \frac{\delta}{\delta-1}$, we have*

$$\Lambda_+(\theta(\vartheta_\star)) = \Lambda(\tau_r).$$

Furthermore if $\psi_1(\tau_r) < \delta/(\delta-1)$, then,

$$\left.\frac{\mathrm{d}\Lambda_+(\theta(\vartheta))}{\mathrm{d}\vartheta}\right|_{\vartheta=\vartheta_\star} = 0,$$

Case 2 *If $\psi_1(\tau_r) > \frac{\delta}{\delta-1}$, we have*

$$\Lambda_+(\theta(\vartheta_\star)) = \Lambda(\theta_\star),$$

and,

$$\left.\frac{\mathrm{d}\Lambda_+(\theta(\vartheta))}{\mathrm{d}\vartheta}\right|_{\vartheta=\vartheta_\star} =$$

$$\frac{1}{\vartheta_\star^2} \cdot \frac{\delta}{\delta-1} \cdot \left(\frac{\delta}{\delta-1} - \psi_2(\theta_\star)\right) \cdot \frac{1}{\psi_3^2(\theta_\star) - \frac{\delta^2}{(\delta-1)^2}}.$$

where $\theta_\star > 1$ is the unique $\theta \geq \tau_r$ that satisfies $\psi_1(\theta) = \frac{\delta}{\delta-1}$.

We are now in the position to prove our main result (restated below for convenience). Recall the definitions of the functions $\Lambda(\tau), \psi_1(\tau), \psi_2(\tau), \psi_3(\tau)$ from Section 3.2.

Theorem 1 *Define $\tau_r \triangleq \arg\min_{\tau\in[1,\infty)} \Lambda(\tau)$. Also, let $\theta_\star$ denote the unique value of $\theta > \tau_r$ that satisfies $\psi_1(\theta) = \frac{\delta}{\delta-1}$. Then, we have*

$$\lambda_1(\boldsymbol{M}) \xrightarrow{a.s.} \begin{cases} \Lambda(\tau_r), & \text{if } \psi_1(\tau_r) \leq \frac{\delta}{\delta-1}, \\ \Lambda(\theta_\star), & \text{if } \psi_1(\tau_r) > \frac{\delta}{\delta-1}. \end{cases}$$

Furthermore,

$$|\boldsymbol{e}_1^* \hat{\boldsymbol{x}}|^2 \xrightarrow{a.s.} \begin{cases} 0, & \text{if } \psi_1(\tau_r) < \frac{\delta}{\delta-1}, \\ \frac{\left(\frac{\delta}{\delta-1}\right)^2 - \frac{\delta}{\delta-1}\cdot\psi_2(\theta_\star)}{\psi_3(\theta_\star)^2 - \frac{\delta}{\delta-1}\cdot\psi_2(\theta_\star)}, & \text{if } \psi_1(\tau_r) > \frac{\delta}{\delta-1}. \end{cases}$$

Proof. We start with the analysis of the largest eigenvalue. We recall the claim of Corollary 7, which tells us that $\lambda_1(\boldsymbol{M})$ is given by $L_m(\vartheta_m)$ where ϑ_m denotes the solution of $L_m(\vartheta) = 1/\vartheta + a_m$ and $a_m = \boldsymbol{A}_1^* \boldsymbol{T} \boldsymbol{A}_1$.

We also know that there exists a probability 1 event $\mathcal{E}$, on which, $L_m(\vartheta) \xrightarrow{\text{a.s.}} \Lambda_+(\theta(\vartheta))$ (Lemma 8) and $a_m \xrightarrow{\text{a.s.}} \mathbb{E}[|Z|^2 T]$ (see Lemma 12 in Section 3.4.3).

We claim that on $\mathcal{E}$, $\vartheta_m \to \vartheta_\star$, where $\vartheta_\star$ is the solution of the limiting fixed point equation $\Lambda_+(\theta(\vartheta)) = 1/\vartheta + \mathbb{E}[|Z|^2 T]$ (which was analyzed in Lemma 9). To see this let $\overline{\vartheta} = \limsup \vartheta_m$. Consider a subsequence $\vartheta_{m_k} \to \overline{\vartheta}$. Then applying Lemma 3 (in Appendix E) of [29], we obtain,

$$\begin{aligned} 0 &= \lim_{k\to\infty} \left(L_{m_k}(\vartheta_{m_k}) - \frac{1}{\vartheta_{m_k}} - a_{m_k} \right) \\ &= \Lambda_+(\theta(\overline{\vartheta})) - \frac{1}{\overline{\vartheta}} - \mathbb{E}|Z|^2 T. \end{aligned}$$

That is, $\overline{\vartheta}$ is also a solution to the limiting fixed point equation $\Lambda_+(\theta(\vartheta)) = 1/\vartheta + \mathbb{E}[|Z|^2 T]$. But since this equation has a unique solution (Lemma 9), we have $\limsup \vartheta_m = \overline{\vartheta} = \vartheta_\star$. Likewise, an analogous argument shows $\liminf \vartheta_m = \vartheta_\star$.

Now for any realization in the event $\mathcal{E}$, we have,

$$\lambda_1(\boldsymbol{M}) = L_m(\vartheta_m) \overset{\text{(a)}}{\to} \Lambda_+(\theta(\vartheta_\star)).$$

In the above display, in the step marked (a), we again appealed to Lemma 3 (Appendix E) of [29] and the fact that $\vartheta_m \to \vartheta_\star$. Finally, appealing to the alternative characterization of $\Lambda_+(\theta(\vartheta_\star))$ given in Lemma 9 gives us the claim of the theorem.

We now discuss our result about the cosine similarity. We recall that from Corollary 7, we have

$$|\boldsymbol{e}_1^* \hat{\boldsymbol{x}}|^2 \in \left[\frac{\partial_- L_m(\vartheta_m)}{\partial_- L_m(\vartheta_m) + (1/\vartheta_m)^2}, \frac{\partial_+ L_m(\vartheta_n)}{\partial_+ L(\vartheta_m) + (1/\vartheta_m)^2} \right].$$

Appealing to Lemma 4 in Appendix E of [29], we have,

$$\partial_- L_m(\vartheta_m) \to \partial_- \Lambda_+(\theta(\vartheta_\star)), \ \partial_+ L_m(\vartheta_m) \to \partial_+ \Lambda_+(\theta(\vartheta_\star)).$$

The derivative of $\Lambda_+(\theta(\vartheta))$ at $\vartheta = \vartheta_\star$ was calculated in Lemma 9. Plugging this in the above expression gives the statement of the theorem. □

The remainder of this section is dedicated to the proof of Lemmas 8 and 9, and is organized as follows:

- Recall that (cf. 3.7)

  $$L_m(\vartheta) = \lambda_1 \left(\boldsymbol{E}(\vartheta) \boldsymbol{H}_{m-1} \boldsymbol{R} \boldsymbol{H}_{m-1}^* \right),$$

  where

  $$\boldsymbol{E}(\vartheta) \overset{\Delta}{=} \boldsymbol{B}^* (\boldsymbol{T} + \vartheta \boldsymbol{T} \boldsymbol{A}_1 (\boldsymbol{T} \boldsymbol{A}_1)^*) \boldsymbol{B}.$$

  Note that $\boldsymbol{E}(\vartheta)$ is independent of $\boldsymbol{H}_{m-1}$. The spectrum of such a matrix product has been studied in free probability theory, and we collect some results regarding this in Section 3.4.2.

- In order to apply the free probability results, we need to understand the spectrum of $\boldsymbol{E}(\vartheta)$. This is done in Section 3.4.3.

- It turns out that the limiting spectrum measure of $\boldsymbol{E}(\vartheta)\boldsymbol{H}_{m-1}\boldsymbol{R}\boldsymbol{H}_{m-1}^*$ is given by the free convolution (defined in Section 3.4.2) of the measures γ and $\mathcal{L}_T$, where $\gamma \triangleq \frac{1}{\delta}\delta_1 + \left(1 - \frac{1}{\delta}\right)\delta_0$ and $\mathcal{L}_T$ is the law of the random variable $T = \mathcal{T}(|Z|/\sqrt{\delta})$. Section 3.4.4 is devoted to understanding the support of the free convolution.

- Finally, Section 3.4.5 proves lemmas 8 and 9.

3.4.2 Free Probability Background

Our analysis of the spectral estimators relies on a well-studied model in the theory of free probability; We will reduce the problem to the problem of understanding the spectrum of matrices of the form $\boldsymbol{X} = \boldsymbol{E}\boldsymbol{U}\boldsymbol{F}\boldsymbol{U}^*$, where $\boldsymbol{E}$ and $\boldsymbol{F}$ are deterministic matrices and $\boldsymbol{U}$ is a Haar-distributed unitary matrix. Then, the limiting spectral distribution of $\boldsymbol{X}$ is the *free multiplicative convolution* of the limiting spectral distributions of $\boldsymbol{E}$ and $\boldsymbol{F}$. This section is a collection of the results and definitions regarding these aspects. Here is the organization of this section. Section 3.4.2 collects various facts from free harmonic analysis. Section 3.4.2 describes the two fundamental results about the model $\boldsymbol{X} = \boldsymbol{E}\boldsymbol{U}\boldsymbol{F}\boldsymbol{U}^*$ that will be used throughout our paper. Section 3.4.2 reviews some results about the support of singular part of the free convolution of two measures. Throughout this section, we assume that γ and ν are two arbitrary compactly supported probability measures on $[0, \infty)$ and that neither of the two measures is completely concentrated at a single point.

Facts from Free Harmonic Analysis

In this section, we collect some facts from the field of free harmonic analysis. All these results can be found in Chapter 3 of [80] or the papers [79] and [81].

Definition 3. *The Cauchy transform G_γ of γ at z is defined as follows:*

$$G_\gamma(z) = \int \frac{\gamma(\mathrm{d}t)}{z-t}, \ z \in \mathbb{C} \backslash [0, \infty).$$

Definition 4. *The moment generating function of γ, ψ_γ at z is defined as follows:*

$$\psi_\gamma(z) = \int \frac{zt}{1-zt} \gamma(\mathrm{d}t), \ z \in \mathbb{C} \backslash [0, \infty).$$

The Cauchy transform and the moment generating function are related via the relation

$$G_\gamma(z) = \frac{1}{z} \cdot \left(\psi_\gamma \left(\frac{1}{z} \right) + 1 \right).$$

Definition 5. *The η-transform of a measure is defined as,*

$$\eta_\gamma(z) = \frac{\psi_\gamma(z)}{1 + \psi_\gamma(z)}.$$

The Cauchy Transform (and hence the Moment Generating function) uniquely characterizes a measure. The measure can be obtained by the following inversion formula. The particular version we state is taken from Section 3.1 of [79].

Theorem 4. *For $a < b \in [0, \infty)$, we have*

$$\gamma((a,b)) + \frac{1}{2}\gamma(\{a,b\}) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \int_a^b \operatorname{Im}(G_\gamma(x - i\epsilon)) \,\mathrm{d}x.$$

Furthermore, if γ satisfies $\gamma = \gamma_{ac} + \gamma_s$, where γ_{ac} and γ_s denote the absolutely continuous and the singular part of the measure with respect to the Lebesgue measure, then the density of the absolutely continuous part is given by

$$\frac{\mathrm{d}\gamma_{ac}}{\mathrm{d}x}(x) = \lim_{\epsilon \to 0^+} \frac{1}{\pi} \operatorname{Im}(G_\gamma(x - i\epsilon)).$$

Next we recall the definition of the free convolution based on the subordination functions from [82]. The statement we provide below appears in a more general form as Proposition 2.6 in [83].

Definition 6. *Let (γ, ν) be a pair of probability measures. There exist analytic functions w_γ, w_ν defined on $\mathbb{C}\backslash[0, \infty)$ such that, for all $z \in \mathbb{C}^+$ we have*

1. $w_\gamma(z), w_\nu(z) \in \mathbb{C}^+$*;* $w_\gamma(\overline{z}) = \overline{w_\gamma(z)}, w_\nu(\overline{z}) = \overline{w_\nu(z)}$ *and* $\mathsf{Arg}(w_\gamma(z)) \geq \mathsf{Arg}(z), \mathsf{Arg}(w_\nu(z)) \geq \mathsf{Arg}(z)$.

2. *For any $z \in \mathbb{C}^+$, $w_\nu(z)$ is the unique solution in $\mathbb{C}^+$ of the fixed point equation $Q_z(w) = w$, where Q_z is given by*

$$Q_z(w) = \frac{w}{\eta_\nu(w)}\eta_\gamma\left(\frac{z\eta_\nu(w)}{w}\right).$$

   An analogous characterization holds for w_γ with the role of γ and ν changed.

The free convolution of the measures γ and ν denoted by $\gamma \boxtimes \nu$ is the measure whose moment generating function satisfies

$$\psi_{\gamma\boxtimes\nu}(z) = \psi_\gamma(w_\gamma(z)) = \psi_\nu(w_\nu(z)) = \frac{w_\gamma(z)w_\nu(z)}{z - w_\gamma(z)w_\nu(z)}.$$

Remark 15. *We emphasize that each of the subordination functions w_γ, w_ν depend on both the measures γ, ν. This is clear since the function $Q_z(w)$ defining w_ν depends on both ν, γ.*

Note that the above definition defines w_ν and w_γ on $\mathbb{C}\backslash[0, \infty)$. However these functions can be continously extended to $\overline{\mathbb{C}^+} \cup \{\infty\}$ (Lemma 3.2 in [79]). These extensions to the real line will be important for Theorem 3.4.2.

Lemma 10. *The restrictions of subordination functions w_γ, w_ν on $\mathbb{C}^+$ have extensions to $\overline{\mathbb{C}^+} \cup \{\infty\}$ with the following properties:*

1. *$w_\gamma, w_\nu : \overline{\mathbb{C}^+} \cup \{\infty\} \to \overline{\mathbb{C}^+} \cup \{\infty\}$ are continuous.*

2. *If $1/x \in [0,\infty)\backslash Supp(\gamma \boxtimes \nu)$, then the functions w_γ, w_ν continue analytically to a neighborhood of x and*

$$\frac{1}{w_\gamma(x)} = \frac{w_\nu(x)}{x} \cdot \frac{1+\psi_\nu(w_\nu(x))}{\psi_\nu(w_\nu(x))} \in \mathbb{R}\backslash Supp(\gamma),$$
$$\frac{1}{w_\nu(x)} = \frac{w_\gamma(x)}{x} \cdot \frac{1+\psi_\gamma(w_\gamma(x))}{\psi_\gamma(w_\gamma(x))} \in \mathbb{R}\backslash Supp(\nu).$$

Spectrum of $\mathbf{X} = \mathbf{EUFU}^*$

As we discussed before, we will convert the problem of analyzing the spectrum of $\boldsymbol{M}$ to problems involving the spectrum of matrices of the form $\mathbf{X}_N = \mathbf{E}_N\mathbf{U}_N\mathbf{F}_N\mathbf{U}_N^*$, where $\boldsymbol{U}_N$ is a sequence of Haar distributed $N \times N$ random matrices, and $\boldsymbol{E}_N$ and $\boldsymbol{F}_N$ are sequences of deterministic positive semidefinite matrices. In this section, we review two important results from the field of free probability regarding such matrices.

Suppose that $\boldsymbol{E}_N$ and $\boldsymbol{F}_N$ satisfy the following hypotheses:

(i) $\mu_{\boldsymbol{E}_N} \overset{\mathrm{d}}{\to} \mu_e$ and $\mu_{\boldsymbol{F}_N} \overset{\mathrm{d}}{\to} \mu_f$, where μ_e, μ_f are compactly supported measures on $[0,\infty)$.

(ii) $\boldsymbol{E}_N$ has a single outlying eigenvalue θ not contained in $\mathrm{Supp}(\mu_e)$. $\boldsymbol{F}_N$ has no eigenvalues outside $\mathrm{Supp}(\mu_f)$.

(iii) The set of eigenvalues of $\boldsymbol{E}_N$ not equal to θ converge uniformly to $\mathrm{Supp}(\mu_e)$ in the sense,

$$\lim_{N\to\infty} \max_{i:\lambda_i(\boldsymbol{E}_N)\neq\theta} \mathrm{dist}(\lambda_i(\boldsymbol{E}_N), \mathrm{Supp}(\mu_e)) = 0.$$

Our next theorem characterizes the bulk distribution of $\boldsymbol{X}_N$. The first part of this theorem is due to [84] and the second and third parts are due to [79] (Theorem 2.3).

Theorem 5. *Let w_e and w_f denote the subordination functions for the free multiplicative convolution of μ_e and μ_f. Define*

$$\tau_e(1/z) = \frac{1}{w_e(1/z)}, \ K = Supp(\mu_e \boxtimes \nu_f) \cup \tau_e^{-1}(\theta).$$

Then we have, almost surely for large enough N,

1. $\mu_{\boldsymbol{X}_N} \xrightarrow{d} \mu_e \boxtimes \mu_f$.

2. *Given $\epsilon > 0$, we have $\sigma(\boldsymbol{X}_N) \subset K_\epsilon$, where K_ϵ is the ϵ-neighborhood of K and $\sigma(\boldsymbol{X}_N)$ denotes the set of eigenvalues of $\boldsymbol{X}_N$.*

3. *For any $\rho \in \tau_e^{-1}(\theta)$ such that $\exists \epsilon > 0$ with $(\rho - 2\epsilon, \rho + 2\epsilon) \cap K = \{\rho\}$, we have $|\sigma(\boldsymbol{X}_N) \cap (\rho - \epsilon, \rho + \epsilon)| = 1$.*

Remark 16. *The hypothesis in the above theorem can be relaxed (as mentioned in Remark 5.11 of [79]) in the following two ways: 1) $\boldsymbol{E}_N$ is random, independent of $\boldsymbol{U}_N$ and $\boldsymbol{F}_N$ is deterministic, provided $\mu_{\boldsymbol{E}_N} \xrightarrow{d} \mu_e$ occurs almost surely, 2) The spike locations depend on N, θ_N provided $\theta_N \longrightarrow \theta$ almost surely.*

Remark 17. *The above theorem is a simplified version of Theorem 2.3 in [79] which allows for multiple spikes in both $\boldsymbol{E}_N$ and $\boldsymbol{F}_N$.*

Remark 18. *The function τ might not be invertible. In such cases, $\tau^{-1}(\theta)$ can be a non-singleton set, and hence a single spike in $\boldsymbol{E}_N$ can create multiple spikes in $\boldsymbol{X}_N$. But we will see that this doesn't happen in our problem.*

Singular Part of Free Convolution

In the last section we discussed the bulk distribution of $\boldsymbol{X}_N = \boldsymbol{E}_N\boldsymbol{U}_N\boldsymbol{F}_N\boldsymbol{U}_N$. The main objective of this section is to mention a result regarding the largest eigenvalue of $\boldsymbol{X}_N$. We state regularity results for the singular part of $\gamma \boxtimes \nu$ from [85] (Corollary 3.4) and [81] (Theorem 4.1).

Theorem 6 (Singular Part of $\gamma \boxtimes \nu$)**.** *Decompose the singular part of $\gamma \boxtimes \nu$ as $(\gamma \boxtimes \nu)_s = (\gamma \boxtimes \nu)_d + (\gamma \boxtimes \nu)_{sc}$ where $(\gamma \boxtimes \nu)_d$ denotes the discrete part and $(\gamma \boxtimes \nu)_{sc}$ denotes the singular continous part. Then we have,*

1. *There can be at most two atoms. The possible locations of the atoms are:*

   (a) 0*, with* $\gamma \boxtimes \nu(\{0\}) = \max(\gamma(\{0\}), \nu(\{0\}))$.

   (b) *Any $a \in (0, \infty)$ such that there exist $u, v \in (0, \infty)$ with $uv = a$ and $\gamma(\{u\}) + \nu(\{v\}) > 1$ and we have, $\gamma \boxtimes \nu(\{a\}) = \gamma(\{u\}) + \nu(\{v\}) - 1$. Note that there can be atmost one such a.*

2. *Suppose neither of γ, ν is completely concentrated at a single point. We have, Supp$((\gamma \boxtimes \nu)_{sc}) \subset$ Supp$((\gamma \boxtimes \nu)_{ac})$. Hence,*

$$\mathit{Supp}(\gamma \boxtimes \nu) = \mathit{Supp}((\gamma \boxtimes \nu)_{ac}) \cup \mathit{Supp}((\gamma \boxtimes \nu)_d).$$

3.4.3 Analysis of the Spectrum of $\mathbf{E}(\vartheta)$

In order to apply Theorem 5, we need to understand the spectrum of $\boldsymbol{B}^*(\boldsymbol{T} + \vartheta \boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{B}$. This is done in the following lemma.

Lemma 11. *Let*

$$T_{(1)} \geq T_{(2)} \cdots \geq T_{(m)}$$

denote the sorted trimmed measurements. Let $\boldsymbol{E}(\vartheta) \triangleq \boldsymbol{B}^*(\boldsymbol{T} + \vartheta \boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{B}$. Then,

1. *The eigenvalues of $\boldsymbol{E}(\vartheta)$ interlace with $T_{(1)}, T_{(2)} \ldots T_{(m)}$ in the sense,*

$$\lambda_i(\boldsymbol{E}(\vartheta)) \leq T_{(i-1)} \; \forall \; i = 2, 3, \ldots m, \;\; \&$$

$$\lambda_i(\boldsymbol{E}(\vartheta)) \geq T_{(i+1)} \; \forall \; i = 1, 3, \ldots m - 1.$$

2. *$\boldsymbol{E}(\vartheta)$ can have at most one eigenvalue bigger than $T_{(1)}$, which (if it exists) is given by the root of the following equation:*

$$Q_m(\lambda) = \frac{1}{\lambda - a_m - 1/\vartheta}, \ \lambda > \max(a_m + 1/\vartheta, T_{(1)}),$$

where $Q_m(\lambda)$ is defined as

$$Q_m(\lambda) \triangleq \sum_{i=1}^{m} \frac{|A_{1i}|^2}{\lambda - T_i}.$$

3. *Furthermore, $\lambda_1(\boldsymbol{E}(\vartheta)) \leq 1 + \vartheta$ and $\lambda_{m-1}(\boldsymbol{E}(\vartheta)) \geq 0$.*

Proof. Define the matrix $\boldsymbol{E}(\vartheta) = \boldsymbol{B}^*(\boldsymbol{T} + \vartheta \boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{B}$. The main trick will be to choose the orthonormal basis matrix $\boldsymbol{B}$ conveniently, which will make our calculations easier. Recall that the columns of matrix $\boldsymbol{B}$, i.e. $\boldsymbol{B}_1, \boldsymbol{B}_2 ... \boldsymbol{B}_{m-1}$, span the subspace $\boldsymbol{A}_1^{\perp}$. Any basis for subspace $\boldsymbol{A}_1^{\perp}$ can serve as matrix $\boldsymbol{B}$. Hence, we chose the following specific construction of $\boldsymbol{B}$:

$$\boldsymbol{B}_1 = \frac{\boldsymbol{T}\boldsymbol{A}_1 - a_m \boldsymbol{A}_1}{\sqrt{b_m - a_m^2}},$$

where $a_m = \boldsymbol{A}_1^* \boldsymbol{T} \boldsymbol{A}_1$ and $b_m = \boldsymbol{A}_1^* \boldsymbol{T}^2 \boldsymbol{A}_1$. With this choice, we note that

$$\begin{aligned}\boldsymbol{B}^* \boldsymbol{T} \boldsymbol{A}_1 &= [\boldsymbol{B}_1^* \boldsymbol{T} \boldsymbol{A}_1, \boldsymbol{B}_2^* \boldsymbol{T} \boldsymbol{A}_1 ... \boldsymbol{B}_{m-1}^* \boldsymbol{T} \boldsymbol{A}_1]^* \\ &= \sqrt{b_m - a_m^2 \boldsymbol{e}_1}.\end{aligned}$$

Hence $\boldsymbol{E}(\vartheta) = \boldsymbol{B}^* \boldsymbol{T} \boldsymbol{B} + \vartheta(b_m - a_m^2)\boldsymbol{e}_1 \boldsymbol{e}_1^*$. To obtain the eigenvalues of $\boldsymbol{E}(\vartheta)$ we use its characteristic polynomial. To evaluate the characteristic polynomial of $\boldsymbol{E}(\vartheta)$, we connect it to the characteristic polynomial of $\boldsymbol{O}^*\boldsymbol{T}\boldsymbol{O}$, where $\boldsymbol{O} = [\boldsymbol{A}_1, \boldsymbol{B}]$. Note that $\boldsymbol{O}$ is a unitary matrix. First, we have

$$\begin{aligned}\boldsymbol{O}^*\boldsymbol{T}\boldsymbol{O} &= \begin{bmatrix} \boldsymbol{A}_1^*\boldsymbol{T}\boldsymbol{A}_1 & \boldsymbol{A}_1^*\boldsymbol{T}\boldsymbol{B} \\ \boldsymbol{B}^*\boldsymbol{T}\boldsymbol{A}_1 & \boldsymbol{B}^*\boldsymbol{T}\boldsymbol{B} \end{bmatrix} \\ &= \begin{bmatrix} a_m & \sqrt{b_m - a_m^2}\boldsymbol{e}_1^* \\ \sqrt{b_m - a_m^2}\boldsymbol{e}_1 & \boldsymbol{B}^*\boldsymbol{T}\boldsymbol{B} \end{bmatrix}.\end{aligned}$$

Consider the following matrix equation:

$$\begin{aligned}\begin{bmatrix} a_m + \frac{1}{\vartheta} & \mathbf{0}^* \\ \mathbf{0} & \boldsymbol{E}(\vartheta) \end{bmatrix} &= \begin{bmatrix} a_m + \frac{1}{\vartheta} & \mathbf{0}^* \\ \mathbf{0} & \boldsymbol{B}^*\boldsymbol{T}\boldsymbol{B} \end{bmatrix} \\ &\qquad + \vartheta(b_m - a_m^2)\boldsymbol{e}_2\boldsymbol{e}_2^* \\ &= \begin{bmatrix} a_m & \sqrt{b_m - a_m^2}\boldsymbol{e}_1^* \\ \sqrt{b_m - a_m^2}\boldsymbol{e}_1 & \boldsymbol{B}^*\boldsymbol{T}\boldsymbol{B} \end{bmatrix} \\ &\qquad + \begin{bmatrix} 1/\vartheta & -\sqrt{b_m - a_m^2} & \mathbf{0}_{m-2,1}^* \\ -\sqrt{b_m - a_m^2} & \vartheta(b_m - a_m^2) & \mathbf{0}_{m-2,1}^* \\ \mathbf{0}_{m-2,1} & \mathbf{0}_{m-2,1} & \mathbf{0}_{m-2,m-2} \end{bmatrix} \\ &= \boldsymbol{O}^*\boldsymbol{T}\boldsymbol{O} + \begin{bmatrix} 1/\sqrt{\vartheta} \\ -\sqrt{\vartheta(b_m - a_m^2)} \\ \mathbf{0}_{m-2,1} \end{bmatrix}\begin{bmatrix} 1/\sqrt{\vartheta} \\ -\sqrt{\vartheta(b_m - a_m^2)} \\ \mathbf{0}_{m-2,1} \end{bmatrix}^* \\ &= \boldsymbol{O}^*(\boldsymbol{T} + \boldsymbol{u}\boldsymbol{u}^*)\boldsymbol{O}, \end{aligned} \tag{3.9}$$

where

$$\begin{aligned}
\boldsymbol{u} = \boldsymbol{O} \cdot \begin{bmatrix} 1/\sqrt{\vartheta} \\ -\sqrt{\vartheta(b_m - a_m^2)} \\ \mathbf{0}_{m-2,1} \end{bmatrix} &= \frac{1}{\sqrt{\vartheta}}\boldsymbol{A}_1 - \sqrt{\vartheta(b_m - a_m^2)}\boldsymbol{B}_1 \\
&= \left(\frac{1}{\sqrt{\vartheta}} + a_m\sqrt{\vartheta}\right)\boldsymbol{A}_1 - \sqrt{\vartheta}\boldsymbol{T}\boldsymbol{A}_1
\end{aligned}$$

Therefore,

$$|u_i|^2 = \frac{(1 + a_m\vartheta - \vartheta T_i)^2 |A_{1i}|^2}{\vartheta}.$$

Now, we can compute the characteristic polynomial of $\boldsymbol{E}(\vartheta)$. We have

$$\begin{aligned}
&\det(\lambda\boldsymbol{I} - \boldsymbol{E}(\vartheta)) \\
&\quad = \frac{1}{\lambda - a_m - \frac{1}{\vartheta}} \det\left(\lambda\boldsymbol{I} - \begin{bmatrix} a_m + \frac{1}{\vartheta} & \mathbf{0}^* \\ \mathbf{0} & \boldsymbol{E}(\vartheta) \end{bmatrix}\right) \\
&\quad = \frac{1}{\lambda - a_m - 1/\vartheta} \cdot \det(\lambda\boldsymbol{I} - \boldsymbol{T} - \boldsymbol{u}\boldsymbol{u}^*) \\
&\quad = \frac{\det(\lambda I - \boldsymbol{T})}{\lambda - a_m - 1/\vartheta} \cdot (1 - \boldsymbol{u}^*(\lambda\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{u}).
\end{aligned}$$

Note that

$$
\begin{aligned}
1 - \boldsymbol{u}^*(\lambda \boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{u} &= 1 - \sum_{i=1}^{m} \frac{|u_i|^2}{\lambda - T_i} \\
&= 1 - \frac{1}{\vartheta} \sum_{i=1}^{m} \frac{(1 + a_m\vartheta - \lambda\vartheta + (\lambda - T_i)\vartheta)^2 |A_{1i}|^2}{\lambda - T_i} \\
&= 1 - \frac{(1 + a_m\vartheta - \lambda\vartheta)^2}{\vartheta} \cdot \left(\sum_{i=1}^{m} \frac{|A_{1i}|^2}{\lambda - T_i} \right) \\
&\quad - \vartheta \cdot \left(\sum_{i=1}^{m} (\lambda - T_i) \cdot |A_{1i}|^2 \right) - 2(1 + a_m\vartheta - \lambda\vartheta) \\
&= -(1 - \lambda\vartheta + a_m\vartheta) \cdot \left(1 + \frac{1 - \lambda\vartheta + a_m\vartheta}{\vartheta} Q_m(\lambda) \right),
\end{aligned}
$$

Where $Q_m(\lambda)$ is defined in the following way:

$$
Q_m(\lambda) \triangleq \sum_{i=1}^{m} \frac{|A_{1i}|^2}{\lambda - T_i}.
$$

Hence,

$$
\begin{aligned}
&\det(\lambda \boldsymbol{I} - \boldsymbol{E}(\vartheta)) = \\
&\qquad \det(\lambda \boldsymbol{I} - \boldsymbol{T})(\vartheta + (1 - \lambda\vartheta + a_m\vartheta) Q_m(\lambda)).
\end{aligned}
\tag{3.10}
$$

We emphasize that the above equation *does not imply* that $T_1, T_2, \ldots, T_m$ are the eigenvalues of $\boldsymbol{E}(\vartheta)$. This is because while $\det(\lambda \boldsymbol{I} - \boldsymbol{T})$ has zeros at T_i, the function $Q_m(\lambda)$ has poles at T_i. This prevents us from concluding that $\det(\lambda \boldsymbol{I} - \boldsymbol{E}(\vartheta)) = 0$ when $\lambda = T_i$. However, we can make the following observations:

1. By Cauchy's interlacing theorem, we have

$$\begin{aligned}\lambda_1(\boldsymbol{T}+\vartheta(\boldsymbol{T}\boldsymbol{A}_1)(\boldsymbol{T}\boldsymbol{A}_1)^*) \geq T_{(1)} & \\ \geq \lambda_2(\boldsymbol{T}+\vartheta(\boldsymbol{T}\boldsymbol{A}_1)(\boldsymbol{T}\boldsymbol{A}_1)^*) & \\ \geq T_{(2)}. & \qquad (3.11)\end{aligned}$$

   The above is also true for the eigenvalues of:

$$\boldsymbol{O}^*(\boldsymbol{T}+\vartheta(\boldsymbol{T}\boldsymbol{A}_1)(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{O},$$

   since $\boldsymbol{O}$ is a unitary matrix.

2. (3.9) shows that $\boldsymbol{E}(\vartheta)$ is a principal submatrix of

$$\boldsymbol{O}^*(\boldsymbol{T}+\vartheta(\boldsymbol{T}\boldsymbol{A}_1)(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{O}.$$

   Hence, the eigenvalues of $\boldsymbol{E}(\vartheta)$ will interlace the eigenvalues of $\boldsymbol{O}^*(\boldsymbol{T}+\vartheta(\boldsymbol{T}\boldsymbol{A}_1)(\boldsymbol{T}\boldsymbol{A}_1)^*)\boldsymbol{O}$:

$$\begin{aligned}\lambda_1(\boldsymbol{T}+\vartheta(\boldsymbol{T}\boldsymbol{A}_1)(\boldsymbol{T}\boldsymbol{A}_1)^* \geq \lambda_1(\boldsymbol{E}(\vartheta)) & \\ \geq \lambda_2(\boldsymbol{T}+\vartheta(\boldsymbol{T}\boldsymbol{A}_1)(\boldsymbol{T}\boldsymbol{A}_1)^* & \\ \geq \lambda_2(\boldsymbol{E}(\vartheta)). & \qquad (3.12)\end{aligned}$$

   Combining (3.11) and (3.12), one obtains

$$\lambda_2(\boldsymbol{E}(\vartheta)) \leq T_{(1)},\ \lambda_1(\boldsymbol{E}(\vartheta)) \geq T_{(2)}.$$

   This proves statement (1) in the lemma. This means that $\boldsymbol{E}(\vartheta)$ has atmost one eigenvalue bigger than $T_{(1)}$. If $\lambda_1(\boldsymbol{E}(\vartheta)) \leq T_{(1)}$, then it has no outlying eigenvalue, if $\lambda_1(\boldsymbol{E}(\vartheta)) > T_{(1)}$, it has exactly one. We call this eigenvalue an outlying eigenvalue for reasons that will be clear later.

3. The outlying eigenvalue of $\boldsymbol{E}(\vartheta)$ (if it exists) is a root of the characteristic polynomial:

$$\det(\lambda \boldsymbol{I} - \boldsymbol{E}(\vartheta)) =$$
$$\det(\lambda I - \boldsymbol{T}) \cdot (\vartheta + (1 - \lambda\vartheta + a_m\vartheta) Q_m(\lambda)).$$

Since this root lies in $(T_{(1)}, \infty)$, it must be a root of:

$$Q_m(\lambda) = \frac{1}{\lambda - a_m - 1/\vartheta}, \ \lambda > T_{(1)}. \tag{3.13}$$

Observing that:

$$\lambda > T_{(1)} \implies Q_m(\lambda) > 0,$$
$$\lambda > a_m + 1/\vartheta \implies (\lambda - a_m - 1/\vartheta)^{-1} > 0,$$

we conclude the outlying eigenvalue is the unique solution (if it exists) to:

$$Q_m(\lambda) = \frac{1}{\lambda - a_m - 1/\vartheta}, \ \lambda > \max(a_m + 1/\vartheta, T_{(1)}).$$

This proves statement (2).

4. Finally, we observe that $\boldsymbol{E}(\vartheta)$ is a positive semidefinite matrix for all $\vartheta \geq 0$, which shows $\lambda_{m-1}(\boldsymbol{E}(\vartheta)) \geq 0$. Also, we have $\lambda_1(\boldsymbol{E}(\vartheta)) \leq \|\boldsymbol{E}(\vartheta)\| \leq \|\boldsymbol{B}\|^2 \|\boldsymbol{T} + \vartheta \boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*\|$. Note that $\|\boldsymbol{B}\| \leq 1$ and $\|\boldsymbol{T}\| \leq 1$ and $\|\boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*\| = \boldsymbol{A}_1^* \boldsymbol{T}^2 \boldsymbol{A}_1 \leq T_{(1)}^2 \leq 1$. Hence, by the triangle inequality we have $\lambda_1(\boldsymbol{E}(\vartheta)) \leq 1 + \vartheta$. This proves statement (3) of the lemma.

□

The following lemma analyzes the concentration of the function $Q_m(\lambda)$ to the deterministic function $Q(\lambda)$.

Lemma 12. *Suppose $\frac{m}{n} = \delta$. For a Lipschitz function $\mathcal{T}$ whose range is in $[0, 1]$, there exists an event of probability 1, on which the following three statements hold:*

1. $\frac{1}{m}\sum_{i=1}^{m} \delta_{T_i} \xrightarrow{d} \mathcal{L}_T$,

2. $Q_m(\lambda) \to Q(\lambda) \quad \forall\, \lambda \in (1, \infty)$,

3. $a_m \longrightarrow \mathbb{E}|Z|^2 T$.

In the above equations, $Z \sim \mathcal{CN}(0, 1)$, and $T = \mathcal{T}(|Z|/\sqrt{\delta})$. Furthermore, $\mathcal{L}_T$ denotes the law of the random variable T, and

$$Q(\lambda) = \mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right].$$

Proof. It is sufficient to show each item holds almost surely.

1. The argument for this part is a minor modification of the argument sketched in [86]. To prove statement (1) it suffices to show that

$$\frac{1}{m}\sum_{i=1}^{n} \delta_{\sqrt{m}|A_{i1}|} \xrightarrow{d} Z, \tag{3.14}$$

almost surely. Because if we have (3.14), then for every bounded continuous function f,

$$f\left(\mathcal{T}\left(\sqrt{n}|A_{i1}|\right)\right) = g\left(\sqrt{m}|A_{1i}|\right),$$

where $g(x) = f(\mathcal{T}(\frac{|x|}{\sqrt{\delta}}))$ is a bounded continuous function as well. Hence by (3.14),

$$\frac{1}{m}\sum_{i=1}^{m} f(T_i) \to \mathbb{E}\left[g(Z)\right] = \mathbb{E}\left[f\left(\mathcal{T}\left(\frac{Z}{\sqrt{\delta}}\right)\right)\right],$$

which implies $\frac{1}{m}\sum_{i=1}^{m} \delta_{T_i} \xrightarrow{d} \mathcal{L}_T$.

To show (3.14), note that $\boldsymbol{A}_1$ has the same distribution as $\frac{z}{\|z\|}$, where $z = (z_1, ..., z_m)$, and $z_i \overset{i.i.d.}{\sim} \mathcal{CN}(0,1)$. Let Φ denote the cumulative distribution function of a standard normal random variable and define

$$\begin{aligned} F_m(t) &\triangleq \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left(\sqrt{m}|A_{1i}| \leq t\right), \\ G_m(t) &\triangleq \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\left(z_i \leq t\right). \end{aligned}$$

Then, we have

$$F_m(t) \overset{d}{=} G_m\left(t\frac{\|z\|}{\sqrt{m}}\right). \tag{3.15}$$

Moreover,

$$\begin{aligned} & G_m\left(t\frac{\|z\|}{\sqrt{m}}\right) - \Phi(t) = \\ & \quad G_m\left(t\frac{\|z\|}{\sqrt{m}}\right) - \Phi\left(t\frac{\|z\|}{\sqrt{m}}\right) + \Phi\left(t\frac{\|z\|}{\sqrt{m}}\right) - \Phi(t) \\ & \quad \xrightarrow{a.s.} 0 + 0. \end{aligned}$$

$G_m(t\|z\|) - \Phi(t\|z\|)$ goes to 0 almost surely by Glivenko-Cantelli lemma. Furthermore, since

$$\frac{\|z\|}{\sqrt{m}} \xrightarrow{a.s.} 1,$$

and Φ is a continuous function we conclude that

$$\Phi\left(t\frac{\|z\|}{\sqrt{m}}\right) - \Phi(t) \overset{a.s.}{\to} 0.$$

Hence,

$$F_m(t) \to \Phi(t),$$

almost surely which yields (3.14).

2. We now focus on the proof of statement (2). Let

$$C_k \triangleq \left[1+\frac{1}{k}, k\right], \quad k \in \mathbb{N}.$$

We will show that

$$Q_m(\lambda) \to Q(\lambda) \quad \forall\, \lambda \;\in\; C_k, \tag{3.16}$$

almost surely. This means there is a set C'_k, with measure 0, out of which we have the convergence for all $\lambda \in C_k$. If we define $C' \triangleq \bigcup\limits_{k=1}^{\infty} C'_k$, then $Q_m(\lambda) \to Q(\lambda) \quad \forall \lambda \in (1,\infty)$ out of C' and clearly $\mathbb{P}\left(C'\right) = 0$.

First note that $\boldsymbol{A}_1 \overset{d}{=} \frac{z}{\|z\|}$, where

$$\boldsymbol{z} = (z_1, ..., z_m)\,, \quad z_i \overset{i.i.d.}{\sim} \mathcal{CN}(0,1).$$

Define

$$\tilde{Q}_m(\lambda) \triangleq \frac{1}{m}\sum_{i=1}^{m} \frac{|z_i|^2}{\lambda - \mathcal{T}\left(\frac{|z_i|}{\sqrt{\delta}}\right)}. \tag{3.17}$$

Note that for a fixed λ we have $\tilde{Q}_m(\lambda) \to Q(\lambda)$ almost surely by the strong law of large numbers. Since $\tilde{Q}_m(\lambda)$ is a decreasing function in λ and we have $\tilde{Q}_m(\lambda) \to Q(\lambda) \quad \forall \lambda \in C_k \cap \mathbb{Q}$ almost surely, we obtain $\tilde{Q}_m(\lambda) \to Q(\lambda)$ for all $\lambda \in C_k$ with probability 1. Hence, it suffices to show under an event that holds with probability 1,

$$Q_m(\lambda) - \tilde{Q}_m(\lambda) \to 0 \quad \forall \lambda \in C_k. \tag{3.18}$$

To prove (3.18), we will find a sequence τ_m such that $\tau_m \to 0$ as $m \to \infty$, and,

$$\sum_{m\geq 1} \mathbb{P}\left(\sup_{\lambda\in C_k}\left|Q_m(\lambda) - \tilde{Q}_m(\lambda)\right| > \tau_m\right) < \infty.$$

With this, Borel-Cantelli lemma yields that event

$$E = \left\{ \sup_{\lambda \in C_k} \left| Q_m(\lambda) - \tilde{Q}_m(\lambda) \right| > \tau_m \text{ infinitely often} \right\}$$

has measure 0. Out of the event E we have (3.18) as it was desired.

Define the events:

$$E_1 \triangleq \left\{ \sup_{i \le m} |z_i| \le \sqrt{6 \log m} \right\},$$
$$E_{2,\epsilon} \triangleq \left\{ \left| \frac{\|z\|^2}{m} - 1 \right| \le \epsilon \right\},$$

where ϵ is parameter we will set later. Note that,

$$\begin{aligned} &\left| Q_m(\lambda) - \tilde{Q}_m(\lambda) \right| \le \\ &\qquad \sum_{i=1}^{m} \frac{|z_i|^2}{\|z\|^2} \left| \frac{\frac{\|z\|^2}{m}}{\lambda - \mathcal{T}\left(\frac{|z_i|}{\sqrt{\delta}}\right)} - \frac{1}{\lambda - \mathcal{T}\left(\frac{\sqrt{n}}{\|z\|}|z_i|\right)} \right| \\ &\qquad \le \mathsf{I} + \mathsf{II}, \end{aligned}$$

where we defined the terms I, II as:

$$\mathsf{I} = \left| \frac{\|z\|^2}{m} - 1 \right| \cdot \sum_{i=1}^{m} \frac{|z_i|^2}{\|z\|^2} \cdot \left| \frac{1}{\lambda - \mathcal{T}\left(\frac{|z_i|}{\sqrt{\delta}}\right)} \right|$$
$$\mathsf{II} = \sum_{i=1}^{m} \frac{|z_i|^2}{\|z\|^2} \cdot \frac{\left| \mathcal{T}\left(\frac{|z_i|}{\sqrt{\delta}}\right) - \mathcal{T}\left(\frac{\sqrt{n}|z_i|}{\|z\|}\right) \right|}{\left| \lambda - \mathcal{T}\left(\frac{|z_i|}{\sqrt{\delta}}\right) \right| \cdot \left| \lambda - \mathcal{T}\left(\frac{\sqrt{n}|z_i|}{\|z\|}\right) \right|}.$$

Using the fact that $z \in E_1 \cap E_{2,\epsilon}$ and $\lambda \in C_k$, we have,

$$\begin{aligned}
\mathsf{I} &\le k\epsilon, \\
\mathsf{II} &\le k^2 \cdot \max_{i \le n} \left| \mathcal{T}\left(\frac{|z_i|}{\sqrt{\delta}}\right) - \mathcal{T}\big(\frac{\sqrt{n}|z_i|}{\|z\|}\big) \right|.
\end{aligned}$$

Observe that, on the event $E_1 \cap E_{2,\epsilon}$,

$$\begin{aligned}
\left| \frac{|z_i|}{\sqrt{\delta}} - \frac{\sqrt{n}}{\|z\|}|z_i| \right| &\le \frac{|z_i|}{\sqrt{\delta}} \left| 1 - \frac{\sqrt{m}}{\|z\|} \right| \\
&\le \sqrt{6\log(m)} \cdot \left| 1 - \frac{\sqrt{m}}{\|z\|} \right| \\
&\le \sqrt{6\log(m)} \cdot \left| 1 - \frac{m}{\|z\|^2} \right| \\
&\le \sqrt{6\log(m)} \cdot \frac{\epsilon}{1-\epsilon}.
\end{aligned}$$

Since $\mathcal{T}$ was assumed to be Lipchitz,

$$\begin{aligned}
\mathsf{II} &\le k^2 \cdot \max_{i \le n} \left| \mathcal{T}\left(\frac{|z_i|}{\sqrt{\delta}}\right) - \mathcal{T}\big(\frac{\sqrt{n}|z_i|}{\|z\|}\big) \right| \\
&\le k^2 \cdot \|\mathcal{T}\|_{\mathsf{Lip}} \cdot \sqrt{6\log(m)} \cdot \frac{\epsilon}{1-\epsilon},
\end{aligned}$$

where $\|\mathcal{T}\|_{\mathsf{Lip}}$ denotes the Lipchitz constant of $\mathcal{T}$. Hence, when $m \ge e^2$, setting $\epsilon = \frac{1}{\log(m)} \le 0.5$, we obtain, on the event $E_1 \cap E_{2,\epsilon}$

$$\left| Q_m(\lambda) - \tilde{Q}_m(\lambda) \right| \le \tau_m, \ \forall \ \lambda \ \in \ C_k. \tag{3.19}$$

where

$$\tau_m = \frac{k}{\log(m)} + \frac{2k^2 \cdot \|\mathcal{T}\|_{\mathsf{Lip}}}{\sqrt{\log(m)}}.$$

Note that $\tau_m \to 0$ as $m \to \infty$ as required. And,

$$\begin{aligned}\mathbb{P}\left(\sup_{\lambda \in C_k}\left|Q_m(\lambda) - \tilde{Q}_m(\lambda)\right| > \tau_m\right)& \\ &\leq \mathbb{P}\left(E_1^c\right) + \mathbb{P}\left(E_{2,\epsilon}^c\right) \\ &\leq 2 \cdot m^{-2} + 2e^{-\frac{m}{8\log^2(m)}},\end{aligned}$$

where the last step follows from standard bounds on the tail Gaussian random variables and χ^2 random variables. In particular, we have,

$$\sum_{m\geq 1}\mathbb{P}\left(\sup_{\lambda \in C_k}\left|Q_m(\lambda) - \tilde{Q}_m(\lambda)\right| > \tau_m\right) < \infty,$$

as required.

3. The proof is similar to the proof of the second statement. Hence, we skip the details. Note that if we define

$$W_m = \sum_{i=1}^{n}|A_{1i}|^2\,\mathcal{T}(|A_{1i}|\sqrt{n}),$$

then it again converges under the event $E_1 \cap E_{2,\epsilon}$, defined in the proof of statement (2).

□

The next lemma analyzes the properties of the limiting fixed point equation $Q(\lambda) = (\lambda - \mathbb{E}|Z|^2T - 1/\vartheta)^{-1}$. Define the critical value ϑ_c as:

$$\vartheta_c \triangleq \left(1 - \left(\mathbb{E}\left[\frac{|Z|^2}{1-T}\right]\right)^{-1} - \mathbb{E}[|Z|^2T]\right)^{-1} \geq 0.$$

Lemma 13. *Consider the fixed point equation (in λ)*

$$\lambda - \mathbb{E}[|Z|^2T] - 1/\vartheta = \frac{1}{\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]}, \tag{3.20}$$

on the domain:

$$\lambda > \max(1, \mathbb{E}[|Z|^2 T] + 1/\vartheta).$$

We have

1. *If $\vartheta > \vartheta_c$, then the above equation has exactly 1 solution, denoted by $\lambda = \theta(\vartheta)$. Furthermore,*

$$\begin{aligned}\lambda - \mathbb{E}[|Z|^2 T] - 1/\vartheta &> \frac{1}{\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]} \\ &\qquad \forall\, \lambda \in \left(\max(1, \mathbb{E}[|Z|^2 T] + 1/\vartheta), \theta(\vartheta)\right), \\ \lambda - \mathbb{E}[|Z|^2 T] - 1/\vartheta &< \frac{1}{\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]}\ \forall\, \lambda \in \left(\theta(\vartheta), \infty\right).\end{aligned}$$

   Furthermore, we have $\theta(\vartheta)$ is an increasing function of ϑ and $\lim_{\vartheta\to\infty} \theta(\vartheta) = \infty$.

2. *If $\vartheta \le \vartheta_c$, then the equation has no solutions. For any $\vartheta \le \vartheta_c$, we define $\theta(\vartheta) = 1$.*

Proof. The following change of measure simplifies some of the proofs:

$$\begin{aligned}p(z) &\triangleq \frac{|z|^2}{\pi} \exp(-|z|^2), \\ \tilde{\mathbb{E}}[f(Z)] &\triangleq \int f(z) p(z)\, \mathrm{d}z.\end{aligned}$$

Note that $p(z)$ is a proper probability density function since $\int p(z)\, \mathrm{d}z = \mathbb{E}[|Z|^2] = 1$. With this notation, (3.20) can be written as

$$\lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta = \frac{1}{\tilde{\mathbb{E}}\left[\frac{1}{\lambda - T}\right]},\ \lambda > \max(1, \tilde{\mathbb{E}}[T] + 1/\vartheta).$$

Define the random variable $G(\lambda) = (\lambda - T)^{-1}$. Note that $G'(\lambda) = -G^2(\lambda)$. Further, define

$$f(\lambda) \triangleq \frac{1}{\tilde{\mathbb{E}}\left[G(\lambda)\right]};\ \lambda \in [1, \infty).$$

The first two derivatives of $f(\lambda)$ are

$$f'(\lambda) = \frac{\tilde{\mathbb{E}}[G^2]}{\tilde{\mathbb{E}}[G]^2},$$
$$f''(\lambda) = -2 \cdot \frac{\tilde{\mathbb{E}}[G^3]\tilde{\mathbb{E}}[G] - \tilde{\mathbb{E}}[G^2]^2}{\tilde{\mathbb{E}}[G]^3}.$$

First, since $f'(\lambda) \geq 0$, the function $f(\lambda)$ is increasing. By Jensen's Inequality $f'(\lambda) \geq 1$. Since the equality holds if and only if G is deterministic, and we have assumed that the support of T is $[0, 1]$, we conclude that $f(\lambda) > 1$. Noting that $G \geq 0$ and applying Chebychev's association inequality (See Fact 1, Appendix 3.7) with $B = A = G$ and $f(a) = g(a) = a$ gives $f''(\lambda) \leq 0$. Hence $f(\lambda)$ is an increasing, concave function and $f'(\lambda) > 1$.

Next, we claim that $f(\lambda) = \lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta$ can have atmost one solution in $(1, \infty)$. To see this, let λ_1 be the first point at which the two curves intersect. Hence $f(\lambda_1) = \lambda_1 - \tilde{\mathbb{E}}[T] - 1/\vartheta$. Furthermore

$$f'(\lambda) > 1 = \frac{\mathrm{d}(\lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta)}{\mathrm{d}\lambda}.$$

Hence there can be no other intersection point of the two curves after λ_1.

Now consider the following two cases:

Case 1: $\vartheta > \vartheta_c$. First note that since $(1-x)^{-1}$ is a convex function on $(-\infty, 1]$, according to Jensen's Inequality

$$\tilde{\mathbb{E}}\left[\frac{1}{1-T}\right] \geq \frac{1}{1-\tilde{\mathbb{E}}[T]} \geq 0.$$

Hence,

$$\frac{1}{\vartheta_c} = 1 - \left(\tilde{\mathbb{E}}\left[\frac{1}{1-T}\right]\right)^{-1} - \tilde{\mathbb{E}}[T] \geq 0.$$

This shows that $\vartheta_c \geq 0$. Furthermore,

$$\vartheta > \vartheta_c \iff (\lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta)_{\lambda=1} > f(1).$$

On the other hand, we can also compare the limiting behavior of $\lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta$ and $f(\lambda)$ as $\lambda \to \infty$. We have

$$\frac{\lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta}{\lambda} = 1 - \frac{\tilde{\mathbb{E}}[T] + 1/\vartheta}{\lambda},$$

and

$$\begin{aligned}\frac{f(\lambda)}{\lambda} = \frac{1}{\tilde{\mathbb{E}}\left[\frac{1}{1-T/\lambda}\right]} &= \left(\tilde{\mathbb{E}}\left[\sum_{n=0}^{\infty}\left(\frac{T}{\lambda}\right)^n\right]\right)^{-1} \\ &= \left(1 + \tilde{\mathbb{E}}[T]/\lambda + o(1/\lambda)\right)^{-1} \\ &= 1 - \frac{\tilde{\mathbb{E}}[T]}{\lambda} + o(\lambda^{-1}).\end{aligned}$$

Hence, $f(\lambda) > \lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta$ for λ large enough and $f(1) < 1 - \tilde{\mathbb{E}}[T] - 1/\vartheta$. Hence the functions $f(\lambda)$ and $1 - \tilde{\mathbb{E}}[T] - 1/\vartheta$ intersect once in $(1, \infty)$. Finally note that,

$$\begin{aligned}\frac{1}{\vartheta} + \tilde{\mathbb{E}}[T] < \frac{1}{\vartheta_c} + \tilde{\mathbb{E}}[T] &= 1 - \left(\tilde{\mathbb{E}}\left[\frac{1}{1-T}\right]\right)^{-1} \\ &\leq 1.\end{aligned}$$

Hence $f(\lambda) = \lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta$ has exactly one solution in $\lambda \geq \max(1, \tilde{\mathbb{E}}[T] + 1/\vartheta)$ as claimed. By the Implicit Function Theorem, we can compute

$$\theta'(\vartheta) = \frac{1/\vartheta^2}{f'(\theta(\vartheta)) - 1} \geq 0. \tag{3.21}$$

Hence $\theta(\vartheta)$ is an increasing function of ϑ. Finally, we verify that $\lim_{\vartheta\to\infty} \theta(\vartheta) = \infty$. Suppose that this is not the case, i.e. $\theta(\vartheta) \to \theta_\infty < \infty$ as $\vartheta \to \infty$. Recalling the fixed point characterization of $\theta(\vartheta)$, we obtain that θ_∞ satisfies the fixed point equation

$$\theta_\infty - \tilde{\mathbb{E}}[T] = \frac{1}{\tilde{\mathbb{E}}\left[\frac{1}{\theta_\infty - T}\right]}.$$

This means that Jensen's Inequality applied to the strictly convex function $(\theta_\infty - t)^{-1}$ should be tight. This means under the tilted measure ($\tilde{\mathbb{E}}$), T is deterministic. This is not possible since we have assumed that T is supported on $[0, 1]$.

Case 2: $\vartheta \le \vartheta_c$ As in Case 1 we argue (this time with the opposite conclusion) that

$$\vartheta \le \vartheta_c \implies f(1) \ge (\lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta)_{\lambda=1}$$

Furthermore, since $f'(\lambda) > \frac{\mathrm{d}(\lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta)}{\mathrm{d}\lambda} = 1$, $f(\lambda) = \lambda - \tilde{\mathbb{E}}[T] - 1/\vartheta$ has no solution in $(1, \infty)$. □

Combining the above sequence of lemmas, we obtain the following proposition about the spectrum of the matrix $\boldsymbol{E}(\vartheta)$.

Proposition 4. *Let $\boldsymbol{E}(\vartheta) = \boldsymbol{B}^*(\boldsymbol{T} + \vartheta \boldsymbol{T}\boldsymbol{A}_1(\boldsymbol{T}\boldsymbol{A}_1)^*))\boldsymbol{B}$. Then, there exists an event of probability 1, on which we have,*

1. $\mu_{\boldsymbol{E}(\vartheta)} \xrightarrow{d} \mathcal{L}_T$.

2. *If $\vartheta \le \vartheta_c$, $\sigma(\boldsymbol{E}(\vartheta)) \subset [0, 1]$.*

3. *If $\vartheta > \vartheta_c$, then $\lambda_i(\boldsymbol{E}(\vartheta)) \in [0, 1]\ \forall\ i \ge 2$, and,*

$$\lambda_1(\boldsymbol{E}(\vartheta)) \xrightarrow{a.s.} \theta(\vartheta),$$

where $\theta(\vartheta)$ is the unique solution to the equation (in λ):

$$\lambda - \mathbb{E}[|Z|^2 T] - 1/\vartheta = \frac{1}{\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]},$$

in the domain:

$$\lambda > \max(1, \mathbb{E}[|Z|^2 T] + 1/\vartheta).$$

Proof. We restrict ourselves to the event guaranteed by Lemma 12, on which,

1. $a_m \longrightarrow \mathbb{E}|Z|^2 T$
2. $\frac{1}{m}\sum_{i=1}^{m} \delta_{T_i} \xrightarrow{\mathrm{d}} \mathcal{L}_T$
3. $Q_m(\lambda) \rightarrow Q(\lambda)\ \forall\ \lambda \in (1, \infty)$.

Let us denote this event by $\mathcal{E}$. Define the sequence of (random) functions $f_m(\lambda)$ as:

$$f_m(\lambda) = \lambda - a_m - 1/\vartheta - \left(\sum_{i=1}^{m} \frac{|A_{1i}|^2}{\lambda - T_i}\right)^{-1},$$

with the domain:

$$\lambda > \max(1, a_m + 1/\vartheta).$$

Define the (deterministic) function $f(\lambda)$:

$$f(\lambda) = \lambda - \mathbb{E}[|Z|^2 T] - 1/\vartheta - \left(\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]\right)^{-1},$$

with the domain:

$$\lambda > \max(1, \mathbb{E}[|Z|^2 T] + 1/\vartheta).$$

Note that on $\mathcal{E}$, we have $f_m(\lambda) \rightarrow f(\lambda)\ \forall\ \lambda > 1$.

1. By Lemma 11, we know that the eigenvalues of $\boldsymbol{E}(\vartheta)$ interlace with the eigenvalues of the diagonal matrix $\boldsymbol{T}$. On the event $\mathcal{E}$, $\mu_{\boldsymbol{T}} \to \mathcal{L}_T$. Hence indeed $\mu_{\boldsymbol{E}(\vartheta)} \xrightarrow{\text{d}} \mathcal{L}_T$. This proves statement (1) of the proposition.

2. Consider the case $\vartheta \leq \vartheta_c$. By Lemma 11, we already know that $\lambda_2(\boldsymbol{E}(\vartheta)) \leq T_{(1)} \leq 1$ and $\lambda_{m-1}(\boldsymbol{E}(\vartheta)) \geq 0$. Hence to prove (2), it is sufficient to show that

$$\overline{\lambda}_1 \triangleq \limsup_{m\to\infty} \lambda_1(\boldsymbol{E}(\vartheta)) \leq 1, \text{ on } \mathcal{E}.$$

For the sake of contradiction, suppose that there is a realization in $\mathcal{E}$ such that $\overline{\lambda}_1 > 1$. On this realization we consider a subsequence such that $\lambda_1(\boldsymbol{E}(\vartheta)) \to \overline{\lambda}_1$. All the analysis henceforth is along this subsequence. Since for all m large enough $\lambda_1(\boldsymbol{E}(\vartheta)) > 1$, by Lemma 11, we must have $f_m(\lambda_1(\boldsymbol{E}(\vartheta)) = 0$. Applying Lemma 3 from [29] (Appendix E), we obtain

$$0 = f_m(\lambda_1(\boldsymbol{E}(\vartheta)) \to f(\overline{\lambda}_1).$$

Since $\vartheta \leq \vartheta_c$, we know by Lemma 13 that $f(\lambda) = 0$ does not have any solution in $\lambda > \max(1, \mathbb{E}[|Z|^2 T] + 1/\vartheta)$. Hence,

$$1 < \overline{\lambda}_1 \leq \mathbb{E}[|Z|^2 T] + 1/\vartheta.$$

However,

$$\begin{aligned} f(\overline{\lambda}_1) &= \underbrace{\overline{\lambda}_1 - \mathbb{E}[|Z|^2 T] - 1/\vartheta}_{\leq 0} - \left(\underbrace{\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]}_{>0}\right)^{-1} \\ &< 0. \end{aligned}$$

This contradicts $f(\overline{\lambda}_1) = 0$. Hence, $\limsup_{m\to\infty} \lambda_1(\boldsymbol{E}(\vartheta)) \leq 1$, on $\mathcal{E}$. This concludes the proof of statement (2).

3. Now consider the case $\vartheta > \vartheta_c$. Again by Lemma 11, we know $\lambda_i(\boldsymbol{E}(\vartheta)) \in [0, 1]$ for all $i \geq 2$. By Lemma 13, we know that $f(\lambda) = 0$ has a unique solution in $\lambda > \max(1, \mathbb{E}|Z|^2 T + 1/\vartheta)$ denoted by $\theta(\vartheta)$. Fix an ϵ small enough such that $[\theta(\vartheta) - \epsilon, \theta(\vartheta) + \epsilon]$ lies in the domain of $f(\lambda)$. Note that $f(\theta(\vartheta)) = 0$, while $f(\theta(\vartheta) - \epsilon) > 0$ and $f(\theta(\vartheta) + \epsilon) < 0$ (by Lemma 13).

   Since $a_m \to \mathbb{E}|Z|^2 T$, for all m large enough, $[\theta(\vartheta) - \epsilon, \theta(\vartheta) + \epsilon]$ also lies in the domain of $f_m(\lambda)$. By Lemma 12, we have $f_m(\lambda) \to f(\lambda)$ for all $\lambda \in [\theta(\vartheta) - \epsilon, \theta(\vartheta) + \epsilon]$. In particular, we have, for all n large enough $f_m(\theta(\vartheta) - \epsilon) > 0$ while $f_m(\theta(\vartheta) + \epsilon) < 0$. Hence, by Lemma 11, we have $\lambda_1(\boldsymbol{E}(\vartheta)) \in [\theta(\vartheta) - \epsilon, \theta(\vartheta) + \epsilon]$ for all n large enough. Hence indeed, $\lambda_1(\boldsymbol{E}(\vartheta)) \overset{\text{a.s.}}{\to} \theta(\vartheta)$. This proves (3).

□

3.4.4 Analysis of the Support of $\gamma \boxtimes \mathcal{L}_T$

We recall that $\mathcal{L}_T$ is the law of the random variable $T = \mathcal{T}(|Z|/\sqrt{\delta})$, and $\gamma = \frac{1}{\delta}\delta_1 + \left(1 - \frac{1}{\delta}\right)\delta_0$. To keep the notation clean, we will refer to the analytic transforms corresponding to the measure $\mathcal{L}_T$ with the subscript T, for example the Cauchy transform for the measure $\mathcal{L}_T$ will be referred to as G_T.We begin by computing the Cauchy Transform of $\gamma \boxtimes T$.

Lemma 14. *Let $z \in \mathbb{C}^-$. Then, we have,*

$$G_{\gamma \boxtimes T}(z) = \frac{1}{z} \cdot \frac{1 - 1/\delta}{1 - z w_T(1/z)}.$$

In the above display, the subordination function, $w_T(1/z)$, is the unique solution in $\mathbb{C}^+$ to the equation $\Lambda(1/w) = z$, where the function Λ is defined as:

$$\Lambda(\tau) \triangleq \tau - \frac{(1 - 1/\delta)}{\mathbb{E}\left[\frac{1}{\tau - T}\right]}.$$

Proof. First we can compute the moment generating functions:

$$\psi_\gamma(z) = \frac{1}{\delta}\cdot\frac{z}{1-z},$$
$$\psi_T(z) = -1 + \mathbb{E}\left[\frac{1}{1-zT}\right].$$

The η-transforms of the two measures are given by,

$$\eta_\gamma(z) = \frac{z/\delta}{z/\delta - z + 1},$$
$$\eta_T(z) = \frac{\mathbb{E}\left[\frac{zT}{1-zT}\right]}{\mathbb{E}\left[\frac{1}{1-zT}\right]}.$$

Hence, we can compute the function Q_z, given in Definition 6,

$$Q_z(w) = \frac{1/\delta}{(1/\delta - 1)\frac{\mathbb{E}\left[\frac{T}{1-wT}\right]}{\mathbb{E}\left[\frac{1}{1-wT}\right]} + 1/z}.$$

Hence w_T is the unique solution in $\mathbb{C}^+$ of the equation $Q_z(w) = w$. This equation can be simplified to

$$\frac{1}{z} = \Lambda(1/w),$$

where the function Λ is defined as $\Lambda(\tau) \overset{\Delta}{=} \tau - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\tau-T}\right]}$. Hence, we can compute the moment generating function of $\gamma \boxtimes T$ in the following way:

$$\begin{aligned}\psi_{\gamma\boxtimes T}(z) &= \psi_T(w_T(z)) \\ &= -1 + \mathbb{E}\left[\frac{1}{1-w_T(z)T}\right] \\ &\overset{\text{(a)}}{=} -1 + \frac{1-1/\delta}{1-w_T(z)/z}.\end{aligned}$$

In the above display, in the step marked (a), we used the fact that w_T solves $\Lambda(1/w) = 1/z$. Finally, the Cauchy Transform of $\gamma \boxtimes T$ is given by

$$\begin{aligned} G_{\gamma\boxtimes T}(z) &= \frac{1}{z}\left(\psi_{\gamma\boxtimes T}\left(\frac{1}{z}\right) + 1\right) \\ &= \frac{1}{z} \cdot \frac{1 - 1/\delta}{1 - zw_T(1/z)}. \end{aligned}$$

□

Our next goal is to characterize $\mathrm{Supp}(\gamma \boxtimes T)$. Theorem 6 gives a complete characterization of the support of the singular part of $\gamma \boxtimes T$. Hence, we now need to understand the support of the absolutely continuous part of $\gamma \boxtimes T$. According to the Stieltjes Inversion theorem, (Theorem 4) the density of the continuous part is given by

$$\begin{aligned} \frac{\mathrm{d}(\gamma \boxtimes T)_{ac}}{\mathrm{d}x}(x) &= \frac{1}{\pi}\lim_{\epsilon\to 0^+} \mathrm{Im}\ G_{\gamma\boxtimes T}(x - i\epsilon) \\ &= \frac{1}{\pi x}\mathrm{Im}\left(\frac{1 - \frac{1}{\delta}}{1 - x\lim_{\epsilon\to 0^+} w_T(1/(x - i\epsilon))}\right). \end{aligned}$$

Since $\tau_T(x - i\epsilon) \triangleq 1/w_T(1/(x - i\epsilon))$ uniquely solves $\Lambda(\tau) = x - i\epsilon$ in $\mathbb{C}^-$, our interest will be to study the solutions of this equation for $\epsilon \approx 0$. Hence, we begin by studying the solutions of $\Lambda(\tau) = x$. Before doing so, we clarify the definition of $\Lambda(\tau)$ at $\tau = 1$ which is a subtle case because $1 \in \mathrm{Supp}(T)$. We note that the random variable $(1 - T)^{-1}$ is non-negative and hence the expectation $\mathbb{E}[(1 - T)^{-1}]$ is well defined but might be ∞. If it is finite, then $\Lambda(\tau)$ is well defined at $\tau = 1$. If the expectation is ∞, we define $\Lambda(1) = 1$ which is consistent with intepreting $1/\infty = 0$. $\Lambda(\tau)$ is defined at $\tau = 0$ analogously. This definition ensures $\Lambda(\tau)$ is a continuous function on $(-\infty, 0] \cup [1, \infty)$. Next we discuss the solutions of $\Lambda(\tau) = x$. Figure 3.2 shows a typical plot $\Lambda(\tau)$. As is clear from this figure we expect the following two quantities to play major roles in determining the existence of a solution of $\Lambda(\tau) = x$: Define

$$\lambda_l = \max_{\tau\in(-\infty,0]} \Lambda(\tau),\ \ \tau_l = \arg\max_{\tau\in(-\infty,0]} \Lambda(\tau)$$

$$\lambda_r = \min_{\tau\in[1,\infty)} \Lambda(\tau),\ \ \tau_r = \arg\min_{\tau\in[1,\infty)} \Lambda(\tau).$$

Our next lemma proves the properties of $\Lambda(\tau)$ suggested by Figure 3.2.

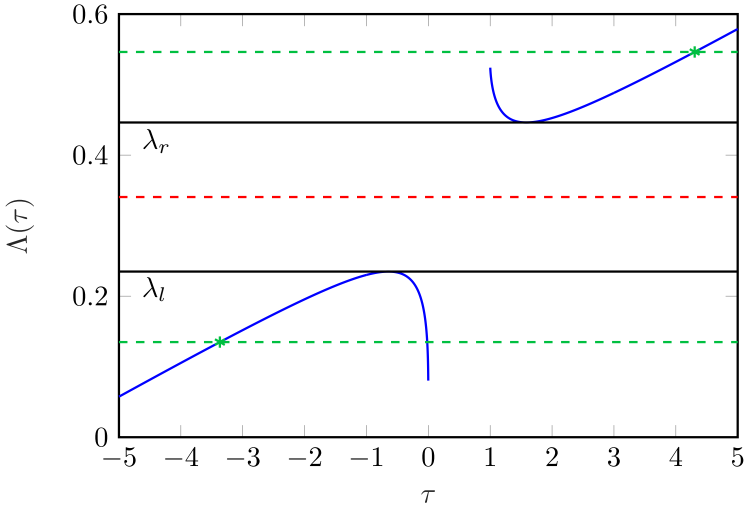


Figure 3.2: An Illustrative plot of the function $\Lambda(\tau)$: When $\lambda_l < x < \lambda_r$, the equation $\Lambda(\tau) = x$ has no solutions. When $x \geq \lambda_r$, the equation $\Lambda(\tau) = x, \Lambda'(\tau) > 0$ has a unique solution in $[1, \infty)$. When $x < \lambda_l$, then $\Lambda(\tau) = x, \Lambda'(\tau) > 0$ has a unique solution in $(-\infty, 0]$.

Lemma 15. *The following statements are true about* $\Lambda(\tau)$*:*

1. $\Lambda(\tau)$ *is a convex function on* $[1, \infty)$ *and a concave function on* $(-\infty, 0]$.

2. $\lim_{\tau\to\infty} \Lambda(\tau) = \infty,\ \lim_{\tau\to-\infty} \Lambda(\tau) = -\infty.$

3. $\lambda_r > \lambda_l \geq 0.$

4. *Consider the 3 mutually exclusive and exhaustive cases:*

   Case A: $x \leq \lambda_l$*. There is at least one and at most two solutions to* $\Lambda(\tau) = x$*. All solutions lie in* $(-\infty, 0]$*. Furthermore, when* $x < \lambda_l$*, there is exactly one solution for the equation* $\Lambda(\tau) = x, \Lambda'(\tau) > 0$*. This unique solution additionally satisfies* $\tau < \tau_l \leq 0$.

Case B: $\lambda_l < x < \lambda_r$. There are no solutions of the equation $\Lambda(\tau) = x,\ \tau \in (-\infty, 0] \cup [1, \infty)$.

Case C: $x \geq \lambda_r$. There is at least one and at most two solutions to $\Lambda(\tau) = x$. All solutions lie in $[1, \infty)$. Furthermore, when, $x > \lambda_r$, there is a unique solution to $\Lambda(\tau) = x, \Lambda'(\tau) > 0$. This solution additionally satisfies $\tau > \tau_r \geq 1$.

Proof. 1. We define the random variable $G(\tau)$,

$$G(\tau) \overset{\Delta}{=} \frac{1}{\tau - T}.$$

We observe that for any $\tau \in [1, \infty)$, $G(\tau) \geq 0$ where as for $\tau \in (-\infty, 0]$, $G(\tau) \leq 0$. It is straightforward to see that $G'(\tau) = -G^2(\tau) \leq 0$. For notational simplicity, we will often short hand $G(\tau)$ as G. We have

$$\Lambda'(\tau) = 1 - \left(1 - \frac{1}{\delta}\right) \cdot \frac{\mathbb{E}G^2}{(\mathbb{E}G)^2},$$
$$\Lambda''(\tau) = 2\left(1 - \frac{1}{\delta}\right) \cdot \frac{(\mathbb{E}G^3) \cdot (\mathbb{E}G) - (\mathbb{E}G^2)^2}{(\mathbb{E}G)^3}.$$

Consider the following two cases, **Case 1: $\tau \in [1, \infty)$.** Applying Chebychev's Association Inequality (Fact 1) with $A = B = G$ and $f(a) = g(a) = a$ gives us that $\Lambda''(\tau) \geq 0$. In fact, an inspection of the proof of the Chebychev's Association Inequality from [87] allows us to rule out the equality case under the assumptions imposed on $\mathcal{T}$, and we have $\Lambda''(\tau) > 0$. Hence, Λ is strictly convex in $(1, \infty)$. Since $\Lambda(\tau)$ is continuous on $[1, \infty)$, we have Λ is convex on $[1, \infty)$ **Case 2: $\tau \in (-\infty, 0]$.** Again, applying Chebychev's Association Inequality with $A = B = -G$ and $f(a) = f(b) = a$ gives us $\Lambda''(\tau) \leq 0$, Hence Λ is concave in this region. As before, an inspection of the proof of Chebychev's Association inequality allows us to rule out the equality case under the assumptions imposed on $\mathcal{T}$, and we have $\Lambda''(\tau) < 0$. Hence, Λ is strictly concave in $(-\infty, 0)$. Since $\Lambda(\tau)$ is continuous on $(-\infty, 0)$, we have Λ is concave on $(-\infty, 0]$. This concludes the proof of statement (1) in the lemma.

2. Note that,

$$\lim_{\tau \to \infty} \tau - \frac{(1 - 1/\delta)}{\mathbb{E}\left[\frac{1}{\tau - T}\right]} = \tau \left(1 - \frac{(1 - 1/\delta)}{\mathbb{E}\left[\frac{\tau}{\tau - T}\right]} \right) = \infty.$$

   This shows $\lim_{\tau \to \infty} \Lambda(\tau) = \infty$. The claim about the limit as $\tau \to -\infty$ can be analogously obtained. This proves item (2) in the statement of the lemma.

3. The infimum in the definition of λ_r is attained due to item (2) in the statement of the lemma. Analogously, the supremum in the definition of λ_l is attained. Next consider any $\tau_+ \in (1, \infty)$ and any $\tau_- \in (-\infty, 0)$. Since the function $f(t) = (\tau_+ - t)^{-1}$ is convex on $[0, 1]$, according to Jensen's Inequality, we have

$$\begin{aligned} \Lambda(\tau_+) &\geq \tau_+ - \left(1 - \frac{1}{\delta}\right) \cdot (\tau_+ - \mathbb{E}[T]) \\ &= \frac{\tau_+}{\delta} + \left(1 - \frac{1}{\delta}\right) \cdot \mathbb{E}[T]. \end{aligned}$$

   On the other hand, since the function $f(t) = (\tau_- - t)^{-1}$ is concave on $[0, 1]$, we have

$$\begin{aligned} \Lambda(\tau_-) &\leq \tau_- - \left(1 - \frac{1}{\delta}\right) \cdot (\tau_- - \mathbb{E}[T]) \\ &= \frac{\tau_-}{\delta} + \left(1 - \frac{1}{\delta}\right) \cdot \mathbb{E}[T]. \end{aligned}$$

   Hence,

$$\begin{aligned} \Lambda(\tau_+) &\geq \frac{1}{\delta} + \left(1 - \frac{1}{\delta}\right) \cdot \mathbb{E}[T] \\ &> \left(1 - \frac{1}{\delta}\right) \cdot \mathbb{E}[T] \\ &\geq \Lambda(\tau_-). \end{aligned}$$

   Taking the minimum over τ_+ and maximum of τ_- gives us $\lambda_r > \lambda_l$. Furthermore we note that $\Lambda(0^-) \geq 0$. Hence $\lambda_l \geq 0$. This concludes the proof of item (3) in the statement of the lemma.

4. For any $x \in (\lambda_l, \lambda_r)$, $\Lambda(\tau) = x$ doesn't have a solution in $(-\infty, 0] \cup [1, \infty)$ since $\Lambda(\tau) \leq \lambda_l \;\forall\; \tau \leq 0$ and $\Lambda(\tau) \geq \lambda_r \;\forall\; \tau \geq 1$. Now consider any $x \geq \lambda_r$. Since $\lambda(\tau) \leq \lambda_l < \lambda_r \;\forall\; \tau \leq 0$, we know that all solutions of $\Lambda(\tau) = x$ lie in $[1, \infty)$. Since Λ is strictly convex in $(1, \infty)$, there can be atmost 2 solutions. Now consider any $x > \lambda_r$. Let $\tau_r = \arg\min_{\tau \geq 1} \Lambda(\tau)$. Due to strict convexity of $\Lambda(\tau)$, we have $\Lambda'(\tau) > 0$ for any $\tau \in (\tau_r, \infty)$. Hence $\Lambda(\tau)$ is strictly increasing on $[\tau_r, \infty)$. Since $\lambda_r = \Lambda(\tau_r) < x < \Lambda(\infty) = \infty$, we are guaranteed to have exactly one solution to $\Lambda(\tau) = x$ on (τ_r, ∞) which indeed satisfies $\Lambda'(\tau) > 0$. The analysis for the case when $x \leq \lambda_l$ can be done in a similar way. This concludes the proof of item (4) in the statement of the lemma.

□

We are now in the position to characterize the support of $\gamma \boxtimes T$ which is the content of the following proposition.

Proposition 5. *The support of $\gamma \boxtimes T$ is given by*

$$Supp(\gamma \boxtimes T) = [\lambda_l, \lambda_r] \cup Supp((\gamma \boxtimes T)_d),$$

where $(\gamma \boxtimes T)_d$ denotes the discrete part of the measure $\gamma \boxtimes T$. If the random variable T has a density with respect to the Lebesgue measure, then,

$$Supp(\gamma \boxtimes T) = [\lambda_l, \lambda_r].$$

Proof. We first claim that $(\lambda_l, \lambda_r) \subset \text{Supp}(\gamma \boxtimes T)$. Since the support of a measure is closed, this means that $[\lambda_l, \lambda_r] \subset \text{Supp}(\gamma \boxtimes T)$. We prove this claim by contradiction. Suppose that $\exists \lambda \in (\lambda_l, \lambda_r)$ such that $\lambda \notin \text{Supp}(\gamma \boxtimes T)$. To simplify notation, for $z \in \mathbb{C}^-$, we introduce the following reciprocal subordination function $\tau_T(z)$

$$\tau_T(z) \overset{\Delta}{=} \frac{1}{w_T(1/z)}.$$

According to Lemma 10, we have

$$\tau_T(\lambda) \overset{\Delta}{=} \lim_{\epsilon \to 0^+} \tau_T(\lambda - i\epsilon) \in (-\infty, 0) \cup (1, \infty).$$

By Lemma 14, $\tau_T(\lambda - i\epsilon)$ uniquely solves the equation $\Lambda(\tau) = \lambda - i\epsilon$ in $\mathbb{C}^-$. Taking $\epsilon \to 0$, we obtain,

$$\begin{aligned}\lambda &= \lim_{\epsilon \to 0^+} \Lambda(\tau_T(\lambda - i\epsilon)) \\ &= \lim_{\epsilon \to 0^+} \left(\tau_T(\lambda - i\epsilon) - \frac{1 - 1/\delta}{\mathbb{E}\left[\frac{1}{\tau_T(\lambda - i\epsilon) - T}\right]} \right) \\ &\overset{\text{(a)}}{=} \tau_T(\lambda) - \frac{1 - 1/\delta}{\mathbb{E}\left[\frac{1}{\tau_T(\lambda) - T}\right]}.\end{aligned}$$

In the step marked (a), we used the fact that since $\lim_{\epsilon \to 0^+} \tau_T(\lambda - i\epsilon) \notin \operatorname{Supp}(T)$, we have $\exists c > 0$, such that for any ϵ small enough $\operatorname{dist}(\tau_T(\lambda - i\epsilon), \operatorname{Supp}(T)) \geq c$. This gives us a dominating function for an application of the dominated convergence theorem. Hence, we have found a solution for the equation $\lambda = \Lambda(\tau), \tau \in (-\infty, 0) \cup (1, \infty)$. But this contradicts Lemma 15. Hence, we have, $(\lambda_l, \lambda_r) \subset \operatorname{Supp}(\gamma \boxtimes T)$.

Next, we claim that any $x \in [0, \lambda_l) \cup (\lambda_r, \infty)$ is not in the support of the absolutely continuous part of $\gamma \boxtimes T$. To show this, we first compute a first order asymptotic expansion of $\tau_T(x - i\epsilon)$ for $\epsilon \approx 0$. From Lemma 15, we know there exists a unique solution for the equation $\Lambda(\tau) = x, \tau \in (-\infty, 0) \cup (1, \infty)$ and $\Lambda'(\tau) > 0$. We denote this solution by $\tau_\star$. Since $\tau_\star \notin \operatorname{Supp}(T)$, the function $\Lambda(\tau)$ is analytic in the neighborhood (in $\mathbb{C}$) of $\tau_\star$. The implicit function theorem guarantees us a solution $\tau(\epsilon) = \tau_R(\epsilon) + i\tau_I(\epsilon)$ of the equation $\Lambda(\tau) = x - i\epsilon$. However, this $\tau(\epsilon)$ may not be the reciprocal subordination function $\tau_T(x - i\epsilon)$ since we still need to verify it is in $\mathbb{C}^-$. To take care of this, again by the implicit function theorem we have

$$\Lambda'(\tau_\star) \cdot \frac{\mathrm{d}\tau}{\mathrm{d}\epsilon}(0) = -i.$$

This gives us

$$\frac{\mathrm{d}\tau_I}{\mathrm{d}\epsilon}(0) = -\frac{1}{\Lambda'(\tau_\star)} < 0, \ \frac{\mathrm{d}\tau_R}{\mathrm{d}\epsilon}(0) = 0.$$

Hence, we have

$$\tau(\epsilon) = \tau_\star - i\frac{\epsilon}{\Lambda'(\tau_\star)} + o(\epsilon).$$

This verifies that $\tau(\epsilon) \in \mathbb{C}^-$ for ϵ small enough. Finally since $\tau_T(x - i\epsilon)$ is the unique solution to the equation $\Lambda(\tau) = x - i\epsilon$ in $\mathbb{C}^-$, we have

$$\tau_T(x - i\epsilon) = \tau_\star - i\frac{\epsilon}{\Lambda'(\tau_\star)} + o(\epsilon).$$

According to the Stieltjes Inversion Formula, Theorem 4, we obtain

$$\begin{aligned} \frac{\mathrm{d}(\gamma \boxtimes T)_{ac}}{\mathrm{d}x}(x) &= \frac{1}{\pi x} \cdot \mathrm{Im}\left(\frac{1 - \frac{1}{\delta}}{1 - x \cdot \lim_{\epsilon \to 0^+} w_T\left(\frac{1}{x - i\epsilon}\right)}\right) \\ &\overset{\text{(b)}}{=} \frac{1}{\pi x} \cdot \mathrm{Im}\left(\frac{(1 - 1/\delta) \cdot \tau_\star}{\tau_\star - x}\right) = 0. \end{aligned}$$

In the step marked (b), we are relying on the assumption that $\tau_\star \neq x$. To verify this, we recall that

$\tau_\star$ solves, $\Lambda(\tau_\star) = x$ and $\tau_\star \notin [0,1]$. This means that

$$\begin{aligned} |\tau_\star - x| &= \frac{1 - 1/\delta}{\left|\mathbb{E}\left[\frac{1}{\tau_\star - T}\right]\right|} \\ &\geq \frac{1 - 1/\delta}{\mathbb{E}\left[\left|\frac{1}{\tau_\star - T}\right|\right]} \\ &\geq (1 - 1/\delta) \cdot \operatorname{dist}(\tau_\star, [0,1]) > 0. \end{aligned}$$

Hence, we have shown

$$\frac{\mathrm{d}(\gamma \boxtimes T)_{ac}}{\mathrm{d}x}(x) \overset{\text{a.s.}}{=} 0, \forall x \in [0, \lambda_l) \cup (\lambda_r, \infty).$$

This implies,

$$[0, \lambda_l) \cup (\lambda_r, \infty) \subset \mathbb{R} \backslash \operatorname{Supp}((\gamma \boxtimes T)_{ac}).$$

Taking complements, we have $\operatorname{Supp}((\gamma \boxtimes T)_{ac}) \subset [\lambda_l, \lambda_r]$. Hence, we have shown that

$$\begin{aligned} [\lambda_l, \lambda_r] \cup \operatorname{Supp}((\gamma \boxtimes T)_d) \subset \operatorname{Supp}(\gamma \boxtimes T) & \\ &= \operatorname{Supp}((\gamma \boxtimes T)_{ac}) \cup \operatorname{Supp}((\gamma \boxtimes T)_d) \\ &\subset [\lambda_l, \lambda_r] \cup \operatorname{Supp}((\gamma \boxtimes T)_d). \end{aligned}$$

Therefore, $\operatorname{Supp}(\gamma \boxtimes T) = [\lambda_l, \lambda_r] \cup \operatorname{Supp}((\gamma \boxtimes T)_d)$ which proves the claim of the proposition. Finally, when T has a density with respect to Lebesgue measure, Theorem 6 gives us $\operatorname{Supp}((\gamma \boxtimes T)_d) = \emptyset$ which yields the second claim in the proposition. □

Finally we note that in order to apply Theorem 5, it is necessary to understand the set $\tau_T^{-1}(\{\theta\}) \cap (\mathbb{R} \backslash \operatorname{Supp}(\gamma \boxtimes T))$, $\theta \in \mathbb{R}$ (See Theorem 5 for a definition of τ_T). This is done in the following lemma.

Lemma 16. *Let (w_γ, w_T) denote the subordination functions corresponding to the free multiplicative convolution of $\gamma, \mathcal{L}_T$. Define*

$$\tau_T(z) = \frac{1}{w_T(1/z)}.$$

Then, we have

$$\tau_T^{-1}(\{\theta\}) \cap (\mathbb{R}\backslash Supp(\gamma \boxtimes T)) = \begin{cases} \theta \in [\tau_l, \tau_r]: & \emptyset \\ \theta \notin [\tau_l, \tau_r]: & \{\Lambda(\theta)\} \end{cases},$$

where where, $\tau_l \triangleq \arg\max_{\tau \le 0} \Lambda(\tau)$, $\tau_r \triangleq \arg\min_{\tau \ge 1} \Lambda(\tau)$.

Proof. From Proposition 5, we know that $\text{Supp}(\gamma \boxtimes T) = [\lambda_l, \lambda_r]$, where $\lambda_l \overset{\Delta}{=} \max_{\tau \le 0} \Lambda(\tau)$ and $\lambda_r \overset{\Delta}{=} \min_{\tau \ge 1} \Lambda(\tau)$. Furthermore, we showed that for any $x \notin [\lambda_l, \lambda_r]$, the reciprocal subordination function $\tau_T(x)$ is the unique solution to the equations: $\Lambda(\tau) = x, \Lambda'(\tau) > 0, \ \tau \notin [0, 1]$. From Lemma 15, we know that when $x > \lambda_r$, the unique solution to $\Lambda(\tau) = x, \Lambda'(x) > 0$ satisfies $\tau > \tau_r$ and when $x < \lambda_l$, the unique solution satisfies $\tau < \tau_l$. These considerations immediately yield the claim of the lemma. □

3.4.5 Proof of Lemmas 8 and 9

Recall we defined $\Lambda_+(\tau)$ as

$$\Lambda_+(\tau) = \begin{cases} \tau - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\tau - T}\right]} & \text{if } \tau > \tau_r, \\ \min_{\tau \ge 1}\left(\tau - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\tau - T}\right]}\right) & \text{if } \tau \le \tau_r, \end{cases}$$

where $T = \mathcal{T}(|Z|/\sqrt{\delta})$ and $Z \sim \mathcal{CN}(0,1)$, and

$$\tau_r \triangleq \arg\min_{\tau \geq 1} \left(\tau - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\tau - T}\right]} \right).$$

We first prove Lemma 8, which we restated below for convenience.

Lemma 3. *Let* $\vartheta_c \overset{\Delta}{=} \left(1 - \left(\mathbb{E}\left[\frac{|Z|^2}{1-T}\right] \right)^{-1} - \mathbb{E}[|Z|^2 T] \right)^{-1}$. *Define the function* $\theta(\vartheta)$ *as:*

- *When* $\vartheta > \vartheta_c$*: Let* $\theta(\vartheta)$ *be the unique value of* λ *that satisfies the equation:*

$$\lambda - \mathbb{E}[|Z|^2 T] - 1/\vartheta = \left(\mathbb{E}\left[\frac{|Z|^2}{\lambda - T} \right] \right)^{-1},$$

in the interval:

$$\lambda \in \left(\max(1, \mathbb{E}[|Z|^2 T] + 1/\vartheta), \infty \right).$$

- *When* $\vartheta \leq \vartheta_c$*:* $\theta(\vartheta) \overset{\Delta}{=} 1$.

Then, we have $L_m(\vartheta) \xrightarrow{a.s.} \Lambda_+(\theta(\vartheta))$*, where* $L_m(\vartheta)$ *is defined in* (3.7).

Proof. In Proposition 11, we obtained an asymptotic characterization of the spectrum of $\boldsymbol{E}(\vartheta)$. More specifically, we proved that

$$\mu_{\boldsymbol{E}(\vartheta)} \xrightarrow{\mathrm{d}} \mathcal{L}_T,\ \lambda_1(\boldsymbol{E}(\vartheta)) \to \theta(\vartheta).$$

We recall the matrix $\boldsymbol{R}$ was defined as

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I}_{n-1} & \boldsymbol{0}_{n-1,m-1} \\ \boldsymbol{0}_{m-n,n-1} & \boldsymbol{0}_{m-1,m-1} \end{bmatrix}.$$

In particular, $\mu_{\boldsymbol{R}} \xrightarrow{\text{d}} \gamma$, where the measure γ is given by

$$\gamma = \frac{1}{\delta}\delta_1 + \left(1 - \frac{1}{\delta}\right)\delta_0.$$

Applying Theorem 5, we obtain:

1. The spectral measure of $\boldsymbol{E}(\vartheta)\boldsymbol{H}_{m-1}\boldsymbol{R}\boldsymbol{H}_{m-1}^*$ converges to:

$$\mu_{\boldsymbol{E}(\vartheta)\boldsymbol{H}_{m-1}\boldsymbol{R}\boldsymbol{H}_{m-1}^*} \xrightarrow{\text{d}} \gamma \boxtimes \mathcal{L}_T.$$

2. For any $\epsilon > 0$, we have, almost surely, for m large enough that, $\sigma(\boldsymbol{E}(\vartheta)\boldsymbol{H}_{m-1}\boldsymbol{R}\boldsymbol{H}_{m-1}^*) \subset K_\epsilon$, where K_ϵ is the ϵ-neighborhood of the set $K = \text{Supp}(\gamma \boxtimes \mathcal{L}_T) \cup \tau_T^{-1}(\{\theta(\vartheta)\})$.

3. For any $\lambda \in \tau_T^{-1}(\{\theta(\vartheta)\}) \cap (\mathbb{R}\backslash\text{Supp}(\gamma \boxtimes \mathcal{L}_T))$, we have almost surely exactly one eigenvalue of $\boldsymbol{E}(\vartheta)\boldsymbol{H}_{m-1}\boldsymbol{R}\boldsymbol{H}_{m-1}^*$ in a small enough neighborhood of λ for large enough n.

In Proposition 5, we characterized $\text{Supp}(\gamma \boxtimes \mathcal{L}_T)$ as $[\lambda_l, \lambda_r]$, where $\lambda_l = \max_{\tau \le 0} \Lambda(\tau)$, $\lambda_r = \min_{\tau \ge 1} \Lambda(\tau)$ and the function $\Lambda(\tau)$ is given by:

$$\Lambda(\tau) = \tau - \frac{(1 - 1/\delta)}{\mathbb{E}\left[\frac{1}{\tau - T}\right]}.$$

In Lemma 16, we characterized the set:

$$\tau_T^{-1}(\{\theta\}) \cap (\mathbb{R}\backslash\text{Supp}(\gamma \boxtimes T)) = \begin{cases} \emptyset & \theta \in [\tau_l, \tau_r], \\ \{\Lambda(\theta)\} & \theta \notin [\tau_l, \tau_r], \end{cases}$$

where, $\tau_l \triangleq \arg\max_{\tau \le 0} \Lambda(\tau)$, $\tau_r \triangleq \arg\min_{\tau \ge 1} \Lambda(\tau)$. Putting these together, one obtains the following two cases:

Case 1: $\theta(\vartheta) \le \tau_r$. In this case, the set $\tau_T^{-1}(\{\theta\}) \cap (\mathbb{R}\backslash\text{Supp}(\gamma \boxtimes T)) = \emptyset$. The matrix $\boldsymbol{E}(\vartheta)\boldsymbol{H}_{m-1}\boldsymbol{R}\boldsymbol{H}_{m-1}^*$

has no eigenvalues outside the support of the bulk distribution, and

$$L_m(\vartheta) \xrightarrow{\text{a.s.}} \lambda_r = \Lambda(\tau_r).$$

Case 2: $\theta(\vartheta) > \tau_r$. In this case, the set

$$\tau_T^{-1}(\{\theta\}) \cap (\mathbb{R}\backslash\mathrm{Supp}(\gamma \boxtimes T)) = \{\Lambda(\theta(\vartheta))\}.$$

Hence, there is an eigenvalue in the neighborhood of $\Lambda(\theta(\vartheta)))$. Since $\theta(\vartheta) > \tau_r$, and Λ is a strictly increasing function on $[\tau_r, \infty)$ (Lemma 15), we have $\Lambda(\theta(\vartheta)) > \lambda_r$. Hence the eigenvalue in the neighborhood of $\Lambda(\theta(\vartheta))$ is the largest one, and we have

$$L_m(\vartheta) \xrightarrow{\text{a.s.}} \Lambda(\theta(\vartheta)).$$

It is now straightforward to check that the above two cases can be combined into a concise form stated in the claim of the lemma. □

We end this section by proving Lemma 9, restated below for convenience.

Lemma 4. *The following hold for the equation:*

$$\Lambda_+(\theta(\vartheta)) = 1/\vartheta + \mathbb{E}[|Z|^2 T],\ \vartheta > 0.$$

1. *This equation has a unique solution.*

2. *Let $\vartheta_\star$ denote the solution of the above equation. Then:*

Case 1 *If $\psi_1(\tau_r) \leq \frac{\delta}{\delta-1}$, we have*

$$\Lambda_+(\theta(\vartheta_\star)) = \Lambda(\tau_r).$$

Furthermore if $\psi_1(\tau_r) < \delta/(\delta-1)$, then,

$$\left.\frac{\mathrm{d}\Lambda_+(\theta(\vartheta))}{\mathrm{d}\vartheta}\right|_{\vartheta=\vartheta_\star} = 0,$$

Case 2 *If $\psi_1(\tau_r) > \frac{\delta}{\delta-1}$, we have*

$$\Lambda_+(\theta(\vartheta_\star)) = \Lambda(\theta_\star),$$

and,

$$\left.\frac{\mathrm{d}\Lambda_+(\theta(\vartheta))}{\mathrm{d}\vartheta}\right|_{\vartheta=\vartheta_\star} =$$

$$\frac{1}{\vartheta_\star^2}\cdot\frac{\delta}{\delta-1}\cdot\left(\frac{\delta}{\delta-1}-\psi_2(\theta_\star)\right)\cdot\frac{1}{\psi_3^2(\theta_\star)-\frac{\delta^2}{(\delta-1)^2}}.$$

where $\theta_\star > 1$ is the unique $\theta \geq \tau_r$ that satisfies $\psi_1(\theta) = \frac{\delta}{\delta-1}$.

Proof. Before we begin the proof of this lemma, it is helpful to list the conclusions of some of the previous lemmas.

Lemma 13: In this lemma, for $\vartheta > \vartheta_c$ we defined the function $\theta(\vartheta)$ as the unique value of $\lambda > \max(1, \mathbb{E}[|Z|^2 T] + 1/\vartheta)$ that satisfies

$$\lambda - \mathbb{E}[|Z|^2 T] - 1/\vartheta = \frac{1}{\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]}.$$

We also set $\theta(\vartheta) = 1$ when $\vartheta \leq \vartheta_c$. We also showed that $\theta(\vartheta)$ is strictly increasing on $[\vartheta_c, \infty)$ and $\theta(\infty) = \infty$. In particular $\theta(\vartheta)$ has a well defined inverse defined on the domain $[1, \infty)$ given by:

$$\theta^{-1}(\lambda) = \left(\lambda - \mathbb{E}[|Z|^2 T] - \frac{1}{\mathbb{E}\left[\frac{|Z|^2}{\lambda - T}\right]}\right)^{-1}. \tag{3.22}$$

Lemma 15: We defined the function $\Lambda(\tau)$ as

$$\Lambda(\tau) \triangleq \tau - \frac{(1 - 1/\delta)}{\mathbb{E}\left[\frac{1}{\tau - T}\right]}. \tag{3.23}$$

We showed the that $\Lambda(\tau)$ is strictly convex on $[1, \infty)$. We defined (τ_r, λ_r) to be the minimizing argument and the minimum value of $\Lambda(\tau)$ in $[1, \infty)$. In particular $\tau_r \geq 1$. We also showed that $\Lambda(\infty) = \infty$. We further defined $\Lambda_+(\tau)$ in the following way:

$$\Lambda_+(\tau) = \begin{cases} \lambda_r, & \tau \leq \tau_r. \\ \Lambda(\tau), & \tau > \tau_r. \end{cases}$$

Some simple implications of the above assertions are: First, since $\theta(\vartheta)$ and Λ_+ are both non-decreasing continuous functions $\Lambda_+(\theta(\vartheta))$ is non-decreasing and continuous. Second, since $\Lambda(\tau) = \lambda_r$ for $\tau \leq \tau_r$, we have, for all $\vartheta \leq \theta^{-1}(\tau_r)$, $\Lambda_+(\theta(\vartheta)) = \lambda_r$. Third since $\theta(\infty) = \infty$ and $\Lambda(\infty) = \infty$, we have, $\Lambda_+(\theta(\vartheta)) \to \infty$ as $\vartheta \to \infty$. The only possible point of non-differentiability of $\Lambda_+(\theta(\vartheta))$ is at $\vartheta = \theta^{-1}(\tau_r)$. It is straightforward to compute the derivative of $\Lambda(\theta(\vartheta))$ at all other points using implicit function theorem and obtain

$$\frac{\mathrm{d}\Lambda_+(\theta(\vartheta))}{\mathrm{d}\vartheta} = \begin{cases} 0 & \vartheta < \theta^{-1}(\tau_r), \\ \Lambda'(\theta(\vartheta)) \cdot \theta'(\vartheta) & \vartheta > \theta^{-1}(\tau_r). \end{cases} \tag{3.24}$$

The derivatives of Λ, θ can be calculated as,

$$\Lambda'(\tau) = \frac{\delta - 1}{\delta}\left(\frac{\delta}{\delta - 1} - \psi_2(\tau)\right). \tag{3.25}$$

$$\theta'(\vartheta) = \frac{1}{\vartheta^2}\left(\frac{\left(\mathbb{E}\left[\frac{|Z|^2}{\theta(\vartheta) - T}\right]\right)^2}{\mathbb{E}\left[\frac{|Z|^2}{(\theta(\vartheta) - T)^2}\right] - \left(\mathbb{E}\left[\frac{|Z|^2}{\theta(\vartheta) - T}\right]\right)^2}\right). \tag{3.26}$$

A representative plot of the function $\Lambda_+(\theta(\vartheta))$ is shown in Figure 3.3.

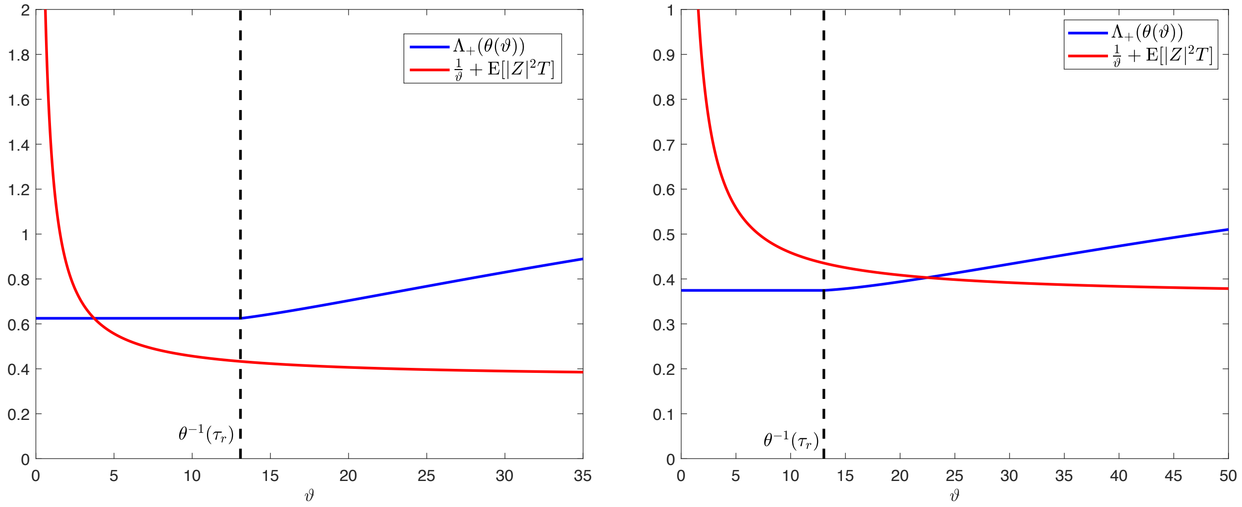


Figure 3.3: Typical Plots of the functions $\Lambda_+(\theta(\vartheta))$ (Blue) and $\mathbb{E}[|Z|^2 T] + \frac{1}{\vartheta}$ (Red). Case 1 (Left): The two functions intersect at the constant part of $\Lambda_+(\theta(\vartheta))$, Case 2 (Right): The The two functions intersect at the increasing part of $\Lambda_+(\theta(\vartheta))$

We are now in a position to prove the claims of the lemma.

1. Since $\Lambda_+(\theta(\vartheta))$ is continuous and non-decreasing and $1/\vartheta + \mathbb{E}[|Z|^2 T]$ is continuous and strictly decreasing, the fixed point equation can have at most one solution. On the other hand comparing the values of the two sides of the fixed point equation at $\vartheta \longrightarrow 0$ and $\vartheta \longrightarrow \infty$ shows that there is at least one solution.

2. Let $\vartheta_\star$ be denote the solution of the fixed point equation $\Lambda_+(\theta(\vartheta)) = 1/\vartheta + \mathbb{E}[|Z|^2 T]$. A typical plot of these two functions is shown in Figure 3.3. The figure shows two possible cases for the intersection of the two curves: *Case 1:* The curves intersect at a point $\vartheta_\star \leq \theta^{-1}(\tau_r)$ (or on the flat part of $\Lambda_+(\theta(\alpha))$. In this case we have, $\Lambda_+(\theta(\vartheta_\star)) = \lambda_r$.

   Case 2: The curves intersect at a point $\vartheta_\star > \theta^{-1}(\tau_r)$ or the rising part of $\Lambda_+(\theta(\alpha)$. We have $\Lambda_+(\theta(\vartheta_\star)) > \lambda_r$. We can distinguish between the two cases by comparing the value of the function $1/\vartheta + \mathbb{E}[|Z|^2 T]$ at $\vartheta = \theta^{-1}(\tau_r)$ with λ_r. In particular, we have,

   Case 1:

$$\Lambda_+(\theta(\vartheta_\star)) = \lambda_r \Leftrightarrow 1/\theta^{-1}(\tau_r) + \mathbb{E}[|Z|^2 T] \leq \lambda_r,$$

Case 2:

$$\Lambda_+(\theta(\vartheta_\star)) > \lambda_r \Leftrightarrow 1/\theta^{-1}(\tau_r) + \mathbb{E}[|Z|^2 T] > \lambda_r.$$

Substituting the formula for $\theta^{-1}(\tau_r)$, mentioned in (3.22), and $\lambda_r = \Lambda(\tau_r)$ and the formula for Λ from (3.23), the 2 cases can be simplified slightly more.

Case 1: This case occurs when

$$\frac{1}{\theta^{-1}(\tau_r)} + \mathbb{E}[|Z|^2 T] \leq \lambda_r \Leftrightarrow \frac{\mathbb{E}\left[\frac{|Z|^2}{\tau_r - T}\right]}{\mathbb{E}\left[\frac{1}{\tau_r - T}\right]} \leq \frac{\delta}{\delta - 1}.$$

In this situation, we have, $\Lambda_+(\theta(\vartheta_\star)) = \lambda_r$. Furthermore, if we additionally have

$$\frac{\mathbb{E}\left[\frac{|Z|^2}{\tau_r - T}\right]}{\mathbb{E}\left[\frac{1}{\tau_r - T}\right]} < \frac{\delta}{\delta - 1}$$

Then $\Lambda_+(\theta(\vartheta))$ is differentiable at $\vartheta_\star$ and, from (3.24), we have

$$\left.\frac{\mathrm{d}\Lambda_+(\theta(\vartheta))}{\mathrm{d}\vartheta}\right|_{\vartheta = \vartheta_\star} = 0.$$

Case 2: This case occurs when

$$\begin{aligned} &\frac{1}{\theta^{-1}(\tau_r)} + \mathbb{E}[|Z|^2 T] > \lambda_r \\ &\qquad\qquad \Leftrightarrow \frac{\mathbb{E}\left[\frac{|Z|^2}{\tau_r - T}\right]}{\mathbb{E}\left[\frac{1}{\tau_r - T}\right]} > \frac{\delta}{\delta - 1}. \end{aligned}$$

In this situation, we have, $\Lambda_+(\theta(\vartheta_\star)) > \lambda_r$. It turns out that we can give a simpler expression for $\Lambda_+(\theta(\vartheta_\star))$. In this case, $\vartheta_\star \geq \theta^{-1}(\tau_r)$ solves,

$$\Lambda(\theta(\vartheta_\star)) = \frac{1}{\vartheta_\star} + \mathbb{E}[|Z|^2 T], \tag{3.27}$$

and $\theta(\vartheta_\star) \geq 1$ is the solution of the equation

$$\mathbb{E}[|Z|^2 T] + \frac{1}{\vartheta_\star} = \theta(\vartheta_\star) - \frac{1}{\mathbb{E}\left[\frac{|Z|^2}{\theta(\vartheta_\star)-T}\right]}. \tag{3.28}$$

By definition the function $\Lambda(\tau(\alpha))$ is

$$\Lambda(\theta(\vartheta_\star)) = \theta(\vartheta_\star) - \frac{(1-1/\delta)}{\mathbb{E}\left[\frac{1}{\theta(\vartheta_\star)-T}\right]}. \tag{3.29}$$

We first eliminate $\vartheta_\star$ from Equations (3.27)-(3.29) and conclude that $\theta_\star \overset{\Delta}{=} \theta(\vartheta_\star)$ solves

$$\frac{\mathbb{E}\left[\frac{|Z|^2}{\theta_\star-T}\right]}{\mathbb{E}\left[\frac{1}{\theta_\star-T}\right]} = \frac{\delta}{\delta-1},\ \theta_\star \geq \tau_r, \tag{3.30}$$

and $\vartheta_\star$ is given by

$$\vartheta_\star = \left(\theta_\star - \frac{1}{\mathbb{E}\left[\frac{|Z|^2}{\theta_\star-T}\right]} - \mathbb{E}[|Z|^2 T]\right)^{-1}.$$

Since the solution to Equations (3.27)-(3.29) was guaranteed to be unique, the solution to (3.30) is guaranteed to be unique. Finally we can compute the derivative of $\Lambda_+(\theta(\vartheta))$ at $\vartheta = \vartheta_\star$. It will be convenient to introduce the random variable $G = (\theta_\star - T)^{-1}$ to write the equations in a compact form. From (3.24)-(3.26), we have

$$\begin{aligned}
&\left.\frac{\mathrm{d}\Lambda_+(\theta(\vartheta))}{\mathrm{d}\vartheta}\right|_{\vartheta=\vartheta_\star} = \Lambda'(\theta_\star)\cdot\theta'(\vartheta_\star)\\
&= \frac{\delta-1}{\delta\vartheta_\star^2}\left(\frac{\delta}{\delta-1}-\psi_2(\theta_\star)\right)\frac{\mathbb{E}[|Z|^2G]^2}{\mathbb{E}[|Z|^2G^2]-\mathbb{E}[|Z|^2G]^2}\\
&\overset{(a)}{=} \frac{\delta\cdot\left(\frac{\delta}{\delta-1}-\psi_2(\theta_\star)\right)}{\vartheta_\star^2\cdot(\delta-1)\cdot\psi_1^2(\theta_\star)}\cdot\frac{\mathbb{E}[|Z|^2G]^2}{\mathbb{E}[|Z|^2G^2]-\mathbb{E}[|Z|^2G]^2}\\
&= \frac{\delta\cdot\left(\frac{\delta}{\delta-1}-\psi_2(\theta_\star)\right)}{\vartheta_\star^2\cdot(\delta-1)}\cdot\frac{\mathbb{E}[G]^2}{\mathbb{E}[|Z|^2G^2]-\mathbb{E}[|Z|^2G]^2}\\
&= \frac{\delta}{\vartheta_\star^2(\delta-1)}\left(\frac{\delta}{\delta-1}-\psi_2(\theta_\star)\right)\frac{1}{\psi_3^2(\theta_\star)-\frac{\delta^2}{(\delta-1)^2}}.
\end{aligned}$$

In the above display, in the step marked (a) we used the fact that $\theta_\star$ satisfies $\psi_1(\theta_\star)=\delta/(\delta-1)$. This concludes the proof of the characterization (2) given in the statement of the lemma.

□

3.5 Conclusions

We analyzed the asymptotic performance of a spectral method for phase retrieval under a random column orthogonal matrix model. Our results provides a rigorous justification for the conjectures in [71], which were obtained by analyzing an expectation propagation algorithm.

3.6 Proof of Proposition 2

This section is devoted to the proof of Proposition 2. We denote the functions $\Lambda,\psi_1,\psi_2,\psi_3$ (recall (3.5)) with $\mathcal{T}=\mathcal{T}_{\mathsf{opt}}$ as $\Lambda_{\mathsf{opt}},\psi_1^{\mathsf{opt}},\psi_2^{\mathsf{opt}},\psi_3^{\mathsf{opt}}$ and those with $\mathcal{T}=\mathcal{T}_{\mathsf{opt},\epsilon}$ as $\Lambda_\epsilon,\psi_1^\epsilon,\psi_2^\epsilon,\psi_3^\epsilon$. Define the random variables:

$$Z\sim\mathcal{CN}(0,1)\,,\ T_{\mathsf{opt}}=\mathcal{T}_{\mathsf{opt}}(|Z|/\sqrt{\delta}),\ T_\epsilon=\mathcal{T}_{\mathsf{opt},\epsilon}(|Z|/\sqrt{\delta}).$$

Next we observe that the function $\mathcal{T}_{\mathsf{opt},\epsilon}$ is a bounded, strictly increasing, Lipchitz function and consequently T_ϵ has a density with respect to the Lebesgue measure. Hence by the rescale and shift argument outlined in Remark 9, Theorem 3 applies to a equivalent modification of $\mathcal{T}_{\mathsf{opt},\epsilon}$ which can used to infer the corresponding result for $\mathcal{T}_{\mathsf{opt},\epsilon}$ (after another rescale and shift argument). This gives us the result:

$$\frac{|\boldsymbol{x}_\star^* \hat{\boldsymbol{x}}_\epsilon|^2}{n} \xrightarrow{\text{a.s.}} \begin{cases} 0, & \psi_1^\epsilon(\tau_r^\epsilon) < \frac{\delta}{\delta-1}, \\ \frac{\left(\frac{\delta}{\delta-1}\right)^2 - \frac{\delta}{\delta-1}\cdot\psi_2^\epsilon(\theta_\star^\epsilon)}{\psi_3^\epsilon(\theta_\star^\epsilon)^2 - \frac{\delta}{\delta-1}\cdot\psi_2^\epsilon(\theta_\star^\epsilon)}, & \psi_1(\tau_r^\epsilon) > \frac{\delta}{\delta-1}. \end{cases}, \tag{3.31}$$

where $\tau_r^\epsilon \overset{\Delta}{=} \arg\min_{\tau\in[1,\infty)} \Lambda_\epsilon(\tau)$ and $\theta_\star^\epsilon$ is the solution to the fixed point equation (in τ): $\psi_1^\epsilon(\tau) = \delta/(\delta-1)$ which is guaranteed to exist uniquely provided $\psi_1(\tau_r^\epsilon) > \delta/(\delta-1)$. First we observe that,

$$\Lambda_\epsilon'(\tau) = 1 - \left(1 - \frac{1}{\delta}\right)\cdot\frac{\mathbb{E}G_\epsilon^2(\tau)}{(\mathbb{E}G_\epsilon(\tau))^2},\ G_\epsilon(\tau) = (\tau - T_\epsilon)^{-1}.$$

In particular, at $\tau = 1$, we have,

$$\begin{aligned} \Lambda_\epsilon'(1) &= 1 - \left(1 - \frac{1}{\delta}\right)\cdot\frac{(1+\epsilon)^2 + 1}{(1+\epsilon)^2} \\ \implies \lim_{\epsilon\downarrow 0}\Lambda'(1) &= \frac{2-\delta}{\delta}, \end{aligned}$$

and,

$$\psi_1^\epsilon(1) = 2 + \epsilon.$$

We consider the following two cases.

Case 1: $1 < \delta < 2$. Lemma 15 shows that $\Lambda_\epsilon(\tau)$ is convex on $[1,\infty)$. When $\delta < 2$, $\Lambda_\epsilon'(1) > 0$ for ϵ small enough, and hence Λ_ϵ is strictly increasing and $\tau_r^\epsilon = 1$. Moreover, in this case, for ϵ

small enough,

$$\frac{\delta}{\delta-1} = 2 + \frac{2-\delta}{\delta-1} > 2 + \epsilon = \psi_1^{\epsilon}(1).$$

Hence, using (3.31),

$$\lim_{\epsilon \downarrow 0} \lim_{\substack{m,n \to \infty \\ m = \delta n}} \frac{|\boldsymbol{x}^* \hat{\boldsymbol{x}}_\epsilon|^2}{n} = 0.$$

Case 2: $\delta > 2$ In this case, for small enough ϵ, $\Lambda_\epsilon'(1) < 0$. Hence the τ_r^ϵ, the minimizer of the convex function Λ_ϵ occurs in the region $(1, \infty)$. This means it satisfies the optimality condition:

$$\Lambda_\epsilon'(\tau_r^\epsilon) = 0 \Leftrightarrow \psi_2(\tau_r^\epsilon) = \frac{\delta}{\delta-1}.$$

Next we claim that, $\forall \tau \in [1, \infty)$,

$$\psi_1^\epsilon(\tau) > \psi_2^\epsilon(\tau) \Leftrightarrow \mathbb{E}[G_\epsilon(\tau)] \cdot \mathbb{E}[|Z|^2 G_\epsilon(\tau)] > \mathbb{E}[G_\epsilon^2(\tau)],$$

which is a consequence of Chebychev's association inequality (Fact 1) with the choice:

$$\begin{aligned} &B = G_\epsilon(\tau),\ A = |Z|, \\ &f(a) = a^2\left(\tau - \mathcal{T}_\epsilon\left(\frac{a}{\sqrt{\delta}}\right)\right),\ g(a) = \left(\tau - \mathcal{T}_\epsilon\left(\frac{a}{\sqrt{\delta}}\right)\right)^{-1}. \end{aligned}$$

In particular we have $\psi_1^\epsilon(\tau_r^\epsilon) > \delta/(\delta-1)$, and hence Theorem 3 gives us:

1. There exists a unique solution $\theta_\star^\epsilon \in (\tau_r^\epsilon, \infty)$ such that $\psi_1^\epsilon(\theta_\star^\epsilon) = \delta/(\delta-1)$,

2. and,

$$\frac{|\boldsymbol{x}^* \hat{\boldsymbol{x}}_\epsilon|^2}{n} \xrightarrow{\text{a.s.}} \frac{\left(\frac{\delta}{\delta-1}\right)^2 - \frac{\delta}{\delta-1} \cdot \psi_2^\epsilon(\theta_\star^\epsilon)}{\psi_3^\epsilon(\theta_\star^\epsilon)^2 - \frac{\delta}{\delta-1} \cdot \psi_2^\epsilon(\theta_\star^\epsilon)}.$$

Next we claim that,

$$1 < \liminf_{\epsilon \downarrow 0} \theta_\star^\epsilon \leq \limsup_{\epsilon \downarrow 0} \theta_\star^\epsilon < \infty.$$

To see this, observe

$$\psi_1^\epsilon(\theta_\star^\epsilon) = \frac{\mathbb{E}\frac{|Z|^2(|Z|^2+\epsilon)}{(\theta_\star^\epsilon-1)(|Z|^2+\epsilon)+1}}{\mathbb{E}\frac{(|Z|^2+\epsilon)}{(\theta_\star^\epsilon-1)(|Z|^2+\epsilon)+1}}.$$

If $\liminf_{\epsilon\downarrow 0} \theta_\star^\epsilon = 1$, one can select a subsequence along which $\psi_1^\epsilon(\theta_\star^\epsilon) \to \mathbb{E}|Z|^4 = 2$ by dominated convergence which contradicts: $\psi_2^\epsilon(\theta_\star^\epsilon) = \delta/(\delta-1) < 2$. Likewise if $\limsup_{\epsilon\downarrow 0} \theta_\star^\epsilon = \infty$, one can find a subsequence along which $\theta_\star^\epsilon \to \infty$ and, by dominated convergence,

$$\psi_1^\epsilon(\theta_\star^\epsilon) = \frac{\mathbb{E}\frac{|Z|^2(|Z|^2+\epsilon)(\theta_\star^\epsilon-1)}{(\theta_\star^\epsilon-1)(|Z|^2+\epsilon)+1}}{\mathbb{E}\frac{(|Z|^2+\epsilon)(\theta_\star^\epsilon-1)}{(\theta_\star^\epsilon-1)(|Z|^2+\epsilon)+1}} \to 1,$$

which contradicts $\psi_1^\epsilon(\theta_\star^\epsilon) = \delta/(\delta-1) < 1 \; \forall \; \delta \; \in \; (2,\infty)$. We can now conclude that,

$$\liminf_{\epsilon\downarrow 0} \theta_\star^\epsilon = \limsup_{\epsilon\downarrow 0} \theta_\star^\epsilon = \theta_\star^{\mathsf{opt}},$$

where $\theta_\star^{\mathsf{opt}}$ is the unique solution to $\psi_1^{\mathsf{opt}}(\tau) = \delta/(\delta-1)$ in $\tau \in (1,\infty)$ guaranteed by Proposition 1 (due to [71]). This is because, by selecting a subsequence along with $\theta_\star^\epsilon \to \liminf_{\epsilon\downarrow 0}\theta_\star^\epsilon$, we can conclude that, along that subsequence,

$$\frac{\delta}{\delta-1} = \psi_1^\epsilon(\theta_\star^\epsilon) \to \psi_1^{\mathsf{opt}}\left(\liminf_{\epsilon\downarrow 0} \theta_\star^\epsilon\right).$$

This implies,

$$\psi_1^{\mathsf{opt}}\left(\liminf_{\epsilon\downarrow 0} \theta_\star^\epsilon\right) = \frac{\delta}{\delta-1},$$

and analogously,

$$\psi_1^{\mathsf{opt}}\left(\limsup_{\epsilon\downarrow 0}\theta_\star^\epsilon\right) = \frac{\delta}{\delta-1}.$$

Since Proposition 1 guarantees that the equation $\psi_1^{\mathsf{opt}}(\tau) = \delta/(\delta-1)$ has a unique solution in $(1,\infty)$ we get,

$$\liminf_{\epsilon\downarrow 0}\theta_\star^\epsilon = \limsup_{\epsilon\downarrow 0}\theta_\star^\epsilon = \theta_\star^{\mathsf{opt}}.$$

Dominated convergence now yields,

$$\psi_i^\epsilon(\theta_\star^\epsilon) \to \psi_i^{\mathsf{opt}}(\theta_\star^{\mathsf{opt}}),\ \text{as } \epsilon\downarrow 0\ \forall\ i = 1,2,3,$$

and consequently, almost surely,

$$\lim_{\epsilon\downarrow 0}\lim_{\substack{m,n\to\infty,\\ m=n\delta}}\frac{|\boldsymbol{x}^*\hat{\boldsymbol{x}}_\epsilon|^2}{n} \overset{\text{a.s.}}{=} \frac{\left(\frac{\delta}{\delta-1}\right)^2 - \frac{\delta}{\delta-1}\cdot\psi_2^{\mathsf{opt}}(\theta_\star^{\mathsf{opt}})}{\psi_3^{\mathsf{opt}}(\theta_\star^{\mathsf{opt}})^2 - \frac{\delta}{\delta-1}\cdot\psi_2^{\mathsf{opt}}(\theta_\star^{\mathsf{opt}})}.$$

The right hand side of the above display can be simplified to:

$$\frac{\left(\frac{\delta}{\delta-1}\right)^2 - \frac{\delta}{\delta-1}\cdot\psi_2^{\mathsf{opt}}(\theta_\star^{\mathsf{opt}})}{\psi_3^{\mathsf{opt}}(\theta_\star^{\mathsf{opt}})^2 - \frac{\delta}{\delta-1}\cdot\psi_2^{\mathsf{opt}}(\theta_\star^{\mathsf{opt}})} = \frac{\theta_\star^{\mathsf{opt}} - 1}{\theta_\star^{\mathsf{opt}} - \frac{1}{\delta}}.$$

This clean formula is due to [71] and we refer the reader to Appendix B in [71] for a proof.

3.7 Miscellaneous results

Fact 1 (Chebychev Association Inequality, [87])**.** *Let A, B be r.v.s and $B \geq 0$. Suppose f, g are two non-decreasing functions. Then,*

$$\mathbb{E}[B]\mathbb{E}[Bf(A)g(A)] \geq \mathbb{E}[f(A)B]\mathbb{E}[g(A)B].$$

Furthermore, if, $\mathbb{P}(B = 0) = 0$ *and,*

$$\mathbb{P}\left(f(A) = x\right) = 0,\ \mathbb{P}\left(g(A) = x\right) = 0,\ \forall\, x\ \in\ \mathbb{R},$$

then, the above inequality is strict.

Proof. The proof of the inequality appears in [87]. Inspecting the proof we can derive a sufficient condition for the inequality to be strict. The proof in [87] shows,

$$\begin{aligned} 2 \cdot (\mathbb{E}[B]\mathbb{E}[Bf(A)g(A)] - \mathbb{E}[f(A)B]\mathbb{E}[g(A)B]) = \\ \mathbb{E}BB'(f(A) - f(A')) \cdot (g(A) - g(A')). \end{aligned}$$

where (B', A') is an independent sample of the random variables (B, A). Since, f, g are increasing $(f(A) - f(A')) \cdot (g(A) - g(A')) \geq 0$ and $B \geq 0, B' \geq 0$. Hence the equality is tight iff:

$$BB'(f(A) - f(A')) \cdot (g(A) - g(A')) \overset{\text{a.s.}}{=} 0,$$

which is ruled out by the assumptions of the claim. □

Chapter 4: Universality of Linearized Message Passing for Phase Retrieval with Structured Sensing Matrices

4.1 Introduction

In the phase retrieval one observes magnitudes of m linear measurements (denoted by $y_1, y_2, ..., y_m$) of an unknown n dimensional signal vector $\boldsymbol{x}$:

$$y_i = |(\boldsymbol{A}\boldsymbol{x})_i|,$$

where $\boldsymbol{A}$ is a $m \times n$ sensing matrix. The phase retrieval problem is a mathematical model of imaging systems which are unable to measure the phase of the measurements. Such imaging systems arise in a variety of applications such as electron microscopy, crystallography, astronomy and optical imaging [7].

Theoretical analyses of the phase retrieval problem seek to design algorithms to recover $\boldsymbol{x}$ (up to a global phase) with the minimum number of measurements. The earliest theoretical analysis modelled the sensing as a random matrix with i.i.d. Gaussian entries and design computationally efficient estimators which recover $\boldsymbol{x}$ with information theoretically rate-optimal $O(n)$ (or nearly optimal $m = O(n\,\text{polylog}(n))$) measurements. A representative, but necessarily incomplete, list of such works includes the analysis of convex relaxations like PhaseLift due to [16, 17], PhaseMax due to [21, 22], and analysis of non-convex optimization based methods due to [28], [25], and [88]. The number of measurements required if the underlying signal has a low dimensional structure has also been investigated [34, 41, 43].

Unfortunately, i.i.d. Gaussian is not realizable in practice; instead, the sensing matrix is usually a variant of the Discrete Fourier Transform (DFT) matrix [89]. Hence, there have been efforts to extend the theory to structured sensing matrices [90, 91, 92, 72, 93, 94]. A popular structured sensing ensemble is the Coded Diffraction Pattern (CDP) ensemble introduced by [92] which is intended to model applications where it is possible to randomize the image acquisition by introducing random masks in front of the object. In this setup, the sensing matrix is given by:

$$\boldsymbol{A}_{\mathsf{CDP}} = \begin{bmatrix} \boldsymbol{F}_n \boldsymbol{D}_1 \\ \boldsymbol{F}_n \boldsymbol{D}_2 \\ \vdots \\ \boldsymbol{F}_n \boldsymbol{D}_L \end{bmatrix},$$

where $\boldsymbol{F}_n$ denotes the $n \times n$ DFT matrix and $\boldsymbol{D}_{1:L}$ are random diagonal matrices representing masks:

$$\boldsymbol{D}_\ell = \mathrm{Diag}\left(\mathrm{e}^{\mathrm{i}\theta_{1,\ell}}, \mathrm{e}^{\mathrm{i}\theta_{2,\ell}}, \cdots, \mathrm{e}^{\mathrm{i}\theta_{n,\ell}}\right),$$

and $\mathrm{e}^{\mathrm{i}\theta_{j,\ell}}$ are random phases. For the CDP ensemble convex relaxation methods like PhaseLift [72] and non-convex optimization based methods [25] are known to recover the signal $\boldsymbol{x}$ with the near optimal $m = O(n \,\mathrm{polylog}(n))$ measurements. Another common structured sensing model is the sub-sampled Fourier sensing model where the sensing matrix is generated as:

$$\boldsymbol{A}_{\mathsf{DFT}} = \boldsymbol{F}_m \boldsymbol{P} \boldsymbol{S},$$

where $\boldsymbol{F}$ is the $m \times m$ Fourier matrix, $\boldsymbol{P}$ is a uniformly random $m \times m$ permutation matrix and $\boldsymbol{S}$ the matrix that selects the first n columns of an $m \times m$ matrix:

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{I}_n \\ \boldsymbol{0}_{m-n,n} \end{bmatrix}. \tag{4.1}$$

This models a common oversampling strategy to ensure injectivity [8]. We also refer the reader to the recent review articles [95, 89, 96, 8] for more discussion regarding good models of practical sensing matrices.

The aforementioned finite sample analyses show that a variety of different methods succeed in solving the phase retrieval problem with the optimal or nearly optimal order of magnitude of measurements. However, in practice, these methods can have a vast difference in performance, which is not captured by the non-asymptotic analyses. Consequently, efforts have been made to complement these results with sharp high dimensional asymptotic analyses which shed light on the performance of different estimators and information theoretic lower bounds in the high dimensional limit $m, n \to \infty, \ n/m \to \kappa$. This provides a high resolution framework to compare different estimators based on the critical value of κ at which they achieve non-trivial performance (i.e. better than a random guess) or exact recovery of $\boldsymbol{x}$. Comparing this to the critical value of κ required information theoretically allows us to reason about the optimality of known estimators. This research program has been executed, to varying extents, for the following unstructured sensing ensembles:

1. Gaussian Ensemble: In this ensemble the entries of the sensing matrix are assumed to be i.i.d. Gaussian (real or complex). This is the most well studied ensemble in the high dimensional asymptotic limit. For this ensemble, precise performance curves for spectral methods [29, 69, 30], convex relaxation methods like PhaseLift [97] and PhaseMax [98], and a class of iterative algorithms called Approximate Message Passing [99] are now well understood. The precise asymptotic limit of the Bayes risk [100] for Bayesian phase retrieval is also known.

2. Sub-sampled Haar Ensemble: In the sub-sampled Haar sensing model, the sensing matrix is generated by picking n columns of a uniformly random orthogonal (or unitary) matrix at random:

$$\boldsymbol{A}_{\mathsf{Haar}} = \boldsymbol{OPS},$$

where $\boldsymbol{O} \sim \text{Unif}\left(\mathbb{U}_m\right)$ (or $\boldsymbol{O} \sim \text{Unif}\left(\mathbb{O}_m\right)$ in the real case) and $\boldsymbol{P}$ is a uniformly random $m \times m$ permutation matrix and $\boldsymbol{S}$ is the matrix defined in (4.1). The sub-sampled Haar model captures a crucial aspect of sensing matrices that arise in practice: namely they have orthogonal columns (note that for both the CDP and the sub-sampled Fourier ensembles we have $\boldsymbol{A}_{\mathsf{DFT}}^* \boldsymbol{A}_{\mathsf{DFT}} = \boldsymbol{A}_{\mathsf{CDP}}^* \boldsymbol{A}_{\mathsf{CDP}} = \boldsymbol{I}_n$). For the sub-sampled Haar sensing model it has been shown that when $\kappa > 0.5$ no estimator performs better than a random guess [31]. Moreover, it is known that spectral estimators can achieve non-trivial performance when $\kappa < 0.5$ [71, 32].

3. Rotationally Invariant Ensemble: This is a broad class of unstructured sensing ensembles that include the Gaussian Ensemble and the sub-sampled Haar ensemble as special cases. Here, it is assumed that the SVD of the sensing matrix is given by:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}},$$

where $\boldsymbol{U}, \boldsymbol{V}$ are independent and uniformly random orthogonal matrices (or unitary in the complex case): $\boldsymbol{U} \sim \mathrm{Unif}\left(\mathbb{O}_m\right), \ \boldsymbol{V} \sim \mathrm{Unif}\left(\mathbb{O}_n\right)$ and $\boldsymbol{S}$ is a deterministic matrix such that the empirical spectral distribution of $\boldsymbol{S}^{\mathsf{T}}\boldsymbol{S}$ converges to a limiting measure μ_S. The analysis of Approximate Message Passing algorithms has been extended to this ensemble [101, 102]. For this ensemble, the non-rigorous replica method from statistical physics can be used to derive conjectures regarding the Bayes risk and performance of convex relaxations as well as spectral methods [103, 104, 105]. Some of these conjectures have been proven rigorously in some special cases [106, 107].

The techniques used to prove the above results rely heavily on the rotational invariance of the underlying matrix ensembles. This makes it difficult to extend these results to structured sensing matrices.

However, numerical simulations reveal an intriguing universality phenomena: It has been observed that the performance curves derived theoretically for sub-sampled Haar sensing provide a nearly perfect fit to the empirical performance on practical sensing ensembles like $\boldsymbol{A}_{\mathsf{CDP}}, \boldsymbol{A}_{\mathsf{DFT}}$. This has been observed by a number of authors in the context of various signal processing problems. It was first pointed out by [108] in the context of ℓ_1 norm minimization for noiseless compressed sensing and then again by [109] for the same setup but for many more structured sensing ensembles. For noiseless compressed sensing both the Gaussian ensemble and the Sub-sampled Haar ensemble lead to identical predictions (and hence the simulations with structured sensing matrices match both of them). However, in noisy compressed sensing, the predictions from the sub-sampled Haar model and the Gaussian model are different. [110] pointed out that structured ensembles generated by sub-sampling deterministic orthogonal matrices empirically behave like Sub-sampled Haar sensing matrices. More recently, [111] have observed this universality phenomena in the context of approximate message passing algorithms for noiseless compressed sensing. In the context of phase retrieval this phenomena was reported by [71] for the performance of the spectral method.

Our Contribution: In this chapter we study the real phase retrieval problem where the sensing matrix is generated by sub-sampling n columns of the $m \times m$ Hadamard-Walsh matrix. Under an average case assumption on the signal vector, our main result (Theorem 7) shows that the dynamics of a class of linearized Approximate message passing schemes for this structured ensemble are asymptotically identical to the dynamics of the same algorithm in the sub-sampled Haar sensing model in the high dimensional limit where m, n diverge to infinity such that ratio $\kappa = n/m \in (0, 1)$ is held fixed. This provides a theoretical justification for the observed empirical universality in this particular setup. In the following section we define the setup we study in more detail.

4.1.1 Setup

Sensing Model

As mentioned in the Introduction, we study the phase retrieval problem where the measurements $y_1, y_2, \dots y_m$ are given by:

$$y_i = (|\boldsymbol{A}\boldsymbol{x}|)_i.$$

The matrix $\boldsymbol{A}$ is called the sensing matrix. We also define $\boldsymbol{z} \triangleq \boldsymbol{A}\boldsymbol{x}$ which we refer to as the signed measurements (which are not observed). We need to introduce the following 3 models for the sensing matrix $\boldsymbol{A}$:

In all the equations below, $\boldsymbol{P}$ is a uniformly random $m \times m$ permutation matrix and $\boldsymbol{S}$ is the selection matrix as defined in (4.1).

Sub-sampled Hadamard Sensing Model: Assume that $m = 2^\ell$ for some $\ell \in \mathbb{N}$. In the sub-sampled Hadamard sensing model the sensing matrix is generated by sub-sampling n columns of a $m \times m$ Hadamard-Walsh matrix $\boldsymbol{H}$ uniformly at random:

$$\boldsymbol{A} = \boldsymbol{HPS}, \tag{4.2}$$

Recall that the Hadamard-Walsh matrix as a closed form formula: For any $i, j \in [m]$, let $\boldsymbol{i}, \boldsymbol{j}$ denote the binary representations of $i-1, j-1$. Hence, $\boldsymbol{i}, \boldsymbol{j} \in \{0,1\}^\ell$. Then the (i,j)-th entry of $\boldsymbol{H}$ is given by:

$$H_{ij} = \frac{(-1)^{\langle \boldsymbol{i}, \boldsymbol{j} \rangle}}{\sqrt{m}}, \tag{4.3}$$

where $\langle \boldsymbol{i}, \boldsymbol{j} \rangle = \sum_{k=1}^{\ell} i_k j_k$. It is well known that $\boldsymbol{H}$ is orthogonal, i.e. $\boldsymbol{H}^\mathsf{T}\boldsymbol{H} = \boldsymbol{I}_m$. This sensing model can be thought of as a real analogue of the sub-sampled Fourier sensing model. Our primary goal is to develop a theory for this sensing model which is not covered by existing results. We believe that our analysis can be extended to the Fourier case without much effort as well as some other deterministic orthogonal matrices like the discrete cosine transform matrix.

Remark 19. *Some authors refer to any orthogonal matrix with* ±1 *entries as a Hadamard matrix. We emphasize that we claim results only about the Hadamard-Walsh construction given in* (4.3) *and not arbitrary Hadamard matrices.*

Sub-sampled Haar Sensing Model: In this model the sensing matrix is generated by sub-sampling n columns, chosen uniformly at random, of a $m \times m$ uniformly random orthogonal matrix:

$$\boldsymbol{A} = \boldsymbol{OPS}, \tag{4.4}$$

where $\boldsymbol{O} \sim \mathrm{Unif}\left(\mathbb{O}_m\right)$. Existing theory applies to this sensing model and our goal will be to transfer these results to the sub-sampled Hadamard model.

Sub-sampled Orthogonal Model: This model includes both sub-sampled Hadamard and Haar models as special cases. In this model the sensing matrix is generated by sub-sampling n columns chosen uniformly at random of a $m \times m$ orthogonal matrix $\boldsymbol{U}$:

$$\boldsymbol{A} = \boldsymbol{UPS}, \tag{4.5}$$

where $\boldsymbol{U}$ is a fixed or random orthogonal matrix. Setting $\boldsymbol{U} = \boldsymbol{O}$ gives the sub-sampled Haar model and setting $\boldsymbol{U} = \boldsymbol{H}$ gives the sub-sampled Hadamard model. Our primary purpose for introducing this general model is that it allows us to handle both the sub-sampled Haar and Hadamard models in a unified way. Additionally, some of our intermediate results hold for any orthogonal matrix $\boldsymbol{U}$ whose entries are delocalized, and we wish to record that when possible.

In addition, we introduce the following matrices which will play an important role in our analysis:

1. We define $\boldsymbol{B} \triangleq \boldsymbol{PSS}^{\mathsf{T}}\boldsymbol{P}^{\mathsf{T}}$. Observe that $\boldsymbol{B}$ is a random diagonal matrix with $\{0, 1\}$ entries. It is easy to check that the distribution of $\boldsymbol{B}$ is described as follows: pick a uniformly random subset $S \subset [m]$ with $|S| = n$ and set:

$$B_{ii} = \begin{cases} 1 : & i \in S \\ 0 : & i \notin S \end{cases}.$$

2. Note that $\mathbb{E}\boldsymbol{B} = \kappa \boldsymbol{I}_m$. We define the zero mean random diagonal matrix $\overline{\boldsymbol{B}} \triangleq \boldsymbol{B} - \kappa \boldsymbol{I}_m$.

3. We define the matrix $\boldsymbol{\Psi} \triangleq \boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^\mathsf{T} = \boldsymbol{A}\boldsymbol{A}^\mathsf{T} - \kappa \boldsymbol{I}_m$.

Finally, note that all the sensing ensembles introduced in this section make sense only when $n \leq m$ or equivalently $\kappa \in [0, 1]$. We will additionally assume that κ lies in the open interval $(0, 1)$.

Algorithm

We study a class of linearized message passing algorithms. This is a class of iterative schemes which execute the following updates:

$$\hat{\boldsymbol{z}}^{(t+1)} \triangleq \left(\frac{1}{\kappa}\boldsymbol{A}\boldsymbol{A}^\mathsf{T} - \boldsymbol{I}\right) \cdot \left(\eta_t(\boldsymbol{Y}) - \frac{\mathbb{E}\mathsf{Tr}(\eta_t(\boldsymbol{Y}))}{m}\boldsymbol{I}\right) \cdot \hat{\boldsymbol{z}}^{(t)}, \tag{4.6a}$$

$$\hat{\boldsymbol{x}}^{(t+1)} \triangleq \boldsymbol{A}^\mathsf{T}\hat{\boldsymbol{z}}^{(t+1)}, \tag{4.6b}$$

where

$$\boldsymbol{Y} = \mathrm{Diag}\left(y_1, y_2 \ldots y_m\right),$$

and $\eta_t : \mathbb{R} \to \mathbb{R}$ are bounded Lipchitz functions that act entry-wise on the diagonal matrix $\boldsymbol{Y}$. The iterates $(\hat{\boldsymbol{z}}^{(t)})_{t\geq 0}$ should be thought as estimates of the signed measurements $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}$. We now provide further context regarding the iteration in (4.6).

Interpretation as Linearized AMP: The iteration (4.6) can be thought of as a linearization of a broad class of non-linear approximate message passing algorithms. These algorithms execute the iteration:

$$\hat{\boldsymbol{z}}^{(t+1)} \triangleq \left(\frac{1}{\kappa}\boldsymbol{A}\boldsymbol{A}^\mathsf{T} - \boldsymbol{I}\right) \cdot H_t(\boldsymbol{y}, \hat{\boldsymbol{z}}^{(t)}), \tag{4.7a}$$

$$\hat{\boldsymbol{x}}^{(t+1)} \triangleq \boldsymbol{A}^\mathsf{T}\hat{\boldsymbol{z}}^{(t+1)}. \tag{4.7b}$$

where $H_t : \mathbb{R}^2 \to \mathbb{R}$ is a bounded Lipchitz function which satisfies the divergence-free property:

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\partial_z H_t(y_i, \hat{z}_i^{(t)}) = 0.$$

Indeed, if H_t was linear in the second (z) argument (or was approximated by its linearization) one obtains the iteration in (4.6). By choosing the function H_t in the iteration appropriately, one can obtain the state-of-the-art performance for phase retrieval with sub-sampled Haar sensing. This algorithm achieves non-trivial (better than random) performance when $\kappa < 2/3$, and exact recovery when $\kappa < 0.63$ [107]. While our analysis currently does not cover the non-linear iteration (4.7), we hope our techniques can be extended to analyze (4.7).

Connection to Spectral Methods: Given that the algorithm we analyze (4.6) does not cover the state-of-the-art algorithm, one can reasonably ask what performance can one achieve with the linearized iteration (4.6). It turns out that the iteration in (4.6) can implement a popular class of spectral methods which estimates the signal vector $\boldsymbol{x}$ as proportional to the leading eigenvector of the matrix:

$$\boldsymbol{M} = \frac{1}{m}\sum_{i=1}^{m} \mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^{\mathsf{T}},$$

where $\boldsymbol{a}_1^T, \boldsymbol{a}_2^T, ..., \boldsymbol{a}_m^T$ denote the rows of $\boldsymbol{A}$ and $\mathcal{T} : \mathbb{R}^{\geq 0} \to (-\infty, 1)$ is a trimming function. The performance of these spectral estimators have been analyzed in the high dimensional limit [71, 32] for the sub-sampled Haar model and they are known to have a non-trivial (better than random) performance when $\kappa < 2/3$. Furthermore, simulations show that the same result holds for sub-sampled Hadamard sensing. In order to connect the iteration (4.6) to the spectral estimator, [71] proposed setting the functions η_t in the following way:

$$\eta_t(y) = \left(\frac{1}{\mu} - \mathcal{T}(y)\right)^{-1}, \tag{4.8}$$

where $\mu \in (0, 1)$ is a tuning parameter. [71] shows that with this choice of η_t, every fixed point of the iteration (4.6) denoted by z^{∞}, $\boldsymbol{A}^{\mathsf{T}}z^{\infty}$ is an eigenvector of the matrix $\boldsymbol{M}$. Furthermore, suppose μ is set to be the solution to the equation:

$$\psi_1(\mu) = \frac{1}{1-\kappa},\ \psi_1(\mu) \triangleq \frac{\mathbb{E}|Z|^2 G}{\mathbb{E}G}, \tag{4.9}$$

where the joint distribution of (Z, G) is given by:

$$Z \sim \mathcal{N}(0,1),\ G = \left(\frac{1}{\mu} - \mathcal{T}(|Z|)\right)^{-1}.$$

Then, [71] have shown that the linearized message passing iterations (4.6) achieve the same performance as the spectral method for the sub-sampled Haar model as $t \longrightarrow \infty$.

The State Evolution Formalism: An important property of the AMP algorithms of (4.6) and (4.7) is that for the sub-sampled Haar model, the dynamics of the algorithm can be tracked by a deterministic scalar recursion known as the state evolution. This was first shown for Gaussian sensing matrices by [99] and subsequently for rotationally invariant ensembles by [102]. We instantiate their result for our problem in the following proposition.

Proposition 6 (State Evolution [102])**.** *Suppose that the sensing matrix is generated from the sub-sampled Haar model and the signal vector is normalized such that* $\|\boldsymbol{x}\|_2^2/m \xrightarrow{P} 1$ *and the iteration* (4.6) *is initialized as:*

$$\hat{\boldsymbol{z}}^{(0)} = \alpha_0 \boldsymbol{z} + \sigma_0 \boldsymbol{w},$$

where $\alpha_0 \in \mathbb{R}, \sigma_0 \in \mathbb{R}^{\geq 0}$ *are fixed and* $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$*. Then for any fixed* $t \in \mathbb{N}$*, as* $m, n \longrightarrow \infty$*,* $n/m \longrightarrow \kappa$*, we have,*

$$\frac{\langle \hat{\boldsymbol{z}}^{(t)}, \boldsymbol{z}\rangle}{m} \xrightarrow{P} \alpha_t,\ \frac{\|\hat{\boldsymbol{z}}^{(t)}\|_2^2}{m} \xrightarrow{P} \alpha_t^2 + \sigma_t^2,$$
$$\frac{\langle \hat{\boldsymbol{x}}^{(t)}, \boldsymbol{x}\rangle}{m} \xrightarrow{P} \alpha_t,\ \frac{\|\hat{\boldsymbol{x}}^{(t)}\|_2^2}{m} \xrightarrow{P} \alpha_t^2 + (1-\kappa)\sigma_t^2,$$

where (α_t, σ_t^2) are given by the recursion:

$$\alpha_{t+1} = (\delta - 1) \cdot \alpha_t \cdot \mathbb{E} Z^2 \overline{\eta}_t(|Z|), \tag{4.10a}$$

$$\sigma_{t+1}^2 = \left(\frac{1}{\kappa} - 1\right) \cdot \left(\alpha_t^2 \cdot \left\{\mathbb{E} Z^2 \overline{\eta}_t^2(|Z|) - (\mathbb{E} Z^2 \overline{\eta}_t(|Z|))^2\right\} + \sigma_t^2 \mathbb{E} \overline{\eta}_t^2(|Z|)\right). \tag{4.10b}$$

In the above display, $Z \sim \mathcal{N}(0,1)$ and $\overline{\eta}_t(z) = \eta_t(z) - \mathbb{E}\eta_t(|Z|)$.

The above proposition lets us track the evolution of some performance metrics like the mean squared error (MSE) and the cosine similarity of the iterates. The proof of Proposition 6 crucially relies on the rotational invariance of the sub-sampled Haar ensemble via Bolthausen's conditioning technique [112] and does not extend to structured sensing ensembles.

A Demonstration of the Universality Phenomena: For the sake of completeness, we provide a self contained demonstration of the universality phenomena that we seek to study in Figure 4.1. In order to generate this figure:

1. We used a 1024 × 256 image (after vectorization, shown as inset in Figure 4.1) as the signal vector. Each of the red, blue, green channels were centered so that that their mean was zero and standard deviation was 1.

2. We set $m = 1024 \times 256$.

3. In order to generate problems with different κ we down-sampled the original image to obtain a new signal with $n \approx m\kappa$ (upto rounding errors).

4. We used a randomly sub-sampled Hadamard matrix for sensing. This was used to construct a phase retrieval problem for each of the red, blue and green channels.

5. We used the linearized message passing configured to implement the spectral estimator (c.f. (4.8) and (4.9)) with the optimal trimming function [30, 71]:

$$\mathcal{T}_\star(y) = 1 - \frac{1}{y^2}.$$

We ran the algorithm for 20 iterations and tracked the squared cosine similarity:

$$\cos^2(\angle(\hat{\boldsymbol{x}}^{(t)},\boldsymbol{x})) \triangleq \frac{|\langle \hat{\boldsymbol{x}}^{(t)},\boldsymbol{x}\rangle|^2}{\|\hat{\boldsymbol{x}}^{(t)}\|_2^2\|\boldsymbol{x}\|_2^2}.$$

We averaged the squared cosine similarity across the RGB channels.

6. We repeated this for 10 different random sensing matrices. The average cosine similarity is represented by + markers in Figure 4.1 and the error bars represent the standard error across 10 repetitions. The solid curves represent the predictions derived from State Evolution (see Proposition 6). We can observe that the State Evolution closely tracks the empirical dynamics.

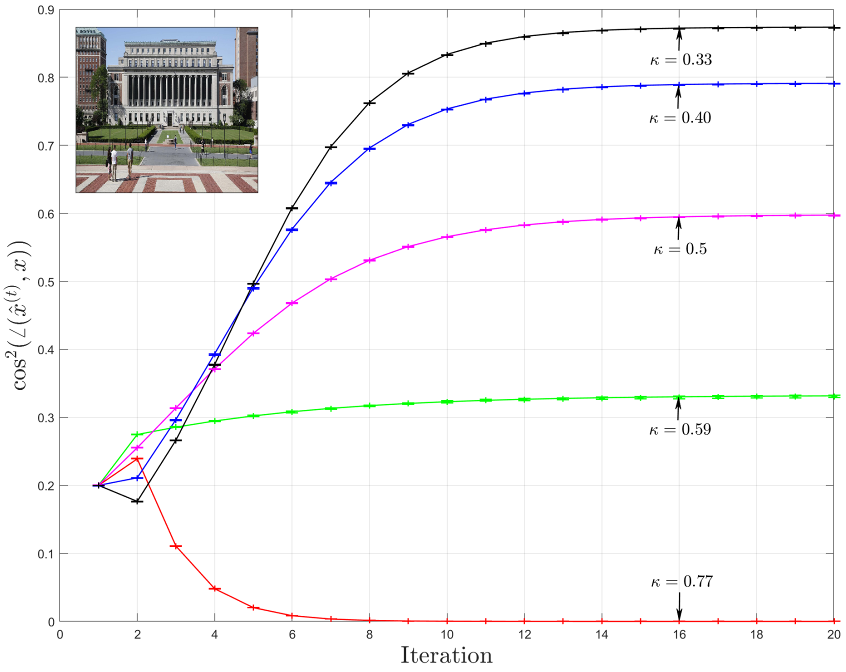


Figure 4.1: Solid Lines: Predicted Dynamics derived using State Evolution (Prop. 6 developed for sub-sampled Haar sensing, + markers: Dynamics of Linearized Message Passing averaged over 10 repetitions with sub-sampled Hadamard sensing and a real image (shown in inset) used as the signal vector. The error bars represent the standard error across repetitions.

Assumption on the signal: It is easy to see that, unlike in the sub-sampled Haar case, the state evolution cannot hold for arbitrary worst case signal vectors for the sub-sampled Hadamard sensing models since the orthogonal signal vectors $\sqrt{m}\boldsymbol{e}_1$ and $\sqrt{m}\boldsymbol{e}_2$ generate the same measurement vector $\boldsymbol{y} = (1, 1 \cdots, 1)^{\mathsf{T}}$. This is a folklore argument for non-indentifiability of the phase retrieval problem for ±1 sensing matrices [95]. Hence we study the universality phenomena under the simplest average case assumption on the signal, namely $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_n/\kappa\right)$.

4.1.2 Notation

Important Sets: $\mathbb{N}, \mathbb{N}_0, \mathbb{R}, \mathbb{C}$ denote the sets of natural numbers, non-negative integers, real numbers, and complex numbers, respectively. $[k]$ denotes the set $\{1, 2, \cdots, k\}$ and $[i : j]$ denotes the set $\{i, i+1, i+2 \cdots, j-1, j\}$. $\mathbb{O}_m$ refers to the set of all $m \times m$ orthogonal matrices and $\mathbb{U}_m$ refers to the set of all $m \times m$ unitary matrices.

Stochastic Convergence: $\xrightarrow{\text{P}}$ denotes convergence in probability. If for a sequence of random variables we have $X_n \xrightarrow{\text{P}} c$ for a deterministic c, we say $\text{p-lim}\, X_n = c$.

Linear Algebraic Aspects: We will use bold face letters to refer to vectors and matrices. For a matrix $\boldsymbol{V} \in \mathbb{R}^{m \times n}$, we adopt the convention of referring to the columns of $\boldsymbol{V}$ by $\boldsymbol{V}_1, \boldsymbol{V}_2 \cdots \boldsymbol{V}_n \in \mathbb{R}^m$ and to the rows by $\boldsymbol{v}_1, \boldsymbol{v}_2 \cdots \boldsymbol{v}_m \in \mathbb{R}^n$. For a vector $\boldsymbol{v}$, $\|\boldsymbol{v}\|_1, \|\boldsymbol{v}\|_2, \|\boldsymbol{v}\|_\infty$ denote the ℓ_1, ℓ_2, and ℓ_∞ norms, respectively. By default, $\|\boldsymbol{v}\|$ denotes the ℓ_2 norm. For a matrix $\boldsymbol{V}$, $\|\boldsymbol{V}\|_{\mathsf{op}}, \|\boldsymbol{V}\|_{\mathsf{Fr}}, \|\boldsymbol{V}\|_\infty$ denote the operator norm, Frobenius norm, and the entry-wise ∞-norm, respectively. For vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$, $\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle$ denotes the inner product $\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle = \sum_{i=1}^n v_{1i} v_{2i}$. For matrices $\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathbb{R}^{m \times n}$, $\langle \boldsymbol{V}_1, \boldsymbol{V}_2 \rangle$ denotes the matrix inner product $\sum_{i=1}^m \sum_{j=1}^n (V_1)_{ij} (V_2)_{ij}$.

Important distributions: $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the scalar Gaussian distribution with mean μ and variance σ^2. $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\mathsf{Bern}(p)$ denotes Bernoulli distribution with bias p. $\mathsf{Binom}(n, p)$ denotes the Binomial distribution with n trials and bias p. For an arbitrary set S, $\text{Unif}\,(S)$ denotes the uniform distribution on the elements of S. For example, $\text{Unif}\,(\mathbb{O}_m)$ denotes the Haar measure on the orthogonal group.

Order Notation and Constants: We use the standard $O(\cdot)$ notation. C will be used to refer to a universal constant independent of all parameters. When the constant C depends on a parameter k

we will make this explicit by using the notation C_k or $C(k)$. We say a sequence $a_n = O(\text{polylog}(n))$ if there exists a fixed, finite constant K such that $a_n \leq O(\log^K(n))$.

4.2 Main Result

Now, we are ready to state our main result.

Theorem 7. *Consider the linear message passing iterations* (4.6). *Suppose that:*

1. *The functions η_t are bounded and Lipchitz.*
2. *The signal is generated from the Gaussian prior: $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{0}, \frac{1}{\kappa}\boldsymbol{I}_n\right)$.*
3. *The sensing matrix is generated from the sub-sampled Hadamard ensemble.*
4. *The iteration* (4.6) *is initialized as:*

$$\hat{\boldsymbol{z}}^{(0)} = \alpha_0 \boldsymbol{z} + \sigma_0 \boldsymbol{w},$$

where $\alpha_0 \in \mathbb{R}, \sigma_0 \in \mathbb{R}_+$ are fixed and $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$.

Then for any fixed $t \in \mathbb{N}$, as $m, n \to \infty$, $n = \kappa m$, we have,

$$\frac{\langle \hat{\boldsymbol{z}}^{(t)}, \boldsymbol{z} \rangle}{m} \xrightarrow{P} \alpha_t, \quad \frac{\|\hat{\boldsymbol{z}}^{(t)}\|_2^2}{m} \xrightarrow{P} \alpha_t^2 + \sigma_t^2,$$
$$\frac{\langle \hat{\boldsymbol{x}}^{(t)}, \boldsymbol{x} \rangle}{m} \xrightarrow{P} \alpha_t, \quad \frac{\|\hat{\boldsymbol{x}}^{(t)}\|_2^2}{m} \xrightarrow{P} \alpha_t^2 + (1-\kappa)\sigma_t^2,$$

where (α_t, σ_t^2) are given by the recursion in (4.10).

Theorem 7 simply states that the dynamics of linearized message passing in the sub-sampled Hadamard model are asymptotically indistinguishable from the dynamics in the sub-sampled Haar model. This provides a theoretical justification for the universality depicted in Figure 4.1.

4.3 Related Work

Gaussian Universality: A number of papers have tried to explain the observations of [108] regarding the universality in performance of ℓ_1 minimization for noiseless linear sensing. For noiseless linear sensing, the Gaussian sensing ensemble, sub-sampled Haar sensing ensemble, and structured sensing ensembles like sub-sampled Fourier sensing ensemble behave identically. Consequently, a number of papers have tried to identify the class of sensing matrices which behave like Gaussian sensing matrices. It has been shown that sensing matrices with i.i.d. entries under mild moment assumptions behave like Gaussian sensing matrices in the context of performance of general (non-linear) Approximate Message Passing schemes [99, 113], the limiting Bayes risk [106], and the performance of estimators based on convex optimization [114, 115]. The assumption that the sensing matrix has i.i.d. entries has been relaxed to the assumption that it has i.i.d. rows (with possible dependence within a row) [97]. Finally, we emphasize that in the presence of noise or when the measurements are non-linear, the structured ensembles that we consider here, obtained by sub-sampling a deterministic orthogonal matrix like the Hadamard-Walsh matrix, no longer behave like Gaussian matrices, but rather like sub-sampled Haar matrices.

A result for highly structured ensembles: While the results mentioned above move beyond i.i.d. Gaussian sensing, the sensing matrices they consider are still largely unstructured and highly random. In particular, they do not apply to the sub-sampled Hadamard ensemble considered here. A notable exception is the work of [116] which considers a random undetermined system of linear equations (in $\boldsymbol{x}$) of the form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}_0$ for a random matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and a k-sparse non-negative vector $\boldsymbol{x}_0 \in \mathbb{R}^n_{\geq 0}$. [116] shows that as $m, n, k \rightarrow \infty$ such that $n/m \rightarrow \kappa_1, k/m \rightarrow \kappa_2$, the probability that $\boldsymbol{x}_0$ is the unique non-negative solution to the system sharply transitions from 0 to 1 depending on the values κ_1, κ_2. Moreover, this transition is universal across a wide range of random $\boldsymbol{A}$, including Gaussian ensembles, random matrices with i.i.d. entries sampled from a symmetric distribution, and highly structured ensembles whose null space is given by a random matrix $\boldsymbol{B} \in \mathbb{R}^{n-m \times n}$ generated by multiplying the columns of a fixed matrix $\boldsymbol{B}_0$ whose columns are in general position by i.i.d.

random signs. The proof technique of [116] uses results from the theory of random polytopes and it is not obvious how to extend their techniques beyond the case of solving under-determined linear equations.

Universality Results in Random Matrix Theory: The phenomena that structured orthogonal matrices, such as Hadamard and Fourier matrices, behave like random Haar matrices in some aspects has been studied in the context of random matrix theory [117] and in particular free probability [80]. A well known result in free probability (see the book of [80] for a textbook treatment) is that if $\boldsymbol{U} \sim \text{Unif}\left(\mathbb{U}(m)\right)$ and $\boldsymbol{D}_1, \boldsymbol{D}_2$ are deterministic $m \times m$ diagonal matrices then $\boldsymbol{U}\boldsymbol{D}_1\boldsymbol{U}^*$ and $\boldsymbol{D}_2$ are asymptotically free and consequently the limiting spectral distribution of matrix polynomials in $\boldsymbol{D}_2$ and $\boldsymbol{U}\boldsymbol{D}_1\boldsymbol{U}^*$ can be described in terms of the limiting spectral distribution of $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. [118, 119] have obtained an extension of this result where a Haar unitary matrix is replaced by $m \times m$ Fourier matrix: If $\boldsymbol{D}_1, \boldsymbol{D}_2$ are independent diagonal matrices then $\boldsymbol{F}_m\boldsymbol{D}_1\boldsymbol{F}_m^*$ is asymptotically free from $\boldsymbol{D}_2$. The result of these authors has been extended to other deterministic orthogonal/unitary matrices (such as the Hadamard-Walsh matrix) conjugated by random signed permutation matrices by [120]. In order to see how the result of [118] connects with ours note that the linearized AMP iterations (4.6) involve 2 random matrices: $\boldsymbol{H}\overline{\boldsymbol{B}}\boldsymbol{H}^\mathsf{T}$ and $q(\boldsymbol{Y})$. Note that if $\boldsymbol{B}$ and the diagonal matrix $q(\boldsymbol{Y})$ were independent, then the result of [118] would imply that $\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\mathsf{T}$ and $q(\boldsymbol{Y})$ are asymptotically free and this could potentially be used to analyze the linearized AMP algorithm. However, the key difficulty is that the measurements $\boldsymbol{y}$ depend on which columns of the Hadamard-Walsh matrix were selected (specified by $\boldsymbol{B}$). Infact, this dependence is precisely what allows the linearized AMP algorithm to recover the signal. However, we still find some of the techniques introduced by [118] useful in our analysis. We also emphasize that asymptotic freeness of $\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\mathsf{T},\ q(\boldsymbol{Y})$ alone seems to be insufficient to characterize the behavior of Linearized AMP algorithms. Asymptotic freeness implies that the expected normalized trace of certain matrix products involving $\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\mathsf{T},\ q(\boldsymbol{Y})$ vanish in the limit $m \rightarrow \infty$. On the other hand, our proof also requires the analysis of certain quadratic forms involving $\boldsymbol{H}\boldsymbol{B}\boldsymbol{H}^\mathsf{T},\ q(\boldsymbol{Y})$ (see Proposition 8) which do not appear to have been studied in the free probability literature.

Non-rigorous Results from Statistical Physics: In the statistical physics literature Cakmak, Opper, Winther, and Fleury [121, 122, 123, 124, 125] have developed an analysis of message passing algorithms for rotationally invariant ensembles via a non-rigorous technique called the dynamical functional theory. These works are interesting because they do not heavily rely on rotational invariance, but instead rely on results from Free probability. Since some of the free probability results have been extended to Fourier and Hadamard matrices [118, 119, 120], there is hope to generalize their analysis beyond rotationally invariant ensembles. However, currently, their results are non-rigorous due to two reasons: 1) due to the use of dynamical field theory, and 2) their application of Free probability results neglects dependence between matrices. In our work, we avoid the use of dynamical functional theory since we analyze linearized AMP algorithms and furthermore, we properly account for dependence that is heuristically neglected in their work.

The Hidden Manifold Model: Lastly, we discuss the recent works of [126, 127, 128], where they study statistical learning problems where the feature matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ (the analogue of the sensing matrix in statistical learning) is generated as:

$$\boldsymbol{A} = \sigma(\boldsymbol{Z}\boldsymbol{F}),$$

where $\boldsymbol{F} \in \mathbb{R}^{d \times n}$ is a generic (possibly structured) deterministic weight matrix and $\boldsymbol{Z} \in \mathbb{R}^{m \times d}$ is an i.i.d. Gaussian matrix. The function $\sigma : \mathbb{R} \to \mathbb{R}$ acts entry-wise on the matrix $\boldsymbol{Z}\boldsymbol{F}$. For this model, the authors have analyzed the dynamics of online (one-pass) stochastic gradient descent (first non-rigorously [126] and then rigorously [128]) and the performance of regularized empirical risk minimization with convex losses (non-rigorously) via the replica method [127] in the high dimensional asymptotic $m, n, d \to \infty$, $n/m \to \kappa_1$, $d/m \to \kappa_2$. Their results show that in this case the feature matrix behaves like a certain correlated Gaussian feature matrix. We note that the feature matrix $\boldsymbol{A}$ here is quite different from the sub-sampled Hadamard ensemble since it uses $O(m^2)$ i.i.d. random variables ($\boldsymbol{Z}$) where as the sub-sampled Hadamard ensemble only uses m i.i.d. random variables (to specify the permutation matrix $\boldsymbol{P}$). However, a technical result proved by the authors

(Lemma A.2 of [126]) appears to be a special case of a classical result of [129, 130] which we find useful to account for the dependence between the matrices $q_t(\boldsymbol{Y}), \boldsymbol{A}$ appearing in the linearized AMP iterations (4.6).

4.4 Proof Overview

Our basic strategy to prove Theorem 7 will be as follows: Throughout the chapter we will assume that Assumptions 1, 2, and 4 of Theorem 7 hold. We will seek to only show that the observables:

$$\frac{\langle \hat{\boldsymbol{z}}^{(t)}, \boldsymbol{z} \rangle}{m}, \ \frac{\|\hat{\boldsymbol{z}}^{(t)}\|_2^2}{m}, \frac{\langle \hat{\boldsymbol{x}}^{(t)}, \boldsymbol{x} \rangle}{m}, \ \frac{\|\hat{\boldsymbol{x}}^{(t)}\|_2^2}{m}, \tag{4.11}$$

have the same limit in probability under both the sub-sampled Haar and the sub-sampled Hadamard sensing models. We will not need to explicitly identify their limits since Proposition 6 already identifies the limit for us, and hence, Theorem 7 will follow.

It turns out the limits of the observables (4.11) depends only on normalized traces and quadratic forms of certain alternating products of the matrices $\boldsymbol{\Psi}$ and $\boldsymbol{Z}$. Hence, we introduce the following definition.

Definition 7 (Alternating Product). *A matrix $\mathcal{A}$ is said to be a alternating product of matrices $\boldsymbol{\Psi}, \boldsymbol{Z}$ if there exist polynomials $p_i : \mathbb{R} \rightarrow \mathbb{R}, \ i \in 1, 2 \ldots, k$, and bounded, Lipchitz functions $q_i : \mathbb{R} \rightarrow \mathbb{R}, \ i \in \{1, 2 \ldots k\}$ such that:*

1. *If $B \sim \mathsf{Bern}(\kappa)$, $\mathbb{E} p_i(B - \kappa) = 0$.*

2. *q_i are even functions i.e. $q_i(\xi) = q_i(-\xi)$ and if $\xi \sim \mathcal{N}(0, 1)$, then, $\mathbb{E} q_i(\xi) = 0$,*

and, $\mathcal{A}$ is one of the following:

1. *Type 1: $\mathcal{A} = p_1(\boldsymbol{\Psi}) q_1(\boldsymbol{Z}) p_2(\boldsymbol{\Psi}) \cdots q_{k-1}(\boldsymbol{Z}) p_k(\boldsymbol{\Psi})$*

2. *Type 2: $\mathcal{A} = p_1(\boldsymbol{\Psi}) q_1(\boldsymbol{Z}) p_2(\boldsymbol{\Psi}) q_2(\boldsymbol{Z}) \cdots p_k(\boldsymbol{\Psi}) q_k(\boldsymbol{Z})$*

3. *Type 3:* $\mathcal{A} = q_1(\mathbf{Z})p_2(\mathbf{\Psi})q_2(\mathbf{Z}) \cdots p_k(\mathbf{\Psi})q_k(\mathbf{Z})$.

4. *Type 4:* $\mathcal{A} = q_1(\mathbf{Z})p_2(\mathbf{\Psi})q_2(\mathbf{Z})p_3(\mathbf{\Psi}) \cdots q_{k-1}(\mathbf{Z})p_k(\mathbf{\Psi})$.

In the above definitions:

1. *The scalar polynomial p_i is evaluated at the matrix $\mathbf{\Psi}$ in the usual sense, for example if $p(\psi) = \psi^2$, then, $p(\mathbf{\Psi}) = \mathbf{\Psi}^2$.*

2. *The functions q_i are evaluated entry-wise on the diagonal matrix $\mathbf{Z}$, i.e.*

$$q_i(\mathbf{Z}) = \mathit{Diag}\left(q_i(z_1), q_i(z_2) \dots q_i(z_m)\right).$$

We note that alternating products are a central notion in free probability [80]. The difference here is that we have additionally constrained the functions p_i, q_i in Definition 7.

Theorem 7 is a consequence of two properties of alternating products which may be of independent interest. These are stated in the following propositions.

Proposition 7. *Let $\mathcal{A}(\mathbf{\Psi}, \mathbf{Z})$ be an alternating product of matrices $\mathbf{\Psi}, \mathbf{Z}$. Suppose the sensing matrix $\boldsymbol{A}$ is generated from the sub-sampled Haar sensing model, or the sub-sampled Hadamard sensing model, or by sub-sampling a deterministic orthogonal matrix $\boldsymbol{U}$ with the property:*

$$\|\boldsymbol{U}\|_\infty \leq \sqrt{\frac{K_1 \log^{K_2}(m)}{m}}, \ \forall m \ \geq K_3,$$

for some fixed constants K_1, K_2, K_3. Then,

$$\mathsf{Tr}(\mathcal{A}(\mathbf{\Psi}, \mathbf{Z}))/m \xrightarrow{P} 0.$$

Proposition 8. *Let $\mathcal{A}(\mathbf{\Psi}, \mathbf{Z})$ be an alternating product of matrices $\mathbf{\Psi}, \mathbf{Z}$. Then for the sub-sampled*

Haar sensing model and for sub-sampled Hadamard ($\boldsymbol{U} = \boldsymbol{H}$) sensing model, we have,

$$\text{p-lim} \ \frac{\langle z, \mathcal{A}z \rangle}{m}$$

exists and is identical for the two models.

Outline of the Remaining Chapter: The remainder of the chapter is organized as follows:

1. In Section 4.5 we provide a proof of Theorem 7 assuming Propositions 7 and 8.
2. In Section 4.6 we introduce some key tools required for the proof of Propositions 7 and 8.
3. The proof of Proposition 7 can be found in Section 4.7.
4. The proof of Proposition 8 can be found in Section 4.8.

4.5 Proof of Theorem 7

In this section we will show the analysis of the observables (4.11) reduces to the analysis of the normalized traces and quadratic forms of alternating products. In particular, we will prove Theorem 7 using Propositions 7 and 8.

Proof of Theorem 7. For simplicity, we will assume the functions η_t do not change with t, i.e. $\eta_t = \eta \ \forall \ t \geq 0$. This is just to simplify notations, and the proof of time varying η_t is exactly the same. Define the function:

$$q(z) = \eta(|z|) - \mathbb{E}_{Z \sim \mathcal{N}(0,1)}[\eta(|Z|)].$$

Note that the linearized message passing iterations (4.6) can be expressed as:

$$\hat{\boldsymbol{z}}^{(t+1)} = \frac{1}{\kappa} \cdot \boldsymbol{\Psi} \cdot q(\boldsymbol{Z}) \cdot \hat{\boldsymbol{z}}^{(t)}.$$

Unrolling the iterations we obtain:

$$\hat{z}^{(t)} = \frac{1}{\kappa^t} \cdot (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot \hat{z}^{(0)}.$$

Note that the initialization is assumed to be of the form: $\hat{z}^{(0)} = \alpha_0 z + \sigma_0 \boldsymbol{w}$, where $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I})$. Hence:

$$\begin{aligned}
\hat{z}^{(t)} &= \alpha_0 \frac{1}{\kappa^t} \cdot (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot z + \sigma_0 \cdot \frac{1}{\kappa^t} \cdot (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot \boldsymbol{w}, \\
\hat{\boldsymbol{x}}^{(t)} &= \boldsymbol{A}^\mathsf{T} \hat{z}^{(t)}.
\end{aligned}$$

We will focus on showing that the limits:

$$\text{p-lim} \frac{\langle \boldsymbol{x}, \hat{\boldsymbol{x}}^{(t)} \rangle}{m}, \ \text{p-lim} \frac{\|\hat{\boldsymbol{x}}^{(t)}\|_2^2}{m}, \tag{4.12}$$

exist and are identical for the two models. The claim for the limits corresponding to $\hat{z}^{(t)}$ are exactly analogous and omitted. Hence, the remainder of the proof is devoted to analyzing the above limits.

Analysis of $\langle \boldsymbol{x}, \hat{\boldsymbol{x}}^{(t)} \rangle$: Observe that:

$$\begin{aligned}
\langle \boldsymbol{x}, \hat{\boldsymbol{x}}^{(t)} \rangle &= \langle \boldsymbol{A}^\mathsf{T} z, \boldsymbol{A}^\mathsf{T} \hat{z}^{(t)} \rangle \\
&= \alpha_0 \frac{1}{\kappa^t} \cdot \underbrace{\langle \boldsymbol{A}^\mathsf{T} z, \boldsymbol{A}^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot z \rangle}_{(T_1)} + \sigma_0 \cdot \frac{1}{\kappa^t} \cdot \underbrace{\langle \boldsymbol{A}^\mathsf{T} z, \boldsymbol{A}^\mathsf{T} \cdot (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot \boldsymbol{w} \rangle}_{(T_2)}.
\end{aligned}$$

We first analyze term (T_1). Observe that:

$$\begin{aligned}
(T_1) &= z^\mathsf{T} \boldsymbol{A} \boldsymbol{A}^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t z \\
&= z^\mathsf{T} \boldsymbol{\Psi} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t z + \kappa z^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t z \\
&= z^\mathsf{T} \boldsymbol{\Psi}^2 (q(\boldsymbol{Z}) \boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) z + \kappa z^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t z \\
&\overset{(a)}{=} z^\mathsf{T} p(\boldsymbol{\Psi}) (q(\boldsymbol{Z}) \boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) z + \kappa(1-\kappa) z^\mathsf{T} (q(\boldsymbol{Z}) \boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) z + \kappa z^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t z.
\end{aligned}$$

In the step marked (a) we defined the polynomial $p(\psi) = \psi^2 - \kappa(1-\kappa)$ which has the property $\mathbb{E}p(B-\kappa) = 0$ when $B \sim \mathsf{Bern}(\kappa)$. One can check that $Z \sim \mathcal{N}(0,1)$, $\mathbb{E}q(Z) = 0$, and q is a bounded, Lipchitz, even function. Hence, each of the terms appearing in step (a) are of the form $\boldsymbol{z}^\mathsf{T}\boldsymbol{\mathcal{A}}\boldsymbol{z}$ for some alternating product $\mathcal{A}$ (Definition 7) of matrices $\boldsymbol{\Psi}, \boldsymbol{Z}$. Consequently, by Proposition 8 we obtain that term (1) divided by m converges to the same limit in probability under both the sub-sampled Haar sensing and the sub-sampled Hadamard sensing model. Next, we analyze (T_2). Note that:

$$\begin{aligned}\frac{\langle \boldsymbol{A}^\mathsf{T}\boldsymbol{z}, \boldsymbol{A}^\mathsf{T}\cdot(\boldsymbol{\Psi}\cdot q(\boldsymbol{Z}))^t\cdot \boldsymbol{w}\rangle}{m} &= \boldsymbol{z}^\mathsf{T}\boldsymbol{A}\boldsymbol{A}^\mathsf{T}(\boldsymbol{\Psi}\cdot q(\boldsymbol{Z}))^t\boldsymbol{w}/m \\ &\overset{\mathrm{d}}{=} \frac{\|(q(\boldsymbol{Z})\boldsymbol{\Psi})^t\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{z}\|_2}{m}\cdot W,\ W \sim \mathcal{N}(0,1),\end{aligned}$$

where $\overset{\mathrm{d}}{=}$ means both sides have a same distribution. Observe that:

$$\begin{aligned}\frac{\|(q(\boldsymbol{Z})\boldsymbol{\Psi})^t\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{z}\|_2}{m} &= \frac{\|(q(\boldsymbol{Z})\boldsymbol{\Psi})^t\boldsymbol{A}\boldsymbol{x}\|_2}{m} \\ &\leq \|(q(\boldsymbol{Z})\boldsymbol{\Psi})^t\boldsymbol{A}\|_{\mathsf{op}}\cdot\frac{\|\boldsymbol{x}\|_2}{m} \\ &\leq \|q(\boldsymbol{Z})\|_{\mathsf{op}}^t\|\boldsymbol{\Psi}\|_{\mathsf{op}}^t\|\boldsymbol{A}\|_{\mathsf{op}}\cdot\frac{\|\boldsymbol{x}\|_2}{m}.\end{aligned}$$

It is easy to check that: $\|q(\boldsymbol{Z})\|_{\mathsf{op}} \leq 2\|\eta\|_\infty < \infty$. Similarly, $\|\boldsymbol{\Psi}\|_{\mathsf{op}} \leq 1,\ \|\boldsymbol{A}\|_{\mathsf{op}} = 1$. Hence,

$$\frac{\|(q(\boldsymbol{Z})\boldsymbol{\Psi})^t\boldsymbol{A}\boldsymbol{A}^\mathsf{T}\boldsymbol{z}\|_2}{m} \leq 2^t\|\eta\|_\infty^t\cdot\sqrt{\frac{\|\boldsymbol{x}\|^2}{m}}\cdot\frac{1}{\sqrt{m}}$$

Observing that $\|\boldsymbol{x}\|^2/m \xrightarrow{\mathrm{P}} 1$ we obtain:

$$\left|\frac{\langle \boldsymbol{A}^\mathsf{T}\boldsymbol{z}, \boldsymbol{A}^\mathsf{T}\cdot(\boldsymbol{\Psi}\cdot q(\boldsymbol{Z}))^t\cdot \boldsymbol{w}\rangle}{m}\right| \leq 2^t\|\eta\|_\infty^t\cdot\sqrt{\frac{\|\boldsymbol{x}\|^2}{m}}\cdot\frac{|W|}{\sqrt{m}} \xrightarrow{\mathrm{P}} 0.$$

Note the above result holds for both subsampled Haar sensing and subsampled Hadamard sensing. This proves that the limit

$$\text{p-lim} \frac{\langle \boldsymbol{x}, \hat{\boldsymbol{x}}^{(t)} \rangle}{m}$$

exists and is identical for the two models.

Analysis of $\|\hat{\boldsymbol{x}}^{(t)}\|^2$: Recalling that:

$$\begin{aligned}
\hat{\boldsymbol{z}}^{(t)} &= \alpha_0 \frac{1}{\kappa^t} \cdot (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot \boldsymbol{z} + \sigma_0 \frac{1}{\kappa^t} \cdot (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot \boldsymbol{w}, \\
\hat{\boldsymbol{x}}^{(t)} &= \boldsymbol{A}^\mathsf{T} \hat{\boldsymbol{z}}^{(t)},
\end{aligned}$$

we can compute:

$$\frac{1}{m} \|\hat{\boldsymbol{x}}^{(t)}\|_2^2 = \frac{1}{\kappa^{2t}} \cdot \left(\alpha_0^2 \cdot (T_3) + 2\alpha_0 \sigma_0 (T_4) + \sigma_0^2 \cdot (T_5) \right),$$

where the terms $(T_3 - T_5)$ are defined as:

$$\begin{aligned}
(T_3) &= \frac{\boldsymbol{z}^\mathsf{T} (q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot \boldsymbol{z}}{m}, \\
(T_4) &= \frac{\boldsymbol{z}^\mathsf{T} (q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot \boldsymbol{w}}{m}, \\
(T_5) &= \frac{\boldsymbol{w}^\mathsf{T} (q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t \cdot \boldsymbol{w}}{m}.
\end{aligned}$$

We analyze each of these terms separately. First, consider (T_3). Our goal will be to decompose the matrix $(q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t$ as:

$$(q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^\mathsf{T} (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t = c_0 \boldsymbol{I} + \sum_{i=1}^{N_t} c_i \mathcal{A}_i,$$

where $\mathcal{A}_i$ are alternating products of the matrices $\boldsymbol{\Psi}, \boldsymbol{Z}$ (see Definition 7) and c_i are some scalar constants. This decomposition has the following properties: 1) It is independent of the choice of the orthogonal matrix $\boldsymbol{U}$ used to generate the sensing matrix. 2) The number of terms in the decomposition N_t depends only on t and not on m, n. In order to see why such a decomposition exists: first recall that $\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{\Psi} + \kappa \boldsymbol{I}_m$. Hence, we can write:

$$(q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t = (q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{\Psi}(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t + \kappa \boldsymbol{z}^{\mathsf{T}}(q(\boldsymbol{Z})\boldsymbol{\Psi})^t(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t$$
$$= (q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) \Psi^3 q(\boldsymbol{Z})(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1} + \kappa \boldsymbol{z}^{\mathsf{T}}(q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) \boldsymbol{\Psi}^2 q(\boldsymbol{Z})(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1}.$$

For any $i \in \mathbb{N}$, we write $\Psi^i = p_i(\boldsymbol{\Psi}) + \mu_i \boldsymbol{I}$, where $\mu_i = \mathbb{E}(B - \kappa)^i$, $B \sim \mathsf{Bern}(\kappa)$, and $p_i(\psi) = \psi^i - \mu_i$. This polynomial satisfies $\mathbb{E} p_i(B - \kappa) = 0$. This gives us:

$$\begin{aligned}(q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t &= (q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) p_3(\boldsymbol{\Psi}) q(\boldsymbol{Z})(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1} \\ &+ \kappa \boldsymbol{z}^{\mathsf{T}}(q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) \boldsymbol{p}_2(\Psi) q(\boldsymbol{Z})(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1} \\ &+ (\mu_3 + \kappa\mu_2) \cdot (q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z})^2 (\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1}.\end{aligned}$$

In the above display, the first two terms on the RHS are in the desired alternating product form. We center the last term. For any $i \in \mathbb{N}$ we define $q_i(z) = q^i(z) - \nu_i$, $\nu_i = \mathbb{E}q(\xi)^i$, $\xi \sim \mathcal{N}(0, 1)$. Hence, $q^i(\boldsymbol{Z}) = q_i(\boldsymbol{Z}) + \nu_i \boldsymbol{I}_m$. Hence:

$$\begin{aligned}(q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^t &= (q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) p_3(\boldsymbol{\Psi}) q(\boldsymbol{Z})(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1} \\ &+ \kappa \boldsymbol{z}^{\mathsf{T}}(q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1} q(\boldsymbol{Z}) \boldsymbol{p}_2(\Psi) q(\boldsymbol{Z})(\boldsymbol{\Psi} q(\boldsymbol{Z}))^{t-1} \\ &+ (\mu_3 + \kappa\mu_2)(q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1} q_2(\boldsymbol{Z})(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1} \\ &+ \nu_2\ (\mu_3 + \kappa\mu_2)(q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1}(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1}.\end{aligned}$$

In the above display, each of the terms in the right hand side is an alternating product except $(\mu_3 + \kappa\mu_2) \cdot (q(\boldsymbol{Z})\boldsymbol{\Psi})^{t-1}(\boldsymbol{\Psi} \cdot q(\boldsymbol{Z}))^{t-1}$. We inductively center this term. Note that this centering procedure does not depend on the choice of the orthogonal matrix $\boldsymbol{U}$ used to generate the sensing matrix. Furthermore, the number of terms is bounded by $N_t \leq N_{t-1} + 3$, so $N_t \leq 1 + 3t$.

Hence, we have obtained the desired decomposition:

$$(q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^\mathsf{T}(\boldsymbol{\Psi}\cdot q(\boldsymbol{Z}))^t = c_0 \boldsymbol{I} + \sum_{i=1}^{N_t} c_i \mathcal{A}_i. \tag{4.13}$$

Therefore, we can write (T_3) as:

$$(T_3) = c_0 \frac{\|\boldsymbol{z}\|^2}{m} + \frac{1}{m}\sum_{i=1}^{N_t} c_i \; \boldsymbol{z}^\mathsf{T}\mathcal{A}_i \boldsymbol{z} = c_0 \frac{\|\boldsymbol{x}\|^2}{m} + \frac{1}{m}\sum_{i=1}^{N_t} c_i \; \boldsymbol{z}^\mathsf{T}\mathcal{A}_i \boldsymbol{z}.$$

Observe that $\|\boldsymbol{x}\|^2/m \xrightarrow{\text{P}} 1$, and Proposition 8 guarantees $\boldsymbol{z}^\mathsf{T}\mathcal{A}_i\boldsymbol{z}/m$ converges in probability to the same limit irrespective of whether $\boldsymbol{U} = \boldsymbol{O}$ or $\boldsymbol{U} = \boldsymbol{H}$. Hence, term (T_3) converges in probability to the same limit for both the subsampled Haar sensing and the subsampled Hadamard sensing model.

Next, we analyze term (T_4). Repeating the arguments we made for the analysis of the term (T_2) we find:

$$\begin{aligned}(T_4) &= \frac{\boldsymbol{z}^\mathsf{T}(q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^\mathsf{T}(\boldsymbol{\Psi}\cdot q(\boldsymbol{Z}))^t \cdot \boldsymbol{w}}{m} \\ &\overset{\text{d}}{=} \frac{\|(q(\boldsymbol{Z})\boldsymbol{\Psi})^t \boldsymbol{A}\boldsymbol{A}^\mathsf{T}(\boldsymbol{\Psi}\cdot q(\boldsymbol{Z}))^t \boldsymbol{z}\|_2}{m} \cdot W \xrightarrow{\text{P}} 0,\end{aligned}$$

where $W \sim \mathcal{N}(0,1)$. Finally, we analyze the term (T_5). Using the decomposition (4.13) we have:

$$(T_5) = c_0 \frac{\|\boldsymbol{w}\|_2^2}{m} + \frac{1}{m}\sum_{i=1}^{N_t} c_i \; \boldsymbol{w}^\mathsf{T}\mathcal{A}_i \boldsymbol{w}.$$

We know that $\|\boldsymbol{w}\|_2^2/m \xrightarrow{\text{P}} 1$. Hence, we focus on analyzing $\boldsymbol{w}^\mathsf{T}\mathcal{A}_i\boldsymbol{w}/m$. We decompose this as:

$$\frac{\boldsymbol{w}^\mathsf{T}\mathcal{A}_i\boldsymbol{w}}{m} = \frac{w^\mathsf{T}\mathcal{A}_i\boldsymbol{w} - \mathbb{E}[w^\mathsf{T}\mathcal{A}_i\boldsymbol{w}|\mathcal{A}_i]}{m} + \frac{\mathbb{E}[w^\mathsf{T}\mathcal{A}_i\boldsymbol{w}|\mathcal{A}_i]}{m}.$$

Observe that:

$$\frac{\mathbb{E}[\boldsymbol{w}^\mathsf{T}\mathcal{A}_i\boldsymbol{w}|\mathcal{A}_i]}{m} = \frac{\kappa \cdot \mathsf{Tr}(\mathcal{A}_i)}{m} \xrightarrow{\text{P}} 0 \quad \text{(By Proposition 7).}$$

On the other hand, using the Hanson-Wright Inequality (Fact 2) together with the estimates

$$\|\mathcal{A}_i\|_{\mathsf{op}} \leq C(\mathcal{A}_i), \ \|\mathcal{A}_i\|_{\mathsf{Fr}} \leq \sqrt{m} \cdot C(\mathcal{A}_i),$$

for a fixed constant $C(\mathcal{A}_i)$ (independent of m, n) depending only on the formula for $\mathcal{A}_i$, we obtain:

$$\mathbb{P}\left(\left|\boldsymbol{w}^\mathsf{T}\mathcal{A}_i\boldsymbol{w} - \mathbb{E}[w^\mathsf{T}\mathcal{A}_i\boldsymbol{w}|\mathcal{A}_i]\right| > mt \ \middle| \ \mathcal{A}_i\right) \leq 2\exp\left(-\frac{c}{C(\mathcal{A}_i)} \cdot m \cdot \min(t, t^2)\right) \to 0.$$

Hence,

$$\frac{\boldsymbol{w}^\mathsf{T}\mathcal{A}_i\boldsymbol{w} - \mathbb{E}[w^\mathsf{T}\mathcal{A}_i\boldsymbol{w}|\mathcal{A}_i]}{m} \xrightarrow{\text{P}} 0.$$

This implies $(T_5) \xrightarrow{\text{P}} c_0$ for both the models. This proves the limit :

$$\text{p-lim} \frac{\|\hat{\boldsymbol{x}}^{(t)}\|_2^2}{m}$$

exists and is identical for the two sensing models, which concludes the proof of Theorem 7.

□

4.6 Key Ideas for the Proof of Propositions 7 and 8

In this section, we introduce some key ideas that are important in the proof of Propositions 7 and 8. Recall that we wish to analyze the limit in probability of the normalized trace and the quadratic form. A natural candidate for this limit is the limiting value of their expectation:

$$\begin{aligned}\text{p-lim}\,\frac{1}{m}\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z}) &\overset{?}{=} \lim_{m\to\infty}\frac{1}{m}\mathbb{E}\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z}),\\ \text{p-lim}\,\frac{\langle z,\mathcal{A}z\rangle}{m} &\overset{?}{=} \lim_{m\to\infty}\frac{\mathbb{E}\langle z,\mathcal{A}z\rangle}{m}.\end{aligned}$$

In order to show this, one needs to show that the variance of the normalized trace and the normalized quadratic form converge to 0, which involves analyzing the second moment of these quantities. However, since the analysis of the second moment uses very similar ideas as the analysis of the expectation, we focus on outlining the main ideas in the context of the analysis of expectation.

First, we observe that alternating products can be simplified significantly due to the following property of polynomials of centered Bernoulli random variables.

Lemma 17. *For any polynomial p such that if $B \sim \mathsf{Bern}(\kappa)$, $\mathbb{E}\; p(B-\kappa) = 0$ we have,*

$$p(\boldsymbol{\Psi}) = (p(1-\kappa) - p(-\kappa)) \cdot \boldsymbol{\Psi}.$$

Proof. Observe that since $\boldsymbol{\Psi} = \boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^{\mathsf{T}}$, and $\boldsymbol{U}$ is orthogonal, we have $p(\boldsymbol{\Psi}) = \boldsymbol{U}p(\overline{\boldsymbol{B}})\boldsymbol{U}^{\mathsf{T}}$. Next, observe that:

$$\begin{aligned}p(\overline{B}_{ii}) &= p(1-\kappa)B_{ii} + p(-\kappa)(1-B_{ii})\\ &= (p(1-\kappa)-p(-\kappa))\cdot\overline{B}_{ii} + \underbrace{\kappa p(1-\kappa) + (1-\kappa)p(-\kappa)}_{=0},\end{aligned}$$

where the last step follows from the assumption $\mathbb{E}\; p(B-\kappa) = 0$. Hence, $p(\overline{\boldsymbol{B}}) = (p(1-\kappa)-p(-\kappa))\overline{\boldsymbol{B}}$ and $p(\boldsymbol{\Psi}) = (p(1-\kappa)-p(-\kappa))\boldsymbol{\Psi}$. □

Hence, without loss of generality we can assume that each of the p_i in an alternating product satisfy $p_i(\xi) = \xi$.

4.6.1 Partitions

Note that the expected normalized trace and the expected quadratic form in Propositions 7 and 8 can be expanded as follows:

$$\frac{1}{m}\mathbb{E}\mathsf{Tr}\mathcal{A}(\mathbf{\Psi}, \mathbf{Z}) = \frac{1}{m}\sum_{a_1,a_2,\dots a_k=1}^{m} \mathbb{E}[(\mathbf{\Psi})_{a_1,a_2} q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k})(\mathbf{\Psi})_{a_k,a_1}],$$
$$\frac{\mathbb{E}\langle z, \mathcal{A}z\rangle}{m} = \frac{1}{m}\sum_{a_{1:k+1}\in[m]} \mathbb{E}[z_{a_1}(\mathbf{\Psi})_{a_1,a_2} q_1(z_{a_2})(\mathbf{\Psi})_{a_2,a_3} \cdots q_{k-1}(z_{a_k})(\mathbf{\Psi})_{a_k,a_{k+1}} z_{a_{k+1}}].$$

Some Notation: Let $\mathcal{P}([k])$ denotes the set of all partitions of a discrete set $[k]$. We use $|\pi|$ to denote the number of blocks in π. Recall that a partition $\pi \in \mathcal{P}([k])$ is simply a collection of disjoint subsets of $[k]$ whose union is $[k]$ i.e.

$$\pi = \{\mathcal{V}_1, \mathcal{V}_2 \dots \mathcal{V}_{|\pi|}\}, \ \sqcup_{t=1}^{|\pi|}\mathcal{V}_t = [k].$$

The symbol $\sqcup$ is exclusively reserved for representing a set as a union of disjoint sets. For any element $s \in [k]$, we use the notation $\pi(s)$ to refer to the block that s lies in. That is, $\pi(s) = \mathcal{V}_i$ iff $s \in \mathcal{V}_i$. For any $\pi \in \mathcal{P}([k])$, define the set $C(\pi)$ the set of all vectors $\boldsymbol{a} \in [m]^k$ which are constant exactly on the blocks of π:

$$C(\pi) \overset{\text{def}}{=} \{\boldsymbol{a} \in [m]^k : a_s = a_t \Leftrightarrow \pi(s) = \pi(t)\}.$$

Consider any $\boldsymbol{a} \in C(\pi)$. If $\mathcal{V}_i$ is a block in π, we use $a_{\mathcal{V}_i}$ to denote the unique value the vector $\boldsymbol{a}$ assigns to the all the elements of $\mathcal{V}_i$.

The rationale for introducing this notation is the observation that:

$$[m]^k = \bigsqcup_{\pi\in\mathcal{P}([k])} C(\pi),$$

and hence we can write the normalized trace and quadratic forms as:

$$\frac{\mathbb{E}\mathsf{Tr}\mathcal{A}(\mathbf{\Psi}, \mathbf{Z})}{m} = \frac{1}{m} \sum_{\pi \in \mathcal{P}([k])} \sum_{\mathbf{a} \in \mathcal{C}(\pi)} \mathbb{E}[(\mathbf{\Psi})_{a_1,a_2} q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k})(\mathbf{\Psi})_{a_k,a_1}], \tag{4.14a}$$

$$\frac{\mathbb{E}\langle z, \mathcal{A} z\rangle}{m} = \frac{1}{m} \sum_{\pi \in \mathcal{P}([k+1])} \sum_{\mathbf{a} \in \mathcal{C}(\pi)} \mathbb{E}[z_{a_1}(\mathbf{\Psi})_{a_1,a_2} q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k})(\mathbf{\Psi})_{a_k,a_{k+1}} z_{a_{k+1}}]. \tag{4.14b}$$

This idea of organizing the combinatorial calculations is due to [131] and the rationale for doing so will be clear in a moment.

4.6.2 Concentration

Lemma 18. *Let the sensing matrix $\mathbf{A}$ be generated by sub-sampling an orthogonal matrix $\mathbf{U}$. We have, for any $a, b \in [m]$:*

$$\mathbb{P}\left(|\Psi_{ab}| \geq \epsilon | \mathbf{U}\right) \leq 4 \exp\left(-\frac{\epsilon^2}{8m\|\mathbf{U}\|_\infty^4}\right).$$

Proof. Recall that $\mathbf{\Psi} = \mathbf{U}(\mathbf{B} - \kappa \mathbf{I}_m)\mathbf{U}^\mathsf{T}$, where the distribution of the diagonal matrix

$$\mathbf{B} = \text{Diag}\left(B_{11}, B_{22} \dots B_{mm}\right)$$

is described as follows: First draw a uniformly random subset $S \subset [m]$ with $|S| = n$ and set:

$$B_{ii} = \begin{cases} 0 & : i \notin S \\ 1 & : i \in S \end{cases}.$$

Due to the constraint that $\sum_{i=1}^m B_{ii} = n$, these random variables are not independent. In order to address this issue we couple $\mathbf{B}$ with another random diagonal matrix $\tilde{\mathbf{B}}$ generated as follows:

1. First sample $N \sim \mathsf{Binom}(m, \kappa)$.

2. Sample a subset $\tilde{S} \subset [m]$ with $|\tilde{S}| = N$ as follows:

- If $N \leq n$, then set $\tilde{S}$ to be a uniformly random subset of S of size N.
- If $N > n$ first sample a uniformly random subset A of S^c of size $N - n$ and set $\tilde{S} = S \cup A$.

3. Set $\tilde{\boldsymbol{B}}$ as follows:

$$\tilde{B}_{ii} = \begin{cases} 0 & : i \notin \tilde{S} \\ 1 & : i \in \tilde{S}. \end{cases}.$$

It is easy to check that conditional on N, $\tilde{S}$ is a uniformly random subset of $[m]$ with cardinality N. Since $N \sim \mathsf{Binom}(m, \kappa)$, we have $\tilde{B}_{ii} \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(\kappa)$. Define:

$$T \overset{\text{def}}{=} \Psi_{ab} = \boldsymbol{u}_a^\mathsf{T}(\boldsymbol{B} - \kappa \boldsymbol{I}_m)\boldsymbol{u}_b = \sum_{i=1}^m u_{ai}u_{bi}(B_{ii} - \mathbb{E}B_{ii}),$$

$$\tilde{T} \overset{\text{def}}{=} \boldsymbol{u}_a^\mathsf{T}(\tilde{\boldsymbol{B}} - \kappa \boldsymbol{I}_m)\boldsymbol{u}_b = \sum_{i=1}^m u_{ai}u_{bi}(\tilde{B}_{ii} - \mathbb{E}\tilde{B}_{ii}).$$

Observe that $|T - \tilde{T}| \leq |N - n| \|\boldsymbol{U}\|_\infty^2$. Hence,

$$\begin{aligned} \mathbb{P}\left(|T| \geq \epsilon\right) &\leq \mathbb{P}\left(|\tilde{T}| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(|T - \tilde{T}| \geq \frac{\epsilon}{2}\right) \\ &= \mathbb{P}\left(|\tilde{T}| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(|N - \mathbb{E}N| \geq \frac{\epsilon}{2\|\boldsymbol{U}\|_\infty^2}\right) \\ &\overset{\text{(a)}}{\leq} 4\exp\left(-\frac{\epsilon^2}{8m\|\boldsymbol{U}\|_\infty^4}\right). \end{aligned}$$

In the step marked (a), we used Hoeffding's Inequality. □

Hence the above lemma shows that,

$$\|\boldsymbol{\Psi}\|_\infty \leq O\left(\sqrt{m}\|\boldsymbol{U}\|_\infty^2 \operatorname{polylog}(m)\right),$$

with high probability. Recall that in the subsampled Hadamard model $\boldsymbol{U} = \boldsymbol{H}$ and $\|\boldsymbol{H}\|_\infty = 1/\sqrt{m}$. Similarly, in the subsampled Haar model $\boldsymbol{U} = \boldsymbol{O}$ and $\|\boldsymbol{O}\|_\infty \leq O(\operatorname{polylog}(m)/\sqrt{m})$. Hence, we expect:

$$\|\boldsymbol{\Psi}\|_\infty \leq O\left(\frac{\mathrm{polylog}(m)}{\sqrt{m}}\right), \text{ with high probability.} \tag{4.15}$$

4.6.3 Mehler's Formula

Note that in order to compute the expected normalized trace and quadratic form as given in (4.14), we need to compute:

$$\begin{aligned}
&\mathbb{E}[(\boldsymbol{\Psi})_{a_1,a_2} q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k})(\boldsymbol{\Psi})_{a_k,a_1}],\\
&\mathbb{E}[z_{a_1}(\boldsymbol{\Psi})_{a_1,a_2} q_1(z_{a_2})(\boldsymbol{\Psi})_{a_2,a_3} \cdots q_{k-1}(z_{a_k})(\boldsymbol{\Psi})_{a_k,a_{k+1}} z_{a_{k+1}}].
\end{aligned}$$

Note that by the Tower property:

$$\begin{aligned}
&\mathbb{E}[(\boldsymbol{\Psi})_{a_1,a_2} q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k})(\boldsymbol{\Psi})_{a_k,a_1}] =\\
&\qquad \mathbb{E}\left[(\boldsymbol{\Psi})_{a_1,a_2} \cdots (\boldsymbol{\Psi})_{a_k,a_1} \mathbb{E}[q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k}) | \boldsymbol{A}]\right],
\end{aligned}$$

and analogously for $\mathbb{E}[z_{a_1}(\boldsymbol{\Psi})_{a_1,a_2} q_1(z_{a_2})(\boldsymbol{\Psi})_{a_2,a_3} \cdots q_{k-1}(z_{a_k})(\boldsymbol{\Psi})_{a_k,a_{k+1}} z_{a_{k+1}}]$. Suppose that $\boldsymbol{a} \in \mathcal{C}(\pi)$ for some $\pi \in \mathcal{P}([k])$. Let $\pi = \mathcal{V}_1 \sqcup \mathcal{V}_2 \cdots \sqcup \mathcal{V}_{|\pi|}$. Define:

$$F_{\mathcal{V}_i}(\xi) = \prod_{\substack{j \in \mathcal{V}_i \\ j \neq 1}} q_{j-1}(\xi).$$

Then, we have:

$$\mathbb{E}[q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k}) | \boldsymbol{A}] = \mathbb{E}\left[\prod_{i=1}^{|\pi|} F_{\mathcal{V}_i}(z_{a_{\mathcal{V}_i}}) \middle| \boldsymbol{A}\right].$$

In order to compute the conditional expectation we observe that conditional on $\boldsymbol{A}$, $\boldsymbol{z}$ is a zero mean Gaussian vector with covariance:

$$\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^{\mathsf{T}}|\boldsymbol{A}] = \frac{1}{\kappa}\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} = \frac{1}{\kappa}\boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^{\mathsf{T}} = \boldsymbol{I} + \frac{\boldsymbol{\Psi}}{\kappa}.$$

Note that since $a_{\mathcal{V}_i} \neq a_{\mathcal{V}_j}$ for $i \neq j$, we have as a consequence of (4.15), $\{z_{a_{\mathcal{V}_i}}\}_{i=1}^{|\pi|}$ are weakly correlated Gaussians. Hence we expect,

$$\mathbb{E}[q_1(z_{a_2})\cdots q_{k-1}(z_{a_k})|\boldsymbol{A}] = \prod_{i=1}^{|\pi|} \mathbb{E}_{Z\sim\mathcal{N}(0,1)} F_{\mathcal{V}_i}(Z) + \text{ A small error term,}$$

where the error term is a term that goes to zero as $m \rightarrow \infty$. Mehler's formula given in the proposition below provides an explicit formula for the error term. Observe that in (4.14):

1. the sum over $\pi \in \mathcal{P}([k])$ cannot cause the error terms to add up since $|\mathcal{P}([k])|$ is a constant depending on k but independent of m.

2. On the other hand, the sum over $\boldsymbol{a} \in \mathcal{C}(\pi)$ can cause the errors to add up since:

$$|\mathcal{C}(\pi)| = m \cdot (m-1)\cdots(m-|\pi|+1).$$

It is not obvious right away how accurately the error must be estimated, but it turns out that for the proof of Proposition 7 it suffices to estimate the order of magnitude of the error term. For the proof of Proposition 8 we need to be more accurate and the leading order term in the error needs to be tracked precisely.

Before we state Mehler's formula we recall some preliminaries regarding Fourier analysis on the Gaussian space. Let $Z \sim \mathcal{N}(0,1)$. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be such that $\mathbb{E}f^2(Z) < \infty$, i.e. $f \in L^2(\mathcal{N}(0,1))$. The Hermite polynomials $\{H_j : j \in \mathbb{N}_0\}$ form an orthogonal polynomial basis for $L^2(\mathcal{N}(0,1))$.

The polynomial H_j is a degree j polynomial. They satisfy the orthogonality property:

$$\mathbb{E} H_i(Z) H_j(Z) = i! \cdot \delta_{ij}.$$

The first few Hermite polynomials are given by:

$$H_0(z) = 1, \ H_1(z) = z, \ H_2(z) = z^2 - 1.$$

Proposition 9 ([129, 130]). *Consider a k dimensional Gaussian vector $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, such that $\Sigma_{ii} = 1$ for all $i \in [k]$. Let $f_1, f_2, \dots, f_k : \mathbb{R} \to \mathbb{R}$ be k arbitrary functions whose absolute value can be upper bounded by a polynomial. Then,*

$$\left| \mathbb{E}\left[\prod_{i=1}^{k} f_i(z_i) \right] - \sum_{\substack{\boldsymbol{w} \in \mathcal{G}(k) \\ \|\boldsymbol{w}\| \leq t}} \left(\prod_{i=1}^{k} \hat{f}_i(\mathsf{d}_i(\boldsymbol{w})) \right) \cdot \frac{\boldsymbol{\Sigma}^{\boldsymbol{w}}}{\boldsymbol{w}!} \right| \leq C \left(1 + \frac{1}{\lambda_{\min}^{4t+4}(\boldsymbol{\Sigma})} \right) \left(\max_{i \neq j} |\Sigma_{ij}| \right)^{t+1},$$

where:

1. *$\mathcal{G}(k)$ denotes the set of undirected weighted graphs with non-negative integer weights on k nodes with no self loops.*

2. *An element $\boldsymbol{w} \in \mathcal{G}(k)$ is represented by a $k \times k$ symmetric matrix $\boldsymbol{w}$ with $w_{ij} = w_{ji} \in \mathbb{N} \cup \{0\}$, and $w_{ii} = 0$.*

3. *$\mathsf{d}_i(\boldsymbol{w})$ denotes the degree of node i: $\mathsf{d}_i(\boldsymbol{w}) = \sum_{j=1}^{k} w_{ij}$.*

4. *$\|\boldsymbol{w}\|$ denotes the total weight of the graph defined as:*

$$\|\boldsymbol{w}\| \overset{def}{=} \sum_{i<j} w_{ij} = \frac{1}{2} \sum_{i=1}^{k} \mathsf{d}_i(\boldsymbol{w}).$$

5. *The coefficients $\hat{f}_i(j)$ are defined as: $\hat{f}_i(j) = \mathbb{E} f_i(Z) H_j(Z)$ where $Z \sim \mathcal{N}(0, 1)$.*

6. $\boldsymbol{\Sigma}^{\boldsymbol{w}}, \boldsymbol{w}!$ *denote the entry-wise powering and factorial:*

$$\boldsymbol{\Sigma}^{\boldsymbol{w}} = \prod_{i<j} \Sigma_{ij}^{w_{ij}}, \ \boldsymbol{w}! = \prod_{i<j} w_{ij}!$$

7. $C = C_{t,k,f_{1:k}}$ *is a finite constant depending only on the* t, k, *and the functions* $f_{1:k}$ *but is independent of* $\boldsymbol{\Sigma}$.

This result is essentially due to [129] in the case $k = 2$, and the result for general k was obtained by [130]. Actually the results of these authors show that the pdf of $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ denoted by $\psi(\boldsymbol{z}; \boldsymbol{\Sigma})$ has the following Taylor expansion around $\boldsymbol{\Sigma} = \boldsymbol{I}_k$:

$$\psi(\boldsymbol{z}; \boldsymbol{\Sigma}) = \psi(\boldsymbol{z}; \boldsymbol{I}_k) \cdot \left(\sum_{\boldsymbol{w} \in \mathcal{G}(k)} \frac{\boldsymbol{\Sigma}^{\boldsymbol{w}}}{\boldsymbol{w}!} \cdot \prod_{i=1}^{k} H_{\mathsf{d}_i(\boldsymbol{w})}(z_i) \right).$$

In Appendix 4.10.5 of the supplementary materials we check that this Taylor's expansion can be integrated, and estimate the truncation error to obtain Proposition 9.

At this point, we have introduced all the tools used in the proof of Proposition 7 and we refer the reader to Section 4.7 for the proof of Proposition 7.

4.6.4 Central Limit Theorem

We introduce the following definition.

Definition 8 (Matrix Moment)**.** *Let* $\boldsymbol{M}$ *be a symmetric matrix. Given:*

1. *A partition* $\pi \in \mathcal{P}([k])$ *with blocks* $\pi = \{\mathcal{V}_1, \mathcal{V}_2, \cdots, \mathcal{V}_{|\pi|}\}$.

2. *A* $k \times k$ *symmetric weight matrix* $\boldsymbol{w} \in \mathcal{G}(k)$ *with non-negative valued entries and* $w_{ii} = 0 \ \forall \ i \in [k]$.

3. *A vector* $\boldsymbol{a} \in \mathcal{C}(\pi)$.

Define the $(\boldsymbol{w}, \pi, \boldsymbol{a})$ - matrix moment of the matrix $\boldsymbol{M}$ as:

$$\mathcal{M}(\boldsymbol{M}, \boldsymbol{w}, \pi, \boldsymbol{a}) \overset{def}{=} \prod_{i,j \in [k], i<j} M_{a_i,a_j}^{w_{ij}}.$$

By defining:

$$W_{st}(\boldsymbol{w}, \pi) \overset{def}{=} \sum_{\substack{i,j \in [k], i<j \\ \{\pi(i), \pi(j)\} = \{\mathcal{V}_s, \mathcal{V}_t\}}} w_{ij},$$

we can write $\mathcal{M}(\boldsymbol{M}, \boldsymbol{w}, \pi, \boldsymbol{a})$ in the form:

$$\mathcal{M}(\boldsymbol{M}, \boldsymbol{w}, \pi, \boldsymbol{a}) = \prod_{\substack{s,t \in [|\pi|] \\ s \leq t}} M_{a_{\mathcal{V}_s}, a_{\mathcal{V}_t}}^{W_{st}(\boldsymbol{w}, \boldsymbol{\pi})}.$$

Remark 20 (Graph Interpretation)**.** *It is often useful to interpret the tuple $(\boldsymbol{w}, \pi, \boldsymbol{a})$ in terms of graphs:*

1. *$\boldsymbol{w}$ represents the adjacency matrix of an undirected weighted graph on the vertex set $[k]$ with no self-edges ($w_{ii} = 0$). We say an edge exists between nodes $i, j \in [k]$ if $w_{ij} \geq 1$ and the weight of the edge is given by w_{ij}.*

2. *The partition π of the vertex set $[k]$ represents a community structure on the graph. Two vertices $i, j \in [k]$ are in the same community iff $\pi(i) = \pi(j)$.*

3. *$\boldsymbol{a}$ represents a labelling of the vertices $[k]$ with labels in the set $[m]$ which respects the community structure.*

4. *The weights $W_{st}(\boldsymbol{w}, \pi)$ simply denote the total weight of edges between communities s, t.*

The rationale for introducing this definition is as follows: When we use Mehler's formula to compute $\mathbb{E}[q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k}) | \boldsymbol{A}]$ and $\mathbb{E}[z_{a_1} q_1(z_{a_2}) \cdots q_{k-1}(z_{a_k}) z_{a_{k+1}} | \boldsymbol{A}]$, and substitute the resulting expression in (4.14), it expresses:

$$\frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})}{m}, \ \frac{\mathbb{E}\langle z, \mathcal{A}z\rangle}{m},$$

in terms of the matrix moments $\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$.

For the proof of Proposition 7 it suffices to upper bound $|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})|$. We do so in the following lemma.

Lemma 19. *Consider an arbitrary matrix moment $\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ of $\boldsymbol{\Psi}$. There exists a universal constant C (independent of $m, \boldsymbol{a}, \pi, \boldsymbol{w}$) such that,*

$$\mathbb{E}|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})| \leq \left(\sqrt{\frac{C\|\boldsymbol{w}\| \log^2(m)}{m}}\right)^{\|\boldsymbol{w}\|},$$

for both the sub-sampled Haar and the sub-sampled Hadamard sensing model.

The claim of the lemma is not surprising in light of (4.15). The complete proof follows from the concentration inequality in Lemma 18, which can be found in Appendix 4.10.3 of the supplementary materials.

On the other hand, to prove Proposition 8 we need a more refined analysis and we need to estimate the leading order term in $\mathbb{E}\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$. In order to do so, we first consider any fixed entry of $\sqrt{m}\boldsymbol{\Psi}$:

$$\sqrt{m}\Psi_{ab} = \sqrt{m}(\boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^{\mathsf{T}})_{ab} = \sum_{i=1}^{m} \sqrt{m} \cdot u_{ai} \cdot u_{bi}(B_{ii} - \kappa).$$

Observe that:

1. $B_{ii} - \kappa$ are centered and weakly dependent.

2. $\sqrt{m}u_{ai}u_{bi} = O(m^{-\frac{1}{2}})$ under both the sub-sampled Haar model and the sub-sampled Hadamard model.

Consequently, we expect $\sqrt{m}\Psi_{ab}$ to converge to a Gaussian random variable and hence, we expect that:

$$\mathbb{E}\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$$

to converge to a suitable Gaussian moment. In order to show that the normalized quadratic form $\mathbb{E}\langle z, \mathcal{A}z\rangle/m$ converges to the same limit under both the sensing models, we need to understand what is the limiting value of $\mathbb{E}\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ under both the models. Understanding this uses the following simple but important property of Hadamard matrices.

Lemma 20. *For any $i, j \in [m]$, we have:*

$$\sqrt{m}\boldsymbol{h}_i \odot \boldsymbol{h}_j = \boldsymbol{h}_{i \oplus j},$$

where $\odot$ denotes the entry-wise multiplication of vectors, and $i \oplus j \in [m]$ denotes the result of the following computation:

Step 1: *Compute $\boldsymbol{i}, \boldsymbol{j} \in \{0,1\}^m$ which are the binary representations of $(i-1)$ and $(j-1)$ respectively.*

Step 2: *Compute $\boldsymbol{i} + \boldsymbol{j}$ by adding $\boldsymbol{i}, \boldsymbol{j}$ bit-wise (modulo 2).*

Step 3: *Compute the number in $[0 : m-1]$ whose binary representation is given by $\boldsymbol{i} + \boldsymbol{j}$.*

Step 4: *Add one to the number obtained in Step 3 to obtain $i \oplus j \in [m]$.*

Proof. Recall by the definition of the Hadamard matrix, we have,

$$h_{ik} = \frac{1}{\sqrt{m}}(-1)^{\langle \boldsymbol{i}, \boldsymbol{k}\rangle}, \ h_{jk} = \frac{1}{\sqrt{m}}(-1)^{\langle \boldsymbol{j}, \boldsymbol{k}\rangle}.$$

Hence,

$$\sqrt{m}(\boldsymbol{h}_i \odot \boldsymbol{h}_j)_k = \frac{(-1)^{\langle \boldsymbol{i}+\boldsymbol{j}, \boldsymbol{k} \rangle}}{\sqrt{m}} = (\boldsymbol{h}_{i \oplus j})_k,$$

as claimed. □

Due to the structure in Hadamard matrices, $\mathbb{E}\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ might not always converge to the same limit under the subsampled Haar and the Hadamard models. There are two kinds of exceptions:

Exception 1: Note that for the subsampled Hadamard Model,

$$\sqrt{m}\Psi_{aa} = \sqrt{m} \sum_{i=1}^{m} \overline{B}_{ii} |h_{ai}|^2 = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \overline{B}_{ii} = 0.$$

In contrast, under the subsampled Haar model, it can be shown that $\sqrt{m}\Psi_{aa}$ converges to a non-degenerate Gaussian. These exceptions are ruled out by requiring the weight matrix $\boldsymbol{w}$ to be dissassortative with respect to π (See definition below).

Exception 2: Define $\overline{\boldsymbol{b}} \in \mathbb{R}^m$ to be the vector formed by the diagonal entries of $\overline{\boldsymbol{B}}$. Observe that for the subsampled Hadamard model:

$$\sqrt{m}\Psi_{ab} = \langle \overline{\boldsymbol{b}}, \sqrt{m}\boldsymbol{h}_a \odot \boldsymbol{h}_b \rangle = \langle \overline{\boldsymbol{b}}, \boldsymbol{h}_{a \oplus b} \rangle.$$

Consequently, if two distinct pairs (a_1, b_1) and (a_2, b_2) are such that $a_1 \oplus b_1 = a_2 \oplus b_2$, then $\sqrt{m}\Psi_{a_1,b_1}$ and $\sqrt{m}\Psi_{a_2,b_2}$ are perfectly correlated in the subsampled Hadamard model. In contrast, unless $(a_1, b_1) = (a_2, b_2)$, it can be shown they are asymptotically uncorrelated in the subsampled Haar model. This exception is ruled out by requiring the labelling $\boldsymbol{a}$ to be conflict free with respect to $(\boldsymbol{w}, \pi)$ (defined below).

Definition 9 (Disassortative Graphs)**.** *We say the weight matrix* $\boldsymbol{w}$ *is disassortative with respect to the partition* π *if:* $\forall\ i, j \in [k],\ i < j$ *such that* $\pi(i) = \pi(j)$*, we have* $w_{ij} = 0$*. This is equivalent to*

$W_{ss}(\boldsymbol{w}, \pi) = 0$ for all $s \in [|\pi|]$. In terms of the graph interpretation, this means that there are no intra-community edges in the graph. For any $\pi \in \mathcal{P}([k])$,we denote the set of all weight matrices dissortive with respect to π by $\mathcal{G}_{\mathsf{DA}}(\pi)$:

$$\mathcal{G}_{\mathsf{DA}}(\pi) \overset{def}{=} \{\boldsymbol{w} \in \mathcal{G}(k) : W_{ss}(\boldsymbol{w}, \pi) = 0 \ \forall \ s \ \in [|\pi|]\}.$$

Definition 10 (Conflict Freeness)**.** *Let $\pi \in \mathcal{P}([k])$ be a partition and let $\boldsymbol{w} \in \mathcal{G}_{\mathsf{DA}}(\pi)$ be a weight matrix disassortative with respect to π. Let $s_1 < t_1$ and $s_2 < t_2$ be distinct pairs of communities: $s_1, s_2, t_1, t_2 \in [|\pi|]$, $(s_1, t_1) \neq (s_2, t_2)$. We say a labelling $\boldsymbol{a} \in \mathcal{C}(\pi)$ has a conflict between distinct community pairs (s_1, t_1) and (s_2, t_2) if:*

1. $W_{s_1,t_1}(\boldsymbol{w}, \pi) \geq 1, \ W_{s_2,t_2}(\boldsymbol{w}, \pi) \geq 1.$

2. $a_{\mathcal{V}_{s_1}} \oplus a_{\mathcal{V}_{t_1}} = a_{\mathcal{V}_{s_2}} \oplus a_{\mathcal{V}_{t_2}}.$

We say a labelling $\boldsymbol{a}$ is conflict-free if it has no conflicting community pairs. The set of all conflict free labellings of $(\boldsymbol{w}, \pi)$ is denoted by $\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}, \pi)$.

The following two propositions show that if Exception 1 and Exception 2 are ruled out, then indeed $\mathbb{E}\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ converges to the same Gaussian moment under both the subsampled Haar and the Hadamard models.

Proposition 10. *Consider the sub-sampled Haar model ($\boldsymbol{\Psi} = \boldsymbol{O}\overline{\boldsymbol{B}}\boldsymbol{O}^{\mathsf{T}}$). Fix a partition $\pi \in \mathcal{P}(k)$ and a weight matrix $\boldsymbol{w} \in \mathcal{G}(k)$. Then, there exist constants $K_1, K_2, K_3 > 0$ depending only on $\|\boldsymbol{w}\|$ (independent of m), such that for any $\boldsymbol{a} \in \mathcal{C}(\pi)$ we have:*

$$\left| \mathbb{E}\ \mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) - \prod_{\substack{s,t \in [|\pi|] \\ s \leq t}} \mathbb{E}\left[Z_{st}^{W_{st}(w,\pi)} \right] \right| \leq \frac{K_1 \log^{K_2}(m)}{m^{\frac{1}{4}}}, \ \forall \ m \geq K_3.$$

In the above display, $Z_{st}, \ s \leq t, \ s, t \ \in \ [|\pi|]$ are independent Gaussian random variables with the *distribution:*

$$Z_{st} \sim \begin{cases} s < t: & \mathcal{N}\left(0, \kappa(1-\kappa)\right) \\ s = t: & \mathcal{N}\left(0, 2\kappa(1-\kappa)\right) \end{cases}.$$

Proposition 11. *Consider the sub-sampled Hadamard model ($\boldsymbol{\Psi} = \boldsymbol{H}\overline{\boldsymbol{B}}\boldsymbol{H}^{\mathsf{T}}$). Fix a partition $\pi \in \mathcal{P}(k)$ and a weight matrix $\boldsymbol{w} \in \mathbb{N}_0^{k \times k}$. Then,*

1. *Suppose that $\boldsymbol{w} \notin \mathcal{G}_{\mathsf{DA}}(\pi)$, then,*

$$\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) = 0.$$

2. *Suppose that $\boldsymbol{w} \in \mathcal{G}_{\mathsf{DA}}(\pi)$. Then, there exist constants $K_1, K_2, K_3 > 0$ depending only on $\|\boldsymbol{w}\|$ (independent of m), such that for any conflict free labelling $\boldsymbol{a} \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}, \pi)$, we have:*

$$\left| \mathbb{E}\, \mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) - \prod_{\substack{s,t \in [|\pi|] \\ s<t}} \mathbb{E}\left[Z_\kappa^{W_{st}(\boldsymbol{w},\pi)} \right] \right| \leq \frac{K_1 \log^{K_2}(m)}{m^{\frac{1}{4}}}, \ \forall\, m \geq K_3.$$

In the above display, $Z_\kappa \sim \mathcal{N}\left(0, \kappa(1-\kappa)\right)$.

The proof of these Propositions can be found in Appendix 4.10.3 in the supplementary materials. The proofs use a coupling argument to replace the weakly dependent diagonal matrix $\overline{\boldsymbol{B}}$ with a i.i.d. diagonal entries (as in the proof of Lemma 18) along with a classical Berry Eseen inequality due to [132].

Finally, in order to finish the proof of Proposition 8 regarding the universality of the normalized quadratic form we need to argue the number exceptional labellings under which $\mathbb{E}\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ doesn't converge to the same Gaussian moment under the sub-sampled Hadamard and Haar models are an asymptotically negligible fraction of the total number of labellings.

Lemma 21. *Let $\pi \in \mathcal{P}([k])$ be a partition and $\boldsymbol{w} \in \mathcal{G}_{\mathsf{DA}}(\pi)$ be a weight matrix disassortative with* *respect to* π. *We have,* $|C(\pi)\backslash\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w},\pi)| \leq |\pi|^4 \cdot m^{|\pi|-1}$, *and*

$$\lim_{m\to\infty} \frac{\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w},\pi)}{m^{|\pi|}} = 1.$$

Proof. Let $(s_1, t_1) \neq (s_2, t_2)$ be two distinct community pairs such that:

$$W_{s_1,t_1}(\boldsymbol{w},\pi) \geq 1,\ W_{s_2,t_2}(\boldsymbol{w},\pi) \geq 1.$$

Let $\mathcal{L}_{(\mathsf{s}_1,\mathsf{t}_1;\mathsf{s}_2,\mathsf{t}_2)}(\boldsymbol{w},\pi)$ denote the set of all labellings $\boldsymbol{a} \in C(\pi)$ that have a conflict between distinct community pairs (s_1, t_1) and (s_2, t_2):

$$\mathcal{L}_{(\mathsf{s}_1,\mathsf{t}_1;\mathsf{s}_2,\mathsf{t}_2)}(\boldsymbol{w},\pi) \overset{\text{def}}{=} \{\boldsymbol{a} \in C(\pi) : a_{\mathcal{V}_{s_1}} \oplus a_{\mathcal{V}_{t_1}} = a_{\mathcal{V}_{s_2}} \oplus a_{\mathcal{V}_{t_2}}\}.$$

Then, we note that

$$C(\pi)\backslash\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w},\pi) = \bigcup_{s_1,t_1,s_2,t_2} \mathcal{L}_{(\mathsf{s}_1,\mathsf{t}_1;\mathsf{s}_2,\mathsf{t}_2)}(\boldsymbol{w},\pi),$$

where the union ranges over s_1, t_1, s_2, t_2 such that $1 \leq s_1 < t_1 \leq |\pi|, 1 \leq s_2 < t_2 \leq |\pi|$ and $(s_1, t_1) \neq (s_2, t_2)$ and $W_{s_1,t_1}(\boldsymbol{w},\pi) \geq 1, W_{s_2,t_2}(\boldsymbol{w},\pi) \geq 1$. Next, we bound $|\mathcal{L}_{(\mathsf{s}_1,\mathsf{t}_1;\mathsf{s}_2,\mathsf{t}_2)}(\boldsymbol{w},\pi)|$. Since we know that $(s_1, t_1) \neq (s_2, t_2)$ and $s_1 < t_1$ and $s_2 < t_2$ out of the 4 indices s_1, t_1, s_2, t_2, there must be one index which is different from all the others. Let us assume that this index is t_2 (the remaining cases are analogous). To count $|\mathcal{L}_{(\mathsf{s}_1,\mathsf{t}_1;\mathsf{s}_2,\mathsf{t}_2)}(\boldsymbol{w},\pi)|$ we assign labels to all blocks of π except t_2. The number of ways of doing so is at most $m^{|\pi|-1}$. After we do so, we note that $a_{\mathcal{V}_{t_2}}$ is uniquely determined by the constraint:

$$a_{\mathcal{V}_{s_1}} \oplus a_{\mathcal{V}_{t_1}} = a_{\mathcal{V}_{s_2}} \oplus a_{\mathcal{V}_{t_2}}.$$

Hence, $|\mathcal{L}_{(\mathsf{s}_1,\mathsf{t}_1;\mathsf{s}_2,\mathsf{t}_2)}(\boldsymbol{w},\pi)| \leq m^{|\pi|-1}$. Therefore,

$$|C(\pi)\backslash\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w},\pi)| = \sum_{s_1,t_1,s_2,t_2} |\mathcal{L}_{(\mathsf{s}_1,\mathsf{t}_1;\mathsf{s}_2,\mathsf{t}_2)}(\boldsymbol{w},\pi)| \leq |\pi|^4 m^{|\pi|-1}.$$

Finally, we note that,

$$|C(\pi)| - |C(\pi)\backslash\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w},\pi)| = |\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w},\pi)| \leq |C(\pi)|.$$

$|C(\pi)|$ is given by:

$$|C(\pi)| = m(m-1)\cdots(m-|\pi|+1) = m^{|\pi|}\cdot(1+o_m(1)).$$

Combining this with the already obtained upper bound $|C(\pi)\backslash\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w},\pi)| \leq |\pi|^4 \cdot m^{|\pi|-1}$, we obtain the second claim of the lemma. □

We now have all the tools required to finish the proof of Proposition 8 and we refer the reader to Section 4.8 for the proof of this result.

4.7 Proof of Proposition 7

In this Section we prove Proposition 7.

Let us consider a fixed alternating product $\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})$ as given in Definition 7. As a consequence of Lemma 17 we can assume that all the polynomials $p_i(\xi) = \xi$. We begin by stating a few intermediate lemmas which will be used to prove Proposition 7.

Lemma 22 (A high probability event). *Let $\boldsymbol{U}$ denote the $m \times m$ orthogonal matrix used to generate the sensing matrix . Define the event:*

$$\mathcal{E} = \left\{ \max_{i\neq j} |(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}|)_{ij} \leq \sqrt{32\cdot m\cdot\|\boldsymbol{U}\|_\infty^4\cdot\log(m)}, \right. \\ \left. \max_{i\in[m]} |(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})_{ii} - \kappa| \leq \sqrt{32\cdot m\cdot\|\boldsymbol{U}\|_\infty^4\cdot\log(m)} \right\}. \tag{4.16}$$

Then,

$$\mathbb{P}(\mathcal{E}|\boldsymbol{U}) \geq 1 - 4/m^2.$$

Furthermore, for the subsampled Haar model, when $\boldsymbol{U} = \boldsymbol{O} \sim$ Unif$\left(\mathbb{O}(m)\right)$, we have:

$$\mathbb{P}\left(\left\{\|\boldsymbol{O}\|_\infty \leq \sqrt{\frac{8\log(m)}{m}}\right\} \cap \mathcal{E}\right) \geq 1 - 6/m^2.$$

The above Lemma follows from the concentration result in Lemma 18 and a union bound. Complete details are provided in Appendix 4.10.1 in the supplementary materials.

Lemma 23 (A Continuity Estimate)**.** *Let $\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z})$ be an alternating product of the matrices $\boldsymbol{\Psi}, \mathbf{Z}$ (see Definition 7). Then the map $\mathbf{Z} \mapsto \mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z})/m$ is Lipchitz in Z, i.e. for any two diagonal matrices $\mathbf{Z} = Diag\left(z_1, z_2 \dots, z_m\right)$, $\mathbf{Z}' = Diag\left(z_1', z_2' \dots, z_m'\right)$ we have:*

$$\left|\frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z})}{m} - \frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z}')}{m}\right| \leq \frac{C(\mathcal{A})}{\sqrt{m}} \cdot \|\mathbf{Z} - \mathbf{Z}'\|_{\mathsf{Fr}},$$

where $C(\mathcal{A})$ denotes a constant depending only on the formula for the alternating product $\mathcal{A}$ (independent of m, n).

This lemma follows from a straightforward computation provided in 4.10.1 in the supplementary materals.

Lemma 24 (Analysis of Expectation)**.** *Let the sensing matrix $\mathbf{A}$ be drawn either from the subsampled Haar model or be generated using a deterministic orthogonal matrix $\boldsymbol{U}$ with the property:*

$$\|\boldsymbol{U}\|_\infty \leq \sqrt{\frac{K_1 \log^{K_2}(m)}{m}},$$

for some universal constants $K_1, K_2 \geq 0$, then, we have:

$$\frac{1}{m}\mathbb{E}[\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}))|\boldsymbol{A}] \xrightarrow{P} 0.$$

Lemma 25 (Analysis of Variance)**.** *Let $\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})$ be any alternating product of the matrices $\boldsymbol{\Psi}, \boldsymbol{Z}$. Then,*

$$\mathrm{Var}\left(\frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})}{m}\middle|\boldsymbol{A}\right) \leq \frac{C(\mathcal{A})}{n},$$

where $C(\mathcal{A})$ denotes a constant depending only on the formula for the alternating product $\mathcal{A}$ (independent of m, n).

Proofs of Lemmas 24 and 25 can be found at Section 4.7.1. Before moving forward to the proofs of these lemmas, let us conclude the proof of Proposition 7 assuming Lemmas 24 and 25 are true.

Proof of Proposition 7. We write $\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}))/m$ as:

$$\frac{\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}))}{m} = \mathbb{E}\left[\frac{\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}))}{m}\middle|\boldsymbol{A}\right] + \left(\frac{\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}))}{m} - \mathbb{E}\left[\frac{\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}))}{m}\middle|\boldsymbol{A}\right]\right).$$

We will show each of the two terms on the right hand side converge to zero in probability. Lemma 24 already gives:

$$\mathbb{E}\left[\frac{\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}))}{m}\middle|\boldsymbol{A}\right] \xrightarrow{\mathrm{P}} 0.$$

On the other hand, by Chebychev's Inequality and Lemma 25 we have:

$$\mathbb{P}\left[\left|\frac{\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})) - \mathbb{E}[\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}))|\boldsymbol{A}]}{m}\right| > \epsilon\middle|\boldsymbol{A}\right] \leq \frac{1}{\epsilon^2}\cdot\mathrm{Var}\left(\frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})}{m}\middle|\boldsymbol{A}\right) \leq \frac{C(\mathcal{A})}{n\epsilon^2}.$$

Hence,

$$\mathbb{P}\left[\left|\frac{\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})) - \mathbb{E}[\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z}))|\boldsymbol{A}]}{m}\right| > \epsilon\right] \to 0.$$

This concludes the proof of the proposition. □

4.7.1 Proof of Lemmas 24 and 25

Proof of Lemma 24. Recall the notation regarding partitions introduced in Section 4.6.1. We will organize the proof into various steps.

Step 1: Restricting to a Good Event. We first observe that $\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z}))/m$ is uniformly bounded:

$$\frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})}{m} \leq \|\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})\|_{\mathsf{op}} \leq \prod_{i=1}^{k} \|q_i\|_\infty = C(\mathcal{A}) < \infty,$$

where $\|q_i\|_\infty = \sup_{\xi\in\mathbb{R}} |q_i(\xi)|$, and $C(\mathcal{A})$ denotes a finite constant independent of m, n. Recall the definition of $\mathcal{E}$ in (4.16). If the sensing matrix $\boldsymbol{A}$ was generated by subsampling a deterministic orthogonal matrix $\boldsymbol{U}$ with the property

$$\|\boldsymbol{U}\|_\infty \leq \sqrt{\frac{K_1 \log^{K_2}(m)}{m}},$$

then Lemma 22 gives $\mathbb{P}(\mathcal{E}^c) \leq 4/m^2$. On the other hand, if $\boldsymbol{A}$ was generated by subsampling a uniformly random column orthogonal matrix $\boldsymbol{O}$ then we set $K_1 = 8, K_2 = 1$ and Lemma 22 gives $\mathbb{P}(\mathcal{E}^c) \leq 6/m^2$. Using this event, we decompose $\mathbb{E}[\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})|\boldsymbol{A}]/m$ as:

$$\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})|\boldsymbol{A}]}{m} = \frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})|\boldsymbol{A}]}{m} \cdot \mathbb{I}_{\mathcal{E}} + \frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})|\boldsymbol{A}]}{m} \cdot \mathbb{I}_{\mathcal{E}^c}.$$

Since $\mathbb{P}(\mathcal{E}^c) \to 0$ and $\mathbb{E}[\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})|\boldsymbol{A}]/m < C(\mathcal{A}) < \infty$ is uniformly bounded, we immediately obtain $\mathbb{E}[\mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})|\boldsymbol{A}] \cdot \mathbb{I}_{\mathcal{E}^c}/m \xrightarrow{\mathrm{P}} 0$. Hence, we simply need to show:

$$\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})|\boldsymbol{A}]}{m} \cdot \mathbb{I}_{\mathcal{E}} \xrightarrow{\mathrm{P}} 0.$$

Step 2: Variance Normalization. Recall that $\boldsymbol{Z} = \mathrm{Diag}\,(\boldsymbol{z})\,, \; \boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}/\kappa\right)$. We define the normalized random vector $\tilde{\boldsymbol{z}}$ as:

$$\tilde{z}_i = \frac{z_i}{\sigma_i}, \; \sigma_i^2 = \frac{(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})_{ii}}{\kappa}. \tag{4.17}$$

Note that conditional on $\boldsymbol{A}$, $\tilde{\boldsymbol{z}}$ is a zero mean Gaussian vector with:

$$\mathbb{E}[\tilde{z}_i{}^2|\boldsymbol{A}] = 1, \; \mathbb{E}[\tilde{z}_i\tilde{z}_j|\boldsymbol{A}] = \frac{(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})_{ij}/\kappa}{\sigma_i\sigma_j}.$$

We define the diagonal matrix $\tilde{\boldsymbol{Z}} = \mathrm{Diag}\,(\tilde{\boldsymbol{z}})$. Using the continuity estimate from Lemma 23 we have,

$$\begin{aligned}
\left|\frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})}{m} - \frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}})}{m}\right| &\leq \frac{C(\mathcal{A})}{\sqrt{m}}\|\boldsymbol{z} - \tilde{\boldsymbol{z}}\|_2 \\
&\leq C(\mathcal{A}) \cdot \left(\frac{1}{m}\sum_{i=1}^{m} z_i^2\right)^{\frac{1}{2}} \cdot \left(\max_{i\in[m]}\left|\frac{1}{\sigma_i} - 1\right|\right) \\
&\leq C(\mathcal{A}) \cdot \left(\frac{1}{m}\sum_{i=1}^{m} x_i^2\right)^{\frac{1}{2}} \cdot \left(\max_{i\in[m]}\left|\frac{1}{\sigma_i} - 1\right|\right).
\end{aligned}$$

We observe that $\|\boldsymbol{x}\|^2/m \xrightarrow{\mathrm{P}} \kappa^{-1}$, and on the event $\mathcal{E}$,

$$\max_{i\in[m]}\left|\frac{1}{\sigma_i} - 1\right| \to 0.$$

Hence,

$$\left| \frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})|\boldsymbol{A}]}{m} - \frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}})|\boldsymbol{A}]}{m} \right| \cdot \mathbb{I}_{\mathcal{E}} \overset{\mathrm{P}}{\rightarrow} 0,$$

and hence, to conclude the proof of the lemma we simply need to show:

$$\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}})|\boldsymbol{A}]}{m} \cdot \mathbb{I}_{\mathcal{E}} \overset{\mathrm{P}}{\rightarrow} 0.$$

Step 3: Mehler's Formula. Supposing that alternating product is of the Type 2 form (recall Definition 7):

$$\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}}) = (\boldsymbol{\Psi})q_1(\tilde{\boldsymbol{Z}})(\boldsymbol{\Psi})q_2(\tilde{\boldsymbol{Z}}) \cdots (\boldsymbol{\Psi})q_k(\tilde{\boldsymbol{Z}}).$$

The argument for the other types is very similar and we will sketch it in the end. We expand $\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}})$ as follows:

$$\frac{1}{m}\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}}) = \frac{1}{m} \sum_{a_1, a_2, \dots a_k = 1}^{m} (\boldsymbol{\Psi})_{a_1, a_2} q_1(\tilde{\boldsymbol{Z}})_{a_2, a_2} \cdots (\boldsymbol{\Psi})_{a_k, a_1} q_k(\tilde{\boldsymbol{Z}})_{a_1, a_1}.$$

Next, we observe that:

$$[m]^k = \bigsqcup_{\pi \in \mathcal{P}([k])} C(\pi).$$

Hence we can decompose the above sum as:

$$\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}}) \ |\boldsymbol{A}]}{m} = \sum_{\pi \in \mathcal{P}([k])} \frac{1}{m} \sum_{a \in C(\pi)} (\boldsymbol{\Psi})_{a_1, a_2} \cdots (\boldsymbol{\Psi})_{a_k, a_1} \mathbb{E}[\ q_1(\tilde{z}_{a_2}) \cdots q_k(\tilde{z}_{a_{k+1}}) | \boldsymbol{A}].$$

By the triangle inequality,

$$\left|\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\mathbf{\Psi},\tilde{\mathbf{Z}})\ |\mathbf{A}]}{m}\right| \leq \sum_{\pi\in\mathcal{P}([k])}\frac{1}{m}\sum_{a\in\mathcal{C}(\pi)}|(\mathbf{\Psi})_{a_1,a_2}\cdots(\mathbf{\Psi})_{a_k,a_1}||\mathbb{E}[\ q_1(\tilde{z}_{a_2})\cdots q_k(\tilde{z}_{a_1})|\mathbf{A}]|. \tag{4.18}$$

We first bound $|\mathbb{E}[\ q_1(\tilde{z}_{a_2})q_2(\tilde{z}_{a_3})\cdots q_k(\tilde{z}_{a_1})|\mathbf{A}]|$. Observe that if we denote the blocks of $\pi = \{\mathcal{V}_1, \mathcal{V}_2 \dots \mathcal{V}_{|\pi|}\}$, we can write:

$$\left|\mathbb{E}[\ q_1(\tilde{z}_{a_2})q_2(\tilde{z}_{a_3})\cdots q_k(\tilde{z}_{a_1})|\mathbf{A}]\right| = \left|\mathbb{E}\left[\prod_{i=1}^{|\pi|}\prod_{j\in\mathcal{V}_i} q_{j-1}(\tilde{z}_{a_{\mathcal{V}_i}})\middle|\mathbf{A}\right]\right|.$$

In the above display, we have defined $q_0 \overset{\text{def}}{=} q_k$. Define the functions $\overline{q}_1, \overline{q}_2 \dots \overline{q}_{|\pi|}$ as:

$$\overline{q}_i(\xi) = \prod_{j\in\mathcal{V}_i} q_{j-1}(\xi) - \nu_i,\ \ \nu_i = \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\prod_{j\in\mathcal{V}_i} q_{j-1}(\xi)\right].$$

Hence, we obtain:

$$\begin{aligned}\left|\mathbb{E}[\ q_1(\tilde{z}_{a_2})q_2(\tilde{z}_{a_3})\cdots q_k(\tilde{z}_{a_1})|\mathbf{A}]\right| &= \left|\mathbb{E}\left[\prod_{i=1}^{|\pi|}(\overline{q}_i(z_{a_{\mathcal{V}_i}}) + \nu_i)\middle|\mathbf{A}\right]\right| \\ &\leq \sum_{V\subset[|\pi|]}\left(\prod_{i\notin V}|\nu_i|\right)\cdot\left|\mathbb{E}\left[\prod_{i\in V}\overline{q}_i(\tilde{z}_{a_{\mathcal{V}_i}})\middle|\mathbf{A}\right]\right|. \end{aligned} \tag{4.19}$$

Let $\mathcal{S}(\pi)$ denote the singleton blocks of the partition π: $\mathcal{S}(\pi) = \{i \in [|\pi|] : |\mathcal{V}_i| = 1\}$. Note that for any $i \in \mathcal{S}(\pi)$, $\nu_i = 0$ since the functions q_i satisfy $\mathbb{E}q_i(\xi) = 0$ when $\xi \sim \mathcal{N}(0, 1)$ (Definition 7). Hence,

$$\left|\mathbb{E}[\ q_1(\tilde{z}_{a_2})q_2(\tilde{z}_{a_3})\cdots q_k(\tilde{z}_{a_1})|\mathbf{A}]\right| \leq \sum_{V\subset[|\pi|]:\mathcal{S}(\pi)\subset V}\left(\prod_{i\notin V}|\nu_i|\right)\cdot\left|\mathbb{E}\left[\prod_{i\in V}\overline{q}_i(\tilde{z}_{a_{\mathcal{V}_i}})\middle|\mathbf{A}\right]\right|.$$

Next, we apply Mehler's Formula (Proposition 9) to bound:

$$\left| \mathbb{E}\left[\prod_{i \in V} \overline{q}_i(\tilde{z}_{a_{\mathcal{V}_i}}) \middle| \boldsymbol{A} \right] \right| \mathbb{I}_{\mathcal{E}}.$$

We make the following observations:

1. Recall the distribution of $\tilde{z}$ given in (4.17) and the definition of the event $\mathcal{E}$ in (4.16), we obtain:

   $$\max_{i \neq j} |\mathbb{E}[\tilde{z}_i \tilde{z}_j | \boldsymbol{A}]| \leq \left(\max_{i \neq j} \frac{1}{\kappa \sigma_i \sigma_j} \sqrt{\frac{32 \cdot K_1^2 \cdot \log^{2K_2+1}(m)}{m}} \right).$$

   Note that for large enough m, event $\mathcal{E}$ guarantees $\min_i \sigma_i \geq 1/2$. Hence,

   $$\max_{i \neq j} |\mathbb{E}[\tilde{z}_i \tilde{z}_j | \boldsymbol{A}]| \leq \left(\frac{4}{\kappa} \sqrt{\frac{32 \cdot K_1^2 \cdot \log^{2K_2+1}(m)}{m}} \right).$$

   For any $S \subset [m]$ with $|S| \leq k$, let $\mathbb{E}[\tilde{z}\tilde{z}^{\mathsf{T}} | \boldsymbol{A}]_{S,S}$ be the principal submatrix of the covariance matrix $\mathbb{E}[\tilde{z}\tilde{z}^{\mathsf{T}} | \boldsymbol{A}]$. By Gershgorin's Circle Theorem we have.

   $$\lambda_{\min} \left(\mathbb{E}[\tilde{z}\tilde{z}^{\mathsf{T}} | \boldsymbol{A}]_{S,S} \right) \geq 1 - k \max_{i \neq j} |\mathbb{E}[\tilde{z}_i \tilde{z}_j | \boldsymbol{A}]| \geq \frac{1}{2} \text{ (for } m \text{ large enough).}$$

2. We note that $\overline{q}_i$ satisfy $\mathbb{E}\overline{q}_i(\xi) = 0$ and $\mathbb{E}\xi\overline{q}_i(\xi) = 0$ (since $\overline{q}_i$ are even functions) when $\xi \sim \mathcal{N}(0, 1)$. Hence, the first non-zero term in Mehler's expansion corresponds to $\boldsymbol{w}$ such that:

   $$\mathsf{d}_i(\boldsymbol{w}) \geq 2, \quad \forall\, i \in V,$$

   thus,

   $$\|\boldsymbol{w}\| \geq |V|.$$

Hence, by Mehler's Formula (Proposition 9), we obtain:

$$\left|\mathbb{E}\left[\prod_{i\in V}\overline{q}_i(\tilde{z}_{a_{\mathcal{V}_i}})\middle|\boldsymbol{A}\right]\right|\mathbb{I}_{\mathcal{E}} \leq C\cdot\left(\max_{i\neq j}\mathbb{E}[\tilde{z}_i\tilde{z}_j|\boldsymbol{A}]\right)^{|V|}$$

$$\leq C\cdot\left(\frac{4}{\kappa}\sqrt{\frac{32\cdot K_1^2\cdot\log^{2K_2+1}(m)}{m}}\right)^{|V|},$$

for some finite constant C depending only on k and the functions $q_{1:k}$. Substituting this bound in (4.19) we obtain:

$$\left|\mathbb{E}[\ q_1(\tilde{z}_{a_2})q_2(\tilde{z}_{a_3})\cdots q_k(\tilde{z}_{a_1})|\boldsymbol{A}]\right|\cdot\mathbb{I}_{\mathcal{E}} \leq \sum_{V\subset[|\pi|]}\left(\prod_{i\notin V}|\nu_i|\right)\cdot\left|\mathbb{E}\left[\prod_{i\in V}\overline{q}_i(\tilde{z}_{a_{\mathcal{V}_i}})\middle|\boldsymbol{A}\right]\right|$$

$$\leq C\sum_{V\subset[|\pi|]}\left(\prod_{i\notin V}|\nu_i|\right)\cdot\left(\frac{4}{\kappa}\sqrt{\frac{32\cdot K_1^2\cdot\log^{2K_2+1}(m)}{m}}\right)^{|V|}$$

$$\leq C(\mathcal{A})\cdot\left(\frac{4}{\kappa}\sqrt{\frac{32\cdot K_1^2\cdot\log^{2K_2+1}(m)}{m}}\right)^{|\mathcal{S}(\pi)|}.$$

In the above display, $C(\mathcal{A})$ denotes a finite constant depending only on k and the functions appearing in the definition of $\mathcal{A}$. Substituting this in (4.18):

$$\left|\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\tilde{\boldsymbol{Z}})\ |\boldsymbol{A}]}{m}\right|\mathbb{I}_{\mathcal{E}}$$

$$\leq \sum_{\pi\in\mathcal{P}([k])}\frac{C(\mathcal{A})}{m}\sum_{a\in\mathcal{C}(\pi)}|(\boldsymbol{\Psi})_{a_1,a_2}\cdots(\boldsymbol{\Psi})_{a_k,a_1}|\left(\frac{4}{\kappa}\sqrt{\frac{32\cdot K_1^2\cdot\log^{2K_2+1}(m)}{m}}\right)^{|\mathcal{S}(\pi)|}.$$

Again, recalling the definition of $\mathcal{E}$ in (4.16), we can upper bound $|(\boldsymbol{\Psi})_{a_1,a_2}\cdots(\boldsymbol{\Psi})_{a_k,a_1}|$:

$$\left|\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\tilde{\boldsymbol{Z}})\ |\boldsymbol{A}]}{m}\right|\cdot\mathbb{I}_{\mathcal{E}}\leq\sum_{\pi\in\mathcal{P}([k])}\frac{C(\mathcal{A})}{m}\sum_{a\in C(\pi)}\cdot\left(\sqrt{\frac{\cdot K_1^2\cdot\log^{2K_2+1}(m)}{m}}\right)^{|\mathcal{S}(\pi)|+k}$$

$$=\frac{C(\mathcal{A})}{m}\sum_{\pi\in\mathcal{P}([k])}|C(\pi)|\cdot\left(\sqrt{\frac{\cdot K_1^2\cdot\log^{2K_2+1}(m)}{m}}\right)^{|\mathcal{S}(\pi)|+k}. \quad (4.20)$$

Step 4: Conclusion. Observe that: $|C(\pi)|\leq m^{|\pi|}$. Recall that π has $|\mathcal{S}(\pi)|$ singleton blocks. All remaining blocks of π have at least 2 elements. Hence, we can upper bound $|\pi|$ as follows:

$$|\pi|\leq\frac{k-|\mathcal{S}(\pi)|}{2}+|\mathcal{S}(\pi)|=\frac{k+|\mathcal{S}(\pi)|}{2}.$$

Substituting this in (4.20) along with the trivial bounds $|\mathcal{S}(\pi)|\leq k,\ |\mathcal{P}([k])\leq k^k$, we obtain:

$$\left|\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi},\tilde{\boldsymbol{Z}})\ |\boldsymbol{A}]}{m}\right|\cdot\mathbb{I}_{\mathcal{E}}\leq\frac{C(\mathcal{A})\cdot k^k\cdot(K_1^2\log^{2K_2+1}(m))^k}{m}\to 0,$$

as desired.

Step 5: Other Cases. Recall that we had assumed that the alternating product was of Type 2:

$$\mathcal{A}(\boldsymbol{\Psi},\tilde{\boldsymbol{Z}})=(\boldsymbol{\Psi})q_1(\tilde{\boldsymbol{Z}})(\boldsymbol{\Psi})q_2(\tilde{\boldsymbol{Z}})\cdots(\boldsymbol{\Psi})q_k(\tilde{\boldsymbol{Z}}).$$

The analysis for the other types is analogous, and we briefly sketch these cases:

Type 1: $\mathcal{A}(\boldsymbol{\Psi},\tilde{\boldsymbol{Z}})=(\boldsymbol{\Psi})q_1(\tilde{\boldsymbol{Z}})(\boldsymbol{\Psi})q_2(\tilde{\boldsymbol{Z}})\cdots(\boldsymbol{\Psi})q_k(\tilde{\boldsymbol{Z}})(\boldsymbol{\Psi})$**.** In this case, the normalized trace is expanded as:

$$\begin{aligned}\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}})\ |\boldsymbol{A}]}{m} &= \frac{1}{m}\sum_{a_0,a_1,\dots a_k=1}^{m} \mathbb{E}[(\boldsymbol{\Psi})_{a_0,a_1} q_1(\tilde{\boldsymbol{Z}})_{a_1,a_1}\cdots q_k(\tilde{\boldsymbol{Z}})_{a_k,a_k}(\boldsymbol{\Psi})_{a_k,a_0}|\boldsymbol{A}] \\ &= \frac{1}{m}\sum_{a_0=1}^{m}\sum_{\pi\in\mathcal{P}([k])}\sum_{a\in\mathcal{C}(\pi)}(\boldsymbol{\Psi})_{a_0,a_1}(\boldsymbol{\Psi})_{a_1,a_2}\cdots(\boldsymbol{\Psi})_{a_k,a_0}\mathbb{E}[q_1(\tilde{z}_{a_1})\cdots q_k(\tilde{z}_{a_k})|\boldsymbol{A}].\end{aligned}$$

As before, we can argue on the event $\mathcal{E}$, for any $a_{0:k}$:

$$\begin{aligned}|\mathbb{E}[q_1(\tilde{z}_{a_1})\cdots q_k(\tilde{z}_{a_k})|\boldsymbol{A}]| &\leq O\left(\left(\frac{\mathrm{polylog}(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|}{2}}\right), \\ |(\boldsymbol{\Psi})_{a_0,a_1}(\boldsymbol{\Psi})_{a_1,a_2}\cdots(\boldsymbol{\Psi})_{a_k,a_0}| &\leq O\left(\left(\frac{\mathrm{polylog}(m)}{m}\right)^{\frac{k+1}{2}}\right), \\ |\mathcal{C}(\pi)| &\leq m^{\frac{k+|\mathcal{S}(\pi)|}{2}}, \\ |\mathcal{P}([k])| &\leq k^k.\end{aligned}$$

This gives us:

$$\begin{aligned}\left|\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \tilde{\boldsymbol{Z}})\ |\boldsymbol{A}]}{m}\right|\mathbb{I}_{\mathcal{E}} &\leq \frac{1}{m}\cdot\overbrace{m}^{\text{choices for } a_0}\cdot\overbrace{|\mathcal{P}([k])|}^{\text{choices for } \pi}\cdot\overbrace{|\mathcal{C}(k)|}^{\text{choices for } a_{1:k}}\cdot O\left(\frac{\mathrm{polylog}(m)}{m^{\frac{k+|\mathcal{S}(\pi)|+1}{2}}}\right) \\ &= O\left(\frac{\mathrm{polylog}(m)}{\sqrt{m}}\right) \rightarrow 0.\end{aligned}$$

Type 3: $\mathcal{A} = q_0(\boldsymbol{Z})(\boldsymbol{\Psi})q_1(\boldsymbol{Z})\cdots(\boldsymbol{\Psi})q_k(\boldsymbol{Z})$. This case can be reduced to Type 1 and Type 2. Define $\tilde{q}_k(\xi) = q_0(\xi)q_k(\xi) - \nu,\ \nu = \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\ q_0(\xi)q_k(\xi)$. Then:

$$\begin{aligned}\frac{\mathbb{E}[\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})|\boldsymbol{A}]}{m} &= \frac{\mathbb{E}[\mathsf{Tr}(q_0(\boldsymbol{Z})(\boldsymbol{\Psi})q_1(\boldsymbol{Z})\cdots(\boldsymbol{\Psi})q_k(\boldsymbol{Z}))|\boldsymbol{A}]}{m} \\ &= \frac{\mathbb{E}[\mathsf{Tr}((\boldsymbol{\Psi})q_1(\boldsymbol{Z})\cdots(\boldsymbol{\Psi})q_k(\boldsymbol{Z})q_0(\boldsymbol{Z}))|\boldsymbol{A}]}{m} \\ &= \underbrace{\frac{\mathbb{E}[\mathsf{Tr}((\boldsymbol{\Psi})q_1(\boldsymbol{Z})\cdots(\boldsymbol{\Psi})\tilde{q}_k(\boldsymbol{Z}))|\boldsymbol{A}]}{m}}_{\text{Type 2}} + \nu\underbrace{\frac{\mathbb{E}[\mathsf{Tr}((\boldsymbol{\Psi})q_1(\boldsymbol{Z})\cdots(\boldsymbol{\Psi}))|\boldsymbol{A}]}{m}}_{\text{Type 1}}.\end{aligned}$$

Type 4: $\mathcal{A}(\mathbf{\Psi}, \boldsymbol{Z}) = q_1(\boldsymbol{Z})(\mathbf{\Psi})q_2(\boldsymbol{Z})(\mathbf{\Psi}) \cdots q_k(\boldsymbol{Z})(\mathbf{\Psi})$**.** This case is exactly the same as Type 2, and exactly the same bounds hold.

This concludes the proof of Lemma 24. □

Proof of Lemma 25. We observe that since $\mathbf{\Psi} = \boldsymbol{A}\boldsymbol{A}^\mathsf{T} - \kappa \boldsymbol{I}_m$, conditioning on $\boldsymbol{A}$ fixes $\mathbf{\Psi}$. Hence, the only source of randomness in $\mathcal{A}(\mathbf{\Psi}, \boldsymbol{Z})$ is $\boldsymbol{Z} = \text{Diag}\,(z)\,, z = \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \sim \mathcal{N}\left(0, 1/\kappa\right)$. Define the map $f(\boldsymbol{x}) \overset{\text{def}}{=} \mathsf{Tr}(\mathcal{A}(\mathbf{\Psi}, \text{Diag}\,(\boldsymbol{A}\boldsymbol{x}))/m$. By Lemma 23, we have:

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq \frac{C(\mathcal{A})}{\sqrt{m}} \cdot \|\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}')\|_2 \leq \frac{C(\mathcal{A})\|\boldsymbol{A}\|_{\mathsf{op}}}{\sqrt{m}} \cdot \|\boldsymbol{x} - \boldsymbol{x}'\|_2 = \frac{C(\mathcal{A})}{\sqrt{m}} \cdot \|\boldsymbol{x} - \boldsymbol{x}'\|_2.$$

Hence, f is $C(\mathcal{A})/\sqrt{n}$-Lipchitz. The claim of Lemma follows from the Gaussian Poincare Inequality (see Fact 3). □

4.8 Proof of Proposition 8

In this section, we provide a proof of Proposition 8. The proof follows from the following three results.

Lemma 26 (Continuity Estimates)**.** *We have:*

$$\left|\frac{z^\mathsf{T}\mathcal{A}(\boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^\mathsf{T}, \mathit{Diag}\,(z))z}{m} - \frac{\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^\mathsf{T}, \mathit{Diag}\,(\widetilde{z}))\widetilde{z}}{m}\right| \\ \leq \frac{C(\mathcal{A})}{m} \cdot \left(\|z\|_2^2 \cdot \|z - \widetilde{z}\|_\infty + \|z - \widetilde{z}\|_2 \cdot (\|z\|_2 + \|\widetilde{z}\|_2)\right),$$

where $C(\mathcal{A})$ depends only on k, the $\|\|_\infty$-norms, and Lipchitz constants of the functions appearing in $\mathcal{A}$.

We have relegated the proof of the above continuity estimate to Appendix 4.10.4 in the supplementary materials.

Proposition 12 (Universality of the first moment of the quadratic form)**.** *For both the subsampled Haar sensing model and the subsampled Hadamard sensing model, we have:*

$$\lim_{m\to\infty} \frac{\mathbb{E} z^{\mathsf{T}} \mathcal{A} z}{m} = (1-\kappa)^k \cdot \left(\prod_i \hat{q}_i(2) \right) \cdot \left(\prod_i (p_i(1-\kappa) - p_i(-\kappa)) \right),$$

where the index i in the product ranges over all the p_i, q_i functions appearing in $\mathcal{A}$. In the above display:

$$\hat{q}_i(2) = \mathbb{E} q_i(\xi) H_2(\xi), \ \xi \sim \mathcal{N}(0,1), \tag{4.21}$$

where $H_2(\xi) = \xi^2 - 1$ is the degree 2 Hermite polynomial.

Proposition 13 (Universality of the second moment of the quadratic form)**.** *For both the subsampled Haar sensing model and the subsampled Hadamard sensing model we have:*

$$\lim_{m\to\infty} \frac{\mathbb{E}(z^{\mathsf{T}} \mathcal{A} z)^2}{m^2} = (1-\kappa)^{2k} \cdot \left(\prod_i \hat{q}_i^2(2) \right) \cdot \left(\prod_i (p_i(1-\kappa) - p_i(-\kappa))^2 \right).$$

In the above expression, $\hat{q}_i(2)$ are as defined in (4.21).

We now provide a proof of Proposition 8 using the above results.

Proof of Proposition 8. Note that Propositions 12, 13 together imply that,

$$\mathrm{Var}\left(\frac{z^{\mathsf{T}} \mathcal{A} z}{m} \right) \to 0,$$

for both the sensing models. Hence, by Chebychev's inequality and Proposition 12, we have, for both the sensing models,

$$\text{p-lim} \frac{z^{\mathsf{T}} \mathcal{A} z}{m} = (1-\kappa)^k \cdot \left(\prod_i \hat{q}_i(2) \right) \cdot \left(\prod_i (p_i(1-\kappa) - p_i(-\kappa)) \right).$$

This proves the claim of Proposition 8. □

The remainder of the section is dedicated to the proof of Proposition 12. The proof of Proposition 13 is very similar and can be found in Appendix 4.10.2 in the supplementary materials.

4.8.1 Proof of Proposition 12

We provide a proof of Proposition 12 assuming that alternating form is of Type 1.

$$\mathcal{A}(\mathbf{\Psi}, \mathbf{Z}) = p_1(\mathbf{\Psi})q_1(\mathbf{Z})p_2(\mathbf{\Psi}) \cdots q_{k-1}(\mathbf{Z})p_k(\mathbf{\Psi}).$$

We will outline how to handle the other types at the end of the proof (see Remark 21). Furthermore, in light of Lemma 17 we can further assume that all polynomials $p_i(\psi) = \psi$. Hence, we assume that $\mathcal{A}$ is of the form:

$$\mathcal{A}(\mathbf{\Psi}, \mathbf{Z}) = \mathbf{\Psi} q_1(\mathbf{Z})\mathbf{\Psi} \cdots q_{k-1}(\mathbf{Z})\mathbf{\Psi}.$$

The proof of Proposition 12 consists of various steps which will be organized as separate lemmas. We begin by recall that

$$z \sim \mathcal{N}\left(0, \frac{\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}}{\kappa}\right).$$

Define the event:

$$\mathcal{E} = \left\{ \max_{i \neq j} |(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}|)_{ij} \leq \sqrt{\frac{2048 \cdot \log^3(m)}{m}},\ \max_{i \in [m]} |(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})_{ii} - \kappa| \leq \sqrt{\frac{2048 \cdot \log^3(m)}{m}} \right\}. \tag{4.22}$$

By Lemma 22, we know that $\mathbb{P}(\mathcal{E}^c) \to 0$ for both the subsampled Haar sensing and the subsampled Hadamard model. We define the normalized random vector $\widetilde{z}$ as:

$$\widetilde{z}_i = \frac{z_i}{\sigma_i},\ \sigma_i^2 = \frac{(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})_{ii}}{\kappa}.$$

Note that conditional on $\boldsymbol{A}$, $\widetilde{z}$ is a zero mean Gaussian vector with:

$$\mathbb{E}[\widetilde{z}_i^2|\boldsymbol{A}] = 1,\ \mathbb{E}[\widetilde{z}_i\widetilde{z}_j|\boldsymbol{A}] = \frac{(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{ij}/\kappa}{\sigma_i\sigma_j}.$$

We define the diagonal matrix $\widetilde{\boldsymbol{Z}} = \mathrm{Diag}\left(\widetilde{z}\right)$.

Lemma 27. *We have,*

$$\lim_{m\to\infty}\frac{\mathbb{E}z^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})z}{m} = \lim_{m\to\infty}\frac{\mathbb{E}\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}})\widetilde{z}}{m}\mathbb{I}_{\mathcal{E}},$$

provided the latter limit exists.

The proof of the lemma uses the fact that $\mathbb{P}(\mathcal{E}^c)\to 0$, and that on the event $\mathcal{E}$ since $\sigma_i^2\approx 1$, we have $z\approx\widetilde{z}$ and hence, the continuity estimates of Lemma 26 give the claim of this result. Complete details have been provided in Appendix 4.10.4 in the supplementary materials.

The advantage of Lemma 27 is that $\widetilde{z}_i \sim \mathcal{N}(0,1)$, and on the event $\mathcal{E}$ the coordinates of $\widetilde{z}$ have weak correlations. Consequently, Mehler's Formula (Proposition 9) can be used to analyze the leading order term in $\mathbb{E}[\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}})\widetilde{z}\ \mathbb{I}_{\mathcal{E}}]$. Before we do so, we do one additional preprocessing step.

Lemma 28. *We have:*

$$\lim_{m\to\infty}\frac{\mathbb{E}\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}})\widetilde{z}}{m}\mathbb{I}_{\mathcal{E}} = \lim_{m\to\infty}\frac{\mathbb{E}\langle\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}}),\widetilde{z}\widetilde{z}^\mathsf{T}-\widetilde{\boldsymbol{Z}}^2\rangle\mathbb{I}_{\mathcal{E}}}{m},$$

provided the latter limit exists.

Proof Sketch. Observe that we can write:

$$\begin{aligned}\widetilde{z}^\mathsf{T}\mathcal{A}\widetilde{z} &= \langle\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}}),\widetilde{z}\widetilde{z}^\mathsf{T}\rangle\\ &\overset{(a)}{=} \langle\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}}),\widetilde{z}\widetilde{z}^\mathsf{T}-\widetilde{\boldsymbol{Z}}^2\rangle + \mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}})\cdot q(\widetilde{\boldsymbol{Z}})) + \mathsf{Tr}(\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}})).\end{aligned}$$

In the step marked (a), we defined $q(\xi) = \xi^2 - 1$ which is an even function. Note that we know $|\mathsf{Tr}(\mathcal{A})|/m \leq \|\mathcal{A}\|_{\mathsf{op}} \leq C(\mathcal{A}) < \infty$. Furthermore, by Proposition 7, we know $\mathsf{Tr}(\mathcal{A})/m \xrightarrow{\text{P}} 0$, and hence by Dominated Convergence Theorem $\mathbb{E}\mathsf{Tr}(\mathcal{A})\mathbb{I}_{\mathcal{E}}/m \rightarrow 0$. Additionally, note that $\mathsf{Tr}(\mathcal{A}q(\widetilde{\boldsymbol{Z}}))$ is also an alternating form except for minor issue that $q(\xi)$ is not uniformly bounded and Lipchitz. However, the combinatorial calculations in Proposition 7 can be repeated to show that $\mathbb{E}\mathsf{Tr}(\mathcal{A} \cdot q(\widetilde{\boldsymbol{Z}}))/m \rightarrow 0$. Since we will see a more complicated version of these arguments in the remainder of the proof, we omit the details of this step. □

Note that, so far, Lemmas 27 and 28 show that:

$$\lim_{m \rightarrow \infty} \frac{\mathbb{E}\boldsymbol{z}^{\mathsf{T}}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})\boldsymbol{z}}{m} = \lim_{m \rightarrow \infty} \frac{\mathbb{E}\langle \mathcal{A}(\boldsymbol{\Psi}, \widetilde{\boldsymbol{Z}}), \widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2 \rangle \mathbb{I}_{\mathcal{E}}}{m},$$

provided the latter limit exists. We now focus on analyzing the RHS. We expand

$$\frac{\langle \mathcal{A}(\boldsymbol{\Psi}, \widetilde{\boldsymbol{Z}}), \widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2 \rangle}{m} = \frac{1}{m} \sum_{\substack{a_{1:k+1} \in [m] \\ a_1 \neq a_{k+1}}} \widetilde{z}_{a_1} (\boldsymbol{\Psi})_{a_1, a_2} q_1(\widetilde{z}_{a_2}) \cdots q_{k-1}(\widetilde{z}_{a_k}) (\boldsymbol{\Psi})_{a_k, a_{k+1}} \widetilde{z}_{a_{k+1}}.$$

Recall the notation for partitions introduced in Section 4.6.1. Observe that:

$$\{(a_1 \dots a_{k+1}) \in [m]^{k+1} : \ a_1 \neq a_{k+1}\} = \bigsqcup_{\substack{\pi \in \mathcal{P}([k+1]) \\ \pi(1) \neq \pi(k+1)}} C(\pi).$$

Hence,

$$\frac{\mathbb{E}\langle \mathcal{A}(\boldsymbol{\Psi}, \widetilde{\boldsymbol{Z}}), \widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2 \rangle \cdot \mathbb{I}_{\mathcal{E}}}{m} = \\ \frac{1}{m} \sum_{\substack{\pi \in \mathcal{P}([1:k+1]) \\ \pi(1) \neq \pi(k+1)}} \sum_{a \in C(\pi)} \mathbb{E} \, \widetilde{z}_{a_1} (\boldsymbol{\Psi})_{a_1, a_2} q_1(\widetilde{z}_{a_2}) (\boldsymbol{\Psi})_{a_2, a_3} \cdots q_{k-1}(\widetilde{z}_{a_k}) (\boldsymbol{\Psi})_{a_k, a_{k+1}} \widetilde{z}_{a_{k+1}} \cdot \mathbb{I}_{\mathcal{E}}.$$

Fix a $\pi \in \mathcal{P}([k+1])$ such that $\pi(1) \neq \pi(k+1)$, and consider a labelling $\boldsymbol{a} \in C(\pi)$. By the tower property,

$$\begin{aligned}&\mathbb{E}\widetilde{z}_{a_1}(\boldsymbol{\Psi})_{a_1,a_2}q_1(\widetilde{z}_{a_2})(\boldsymbol{\Psi})_{a_2,a_3}\cdots q_{k-1}(\widetilde{z}_{a_k})(\boldsymbol{\Psi})_{a_k,a_{k+1}}\widetilde{z}_{a_{k+1}}\mathbb{I}_{\mathcal{E}} = \\ &\qquad \mathbb{E}\left[(\boldsymbol{\Psi})_{a_1,a_2}(\boldsymbol{\Psi})_{a_2,a_3}\cdots(\boldsymbol{\Psi})_{a_k,a_{k+1}}\cdot\mathbb{E}[\widetilde{z}_{a_1}q_1(\widetilde{z}_{a_2})q_2(\widetilde{z}_{a_3})\cdots q_{k-1}(\widetilde{z}_{a_k})\widetilde{z}_{a_{k+1}}|\boldsymbol{A}]\mathbb{I}_{\mathcal{E}}\right].\end{aligned}$$

We will now use Mehler's formula (Proposition 9) to evaluate the conditional expectation upto leading order. Note that some of the random variables $\widetilde{z}_{a_{1:k+1}}$ are equal (as given by the partition π). Hence, we group them together and recenter the resulting functions. The blocks corresponding to a_1, a_{k+1} need to be treated specially due to the presence of $\widetilde{z}_{a_1}, \widetilde{z}_{a_{k+1}}$ in the above expectations. Hence, we introduce the following notations:

$$\mathscr{F}(\pi) = \pi(1),\ \mathscr{L}(\pi) = \pi(k+1),\ \mathscr{S}(\pi) = \{i \in [2:k] : |\pi(i)| = 1\}.$$

We label all the remaining blocks of π as $\mathcal{V}_1, \mathcal{V}_2 \dots \mathcal{V}_{|\pi|-|\mathscr{S}(\pi)|-2}$. Hence, the partition π is given by:

$$\pi = \mathscr{F}(\pi) \sqcup \mathscr{L}(\pi) \sqcup \left(\bigsqcup_{i\in\mathscr{S}(\pi)}\{i\}\right) \sqcup \left(\bigsqcup_{t=1}^{|\pi|-|\mathscr{S}(\pi)|-2}\mathcal{V}_i\right).$$

Note that:

$$\widetilde{z}_{a_1}\widetilde{z}_{a_{k+1}}\prod_{i=2}^{k}q_{i-1}(\widetilde{z}_{a_i}) = Q_{\mathscr{F}}(\widetilde{z}_{a_1})Q_{\mathscr{L}}(\widetilde{z}_{a_{k+1}})\left(\prod_{i\in\mathscr{S}(\pi)}q_{i-1}(\widetilde{z}_{a_i})\right)\cdot\prod_{i=1}^{|\pi|-|\mathscr{S}(\pi)|-2}(Q_{\mathcal{V}_i}(z_{a_{\mathcal{V}_i}})+\mu_{\mathcal{V}_i}),$$

where:

$$Q_{\mathscr{F}}(\xi) = \xi \cdot \prod_{i \in \mathscr{F}(\pi), i \neq 1} q_{i-1}(\xi), \tag{4.23}$$

$$Q_{\mathscr{L}}(\xi) = \xi \cdot \prod_{i \in \mathscr{L}(\pi), i \neq k+1} q_{i-1}(\xi), \tag{4.24}$$

$$\mu_{\mathcal{V}_i} = \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\prod_{j \in \mathcal{V}_i} q_{j-1}(\xi) \right], \tag{4.25}$$

$$Q_{\mathcal{V}_i}(\xi) = \prod_{j \in \mathcal{V}_i} q_{j-1}(\xi) - \mu_{\mathcal{V}_i}. \tag{4.26}$$

With this notation in place, we can apply Mehler's formula. The result is summarized in the following lemma.

Lemma 29. *For any $\pi \in \mathcal{P}([k+1])$ such that $\pi(1) \neq \pi(k+1)$, and any labelling $\boldsymbol{a} \in \mathcal{C}(\pi)$ we have:*

$$\mathbb{I}_{\mathcal{E}} \cdot \left| \mathbb{E}[\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) q_2(\widetilde{z}_{a_3}) \cdots q_{k-1}(\widetilde{z}_{a_k}) \widetilde{z}_{a_{k+1}} | \boldsymbol{A}] - \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \cdot \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) \right| \\ \leq C(\mathscr{A}) \cdot \left(\frac{\log^3(m)}{m\kappa^2} \right)^{\frac{2+|\mathcal{S}(\pi)|}{2}}, \tag{4.27a}$$

where $\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ is the matrix moment as defined in Definition 8. The coefficients $g(\boldsymbol{w}, \pi)$ are given by:

$$g(\boldsymbol{w}, \pi) = \frac{1}{\kappa^{\|\boldsymbol{w}\|} \boldsymbol{w}!} \cdot \left(\hat{Q}_{\mathscr{F}}(1) \hat{Q}_{\mathscr{L}}(1) \prod_{i \in \mathcal{S}(\pi)} \hat{q}_{i-1}(2) \right) \cdot \left(\prod_{i \in [|\pi| - |\mathcal{S}(\pi)| - 2]} \mu_{\mathcal{V}_i} \right), \tag{4.27b}$$

and, the set $\mathcal{G}_1(\pi)$ is defined as:

$$\mathcal{G}_1(\pi) \overset{def}{=} \{ \boldsymbol{w} \in \mathcal{G}(k+1) : \mathsf{d}_1(\boldsymbol{w}) = 1, \ \mathsf{d}_{k+1}(\boldsymbol{w}) = 1, \ \mathsf{d}_i(\boldsymbol{w}) = 2 \ \forall \ i \ \in \ \mathcal{S}(\pi), \\ \mathsf{d}_i(\boldsymbol{w}) = 0 \ \forall \ i \ \notin \ \{1, k+1\} \cup \mathcal{S}(\pi) \}. \tag{4.27c}$$

The proof of the lemma is obtained by instantiating Mehler's formula for this situation and identifying the leading order term. Additional details for this step are provided in Appendix 4.10.4 in the supplementary materials.

With this, we return to our analysis of:

$$\frac{\mathbb{E}\langle \mathcal{A}(\boldsymbol{\Psi}, \widetilde{\boldsymbol{Z}}), \widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2\rangle \cdot \mathbb{I}_{\mathcal{E}}}{m} = \\ \frac{1}{m} \sum_{\substack{\pi \in \mathcal{P}([1:k+1]) \\ \pi(1) \neq \pi(k+1)}} \sum_{a \in \mathcal{C}(\pi)} \mathbb{E}\, \widetilde{z}_{a_1} (\boldsymbol{\Psi})_{a_1,a_2} q_1(\widetilde{z}_{a_2}) (\boldsymbol{\Psi})_{a_2,a_3} \cdots q_{k-1}(\widetilde{z}_{a_k}) (\boldsymbol{\Psi})_{a_k,a_{k+1}} \widetilde{z}_{a_{k+1}} \cdot \mathbb{I}_{\mathcal{E}}.$$

We define the following subsets of $\mathcal{P}(k+1)$ as:

$$\mathcal{P}_1([k+1]) \overset{\text{def}}{=} \\ \{\pi \in \mathcal{P}(k+1) :\ \pi(1) \neq \pi(k+1),\ |\pi(1)| = 1, |\pi(k+1)| = 1, |\pi(j)| \leq 2\ \forall\ j\ \in\ [k+1]\}, \tag{4.28a}$$

$$\mathcal{P}_2([k+1]) \overset{\text{def}}{=} \{\pi \in \mathcal{P}(k+1) :\ \pi(1) \neq \pi(k+1)\} \backslash \mathcal{P}_1([k+1]), \tag{4.28b}$$

and the error term which was controlled in Lemma 29:

$$\epsilon(\boldsymbol{\Psi}, \boldsymbol{a}) \overset{\text{def}}{=} \mathbb{I}_{\mathcal{E}} \cdot \left(\mathbb{E}[\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) \cdots q_{k-1}(\widetilde{z}_{a_k}) \widetilde{z}_{a_{k+1}} | \boldsymbol{A}] - \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \cdot \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) \right).$$

With these definitions we consider the decomposition:

$$\frac{\mathbb{E}\langle \mathcal{A}(\boldsymbol{\Psi}, \widetilde{\boldsymbol{Z}}), \widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2\rangle \cdot \mathbb{I}_{\mathcal{E}}}{m} = \\ \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{a \in \mathcal{C}(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \mathbb{E}\left[(\boldsymbol{\Psi})_{a_1,a_2} \cdots (\boldsymbol{\Psi})_{a_k,a_{k+1}} \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) \right] - \mathsf{I} + \mathsf{II} + \mathsf{III},$$

where:

$$\mathsf{I} \overset{\text{def}}{=} \frac{1}{m} \sum_{\substack{\pi \in \mathcal{P}([k+1]) \\ \pi(1) \neq \pi(k+1)}} \sum_{\boldsymbol{a} \in \mathcal{C}(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \mathbb{E}\left[(\boldsymbol{\Psi})_{a_1,a_2} \cdots (\boldsymbol{\Psi})_{a_k,a_{k+1}} \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) \mathbb{I}_{\mathcal{E}^c}\right],$$

$$\mathsf{II} \overset{\text{def}}{=} \frac{1}{m} \sum_{\substack{\pi \in \mathcal{P}([k+1]) \\ \pi(1) \neq \pi(k+1)}} \sum_{\boldsymbol{a} \in \mathcal{C}(\pi)} \mathbb{E}\left[(\boldsymbol{\Psi})_{a_1,a_2} \cdots (\boldsymbol{\Psi})_{a_k,a_{k+1}} \epsilon(\boldsymbol{\Psi}, \boldsymbol{a}) \mathbb{I}_{\mathcal{E}}\right],$$

$$\mathsf{III} \overset{\text{def}}{=} \frac{1}{m} \sum_{\pi \in \mathcal{P}_2([k+1])} \sum_{\boldsymbol{a} \in \mathcal{C}(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \mathbb{E}\left[(\boldsymbol{\Psi})_{a_1,a_2} \cdots (\boldsymbol{\Psi})_{a_k,a_{k+1}} \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})\right].$$

Define $\boldsymbol{\ell}_{k+1} \in \mathcal{G}(k+1)$ to be the weight matrix of a simple line graph, i.e.

$$(\boldsymbol{\ell}_{k+1})_{ij} = \begin{cases} 1 : & |j-i| = 1 \\ 0 : & \text{otherwise} \end{cases}.$$

This decomposition can be written compactly as:

$$\mathsf{I} = \frac{1}{m} \sum_{\substack{\pi \in \mathcal{P}([1:k+1]) \\ \pi(1) \neq \pi(k+1)}} \sum_{\boldsymbol{a} \in \mathcal{C}(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \mathbb{I}_{\mathcal{E}^c}\right],$$

$$\mathsf{II} = \frac{1}{m} \sum_{\substack{\pi \in \mathcal{P}([1:k+1]) \\ \pi(1) \neq \pi(k+1)}} \sum_{\boldsymbol{a} \in \mathcal{C}(\pi)} \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \epsilon(\boldsymbol{\Psi}, \boldsymbol{a}) \mathbb{I}_{\mathcal{E}}\right],$$

$$\mathsf{III} = \frac{1}{m} \sum_{\pi \in \mathcal{P}_2([1:k+1])} \sum_{\boldsymbol{a} \in \mathcal{C}(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right].$$

We will show that $\mathsf{I}, \mathsf{II}, \mathsf{III} \rightarrow 0$. Showing this involves the following components:

1. Bounds on matrix moments $\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right]$, which have been developed in Lemma 19.

2. Controlling the size of the set $|\mathcal{C}(\pi)|$ (since we sum over $\boldsymbol{a} \in \mathcal{C}(\pi)$ in the above terms). Since,

$$|\mathcal{C}(\pi)| = m(m-1)\cdots(m-|\pi|+1) \asymp m^{|\pi|},$$

we need to develop bounds on $|\pi|$. This is done in the following lemma. In contrast, the sums over $\pi \in \mathcal{P}([k+1])$ and $\boldsymbol{w} \in \mathcal{G}_1(\pi)$ are not a cause of concern since $|\mathcal{P}([k+1])|, |\mathcal{G}_1(\pi)|$ depend only on k (which is held fixed), and not on m.

Lemma 30. *For any $\pi \in \mathcal{P}_1([k+1])$, we have:*

$$|\pi| = \frac{k+3+|\mathcal{S}(\pi)|}{2} \implies |C(\pi)| \leq m^{\frac{k+3+|\mathcal{S}(\pi)|}{2}}.$$

For any $\pi \in \mathcal{P}_2([k+1])$, we have:

$$|\pi| \leq \frac{k+2+|\mathcal{S}(\pi)|}{2} \implies |C(\pi)| \leq m^{\frac{k+2+|\mathcal{S}(\pi)|}{2}}.$$

Proof. Consider any $\pi \in \mathcal{P}([k+1])$ such that $\pi(1) \neq \pi(k+1)$. Recall that the disjoint blocks of $|\pi|$ were given by:

$$\pi = \mathscr{F}(\pi) \sqcup \mathscr{L}(\pi) \sqcup \left(\bigsqcup_{i \in \mathcal{S}(\pi)} \{i\} \right) \sqcup \left(\bigsqcup_{t=1}^{|\pi|-|\mathcal{S}(\pi)|-2} \mathcal{V}_i \right).$$

Hence,

$$k+1 = |\mathscr{F}(\pi)| + |\mathscr{L}(\pi)| + |\mathcal{S}(\pi)| + \sum_{t=1}^{|\pi|-|\mathcal{S}(\pi)|-2} |\mathcal{V}_i|.$$

Note that:

$$|\mathscr{F}(\pi)| \geq 1 \qquad \text{(Since } 1 \in \mathscr{F}(\pi)\text{)}, \tag{4.29a}$$

$$|\mathscr{L}(\pi)| \geq 1 \qquad \text{(Since } k+1 \in \mathscr{L}(\pi)\text{)}, \tag{4.29b}$$

$$|\mathcal{V}_i| \geq 2 \qquad \text{(Since } \mathcal{V}_i \text{ are not singletons)}. \tag{4.29c}$$

Hence,

$$k+1 \geq |\mathscr{F}(\pi)| + |\mathscr{L}(\pi)| + |\mathscr{S}(\pi)| + 2|\pi| - 2|\mathscr{S}(\pi)| - 4,$$

which implies:

$$\begin{aligned} |\pi| &\leq \frac{k+5+|\mathscr{S}(\pi)| - |\mathscr{F}(\pi)| - |\mathscr{L}(\pi)|}{2} \\ &\leq \frac{k+3+|\mathscr{S}(\pi)|}{2}, \end{aligned} \tag{4.30}$$

and hence,

$$|C(\pi)| \leq m^{|\pi|} \leq m^{\frac{k+3+|\mathscr{S}(\pi)|}{2}}.$$

Finally, observe that:

1. For any $\pi \in \mathcal{P}_1([k+1])$ each of the inequalities in (4.29) are exactly tight by the definition of $\mathcal{P}_1([k+1])$ in (4.28), and hence:

$$|\pi| = \frac{k+3+|\mathscr{S}(\pi)|}{2}.$$

2. For any $\pi \in \mathcal{P}_2([k+1])$, one of the inequalities in (4.29) must be strict (see (4.28)). Hence, when $\pi \in \mathcal{P}_2([k+1])$, we have the improved bound:

$$|\pi| \leq \frac{k+2+|\mathscr{S}(\pi)|}{2}.$$

This proves the claims of the lemma. □

We will now show that $\mathsf{I}, \mathsf{II}, \mathsf{III} \to 0$.

Lemma 31. *We have,*

$$\mathsf{I} \to 0,\ \mathsf{II} \to 0,\ \mathsf{III} \to 0 \ \text{ as } m \to \infty,$$

and hence:

$$\lim_{m\to\infty} \frac{\mathbb{E} z^{\mathsf{T}} \mathcal{A} z}{m} = \lim_{m\to\infty} \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{a \in \mathcal{C}(\pi)} \sum_{w \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right],$$

provided the latter limit exists.

Proof. First, note that for any $\boldsymbol{w} \in \mathcal{G}_1(\pi)$, we have:

$$\|\boldsymbol{w}\| = \frac{1}{2} \sum_{i=1}^{k+1} \mathsf{d}_i(\boldsymbol{w}) = \frac{1 + 1 + 2|\mathcal{S}(\pi)|}{2} = 1 + |\mathcal{S}(\pi)| \ \text{ (See (4.27)).}$$

Furthermore, recalling that $\boldsymbol{\ell}_{k+1}$ is the weight matrix of a simple line graph, $\|\boldsymbol{\ell}_{k+1}\| = k$. Now, we apply Lemma 19 to obtain:

$$\begin{aligned} |\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \mathbb{I}_{\mathcal{E}^c}\right]| &\leq \sqrt{\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, 2\boldsymbol{w} + 2\boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right]} \sqrt{\mathbb{P}(\mathcal{E}^c)} \\ &\overset{\text{(a)}}{\leq} \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \cdot \sqrt{\mathbb{P}(\mathcal{E}^c)} \\ &\leq \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \cdot \frac{C_k}{m}. \end{aligned}$$

Analogously we can obtain:

$$\begin{aligned} \mathbb{E}|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})| &\leq \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{k}{2}}, \\ \mathbb{E}\left[|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})|\right] &\leq \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \end{aligned}$$

Further, recall that by Lemma 29 we have:

$$|\epsilon(\boldsymbol{\Psi}, \boldsymbol{a})| \leq C(\mathcal{A}) \cdot \left(\frac{\log^3(m)}{m\kappa^2}\right)^{\frac{2+|\mathcal{S}(\pi)|}{2}}.$$

Using these estimates, we obtain:

$$\begin{aligned} |\mathrm{I}| &\leq \frac{C(\mathcal{A})}{m} \cdot \sum_{\substack{\pi:\mathcal{P}([k+1]) \\ \pi(0)\neq\pi(k+1)}} |C(\pi)| \cdot \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \cdot \frac{C_k}{m} \\ &\overset{(a)}{\leq} \frac{C(\mathcal{A})}{m} \cdot \sum_{\substack{\pi:\mathcal{P}([k+1]) \\ \pi(0)\neq\pi(k+1)}} m^{\frac{k+3+|\mathcal{S}(\pi)|}{2}} \cdot \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \cdot \frac{C_k}{m} \\ &= O\left(\frac{\mathrm{polylog}(m)}{m}\right). \end{aligned}$$

In addition:

$$\begin{aligned} |\mathrm{II}| &\leq \frac{C(\mathcal{A})}{m} \cdot \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{k}{2}} \cdot \sum_{\substack{\pi:\mathcal{P}([k+1]) \\ \pi(0)\neq\pi(k+1)}} |C(\pi)| \cdot \left(\frac{\log^3(m)}{m\kappa^2}\right)^{\frac{2+|\mathcal{S}(\pi)|}{2}} \\ &\overset{(a)}{\leq} \frac{C(\mathcal{A})}{m} \cdot \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{k}{2}} \cdot \sum_{\substack{\pi:\mathcal{P}([k+1]) \\ \pi(0)\neq\pi(k+1)}} m^{\frac{k+3+|\mathcal{S}(\pi)|}{2}} \cdot \left(\frac{\log^3(m)}{m\kappa^2}\right)^{\frac{2+|\mathcal{S}(\pi)|}{2}} \\ &= O\left(\frac{\mathrm{polylog}(m)}{\sqrt{m}}\right). \end{aligned}$$

Furthermore:

$$
\begin{aligned}
|\mathsf{III}| &\leq \frac{C(\mathcal{A})}{m} \cdot \sum_{\pi : \mathcal{P}_2([k+1])} |C(\pi)| \cdot \left(\frac{C_k \log^2(m)}{m} \right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \\
&\overset{(a)}{\leq} \frac{C(\mathcal{A})}{m} \cdot \sum_{\pi : \mathcal{P}_2([k+1])} m^{\frac{k+2+|C(\pi)|}{2}} \cdot \left(\frac{C_k \log^2(m)}{m} \right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \\
&= O\left(\frac{\mathrm{polylog}(m)}{\sqrt{m}} \right).
\end{aligned}
$$

In each of the above displays, in the steps marked (a), we used the bounds on $|C(\pi)|$ from Lemma 30. C_k denotes a constant depending only on k and $C(\mathcal{A})$ denotes a constant depending only on k and the functions appearing in $\mathcal{A}$. This concludes the proof of this lemma. □

So far we have shown that:

$$
\lim_{m \to \infty} \frac{\mathbb{E} \boldsymbol{z}^{\mathsf{T}} \mathcal{A} \boldsymbol{z}}{m} = \lim_{m \to \infty} \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{a \in C(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \right],
$$

provided the latter limit exists. Our goal is to show that the limit on the LHS exists and is universal across the subsampled Haar and Hadamard models. In order to do so, we will leverage the fact that the first order term in the expansion of $\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \right]$ is the same for the two models if $\boldsymbol{w} + \boldsymbol{\ell}_{k+1}$ is dissortive with respect to π and if $\boldsymbol{a}$ is a conflict-free labelling (Propositions 10 and 11). Hence, we need to argue that the contribution of terms corresponding to $\boldsymbol{w} : \boldsymbol{w} + \boldsymbol{\ell}_{k+1} \notin \mathcal{G}_{\mathsf{DA}}(\pi)$ and $\boldsymbol{a} \notin \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)$ are negligible. Towards this end, we consider the decomposition:

$$
\begin{aligned}
&\frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{a \in C(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \right] = \\
&\frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w} + \boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \right] + \mathsf{IV} + \mathsf{V},
\end{aligned}
$$

where:

$$\mathsf{IV} \overset{\text{def}}{=} \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{a \in C(\pi)} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1} \notin \mathcal{G}_{\mathsf{DA}}(\pi)}} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right],$$

$$\mathsf{V} \overset{\text{def}}{=} \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in C(\pi) \backslash \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}, \pi)} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right].$$

Lemma 32. *We have* $\mathsf{IV} \rightarrow 0, \mathsf{V} \rightarrow 0$*, as* $m \rightarrow \infty$*, and hence:*

$$\lim_{m \to \infty} \frac{\mathbb{E} z^{\mathsf{T}} \mathcal{A} z}{m} =$$
$$\lim_{m \to \infty} \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}, \pi)} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right],$$

provided the latter limit exists.

Proof. We will prove this in two steps.

Step 1: $\mathsf{IV} \rightarrow 0$**.** We consider the two sensing models separately:

1. Subsampled Hadamard Sensing: In this case, Proposition 11 tells us that if $\boldsymbol{w} + \boldsymbol{\ell}_{k+1} \notin \mathcal{G}_{\mathsf{DA}}(\pi)$, then:

$$\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right] = 0,$$

   and hence, $\mathsf{IV} = 0$.

2. Subsampled Haar Sensing: Observe that, since $\|\boldsymbol{w}\| + \|\boldsymbol{\ell}_{k+1}\| = 1 + |\mathcal{S}(\pi)| + k$, we have:

$$\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right] = \frac{\mathbb{E}\left[\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right]}{m^{\frac{1+|\mathcal{S}(\pi)|+k}{2}}}.$$

By Proposition 10, we know that:

$$\left| \mathbb{E}\left[\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \right] - \prod_{\substack{s,t \in [|\pi|] \\ s \leq t}} \mathbb{E}\left[Z_{st}^{W_{st}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)} \right] \right| \leq \frac{K_1 \log^{K_2}(m)}{m^{\frac{1}{4}}},$$

where K_1, K_2, K_3 are universal constants depending only on k. Note that since $\boldsymbol{w} + \boldsymbol{\ell}_{k+1} \notin \mathcal{G}_{\mathsf{DA}}(\pi)$, we must have some $s \in [|\pi|]$ such that:

$$W_{ss}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi) \geq 1.$$

Recall that $\mathsf{d}_i(\boldsymbol{w}) = 0$ for any $i \notin \{1, k+1\} \cup \mathcal{S}(\pi)$ (since $\boldsymbol{w} \in \mathcal{G}_1(\pi)$), and furthermore, $|\pi(i)| = 1 \; \forall \; i \; \in \; \{1, k+1\} \cup \mathcal{S}(\pi)$ (since $\pi \in \mathcal{P}_1(k+1)$). Hence, we have $\boldsymbol{w} \in \mathcal{G}_{\mathsf{DA}}(\pi)$ and in particular, $W_{ss}(\boldsymbol{w}, \pi) = 0$. Consequently, we must have $W_{ss}(\boldsymbol{\ell}_{k+1}, \pi) \geq 1$. Recall that $\boldsymbol{\ell}_{k+1}$ is the weight matrix of a line graph:

$$(\boldsymbol{\ell}_{k+1})_{ij} = \begin{cases} 1 : & |i - j| = 1 \\ 0 : & \text{otherwise} \end{cases}.$$

Consequently, since $W_{ss}(\boldsymbol{\ell}_{k+1}, \pi) \geq 1$, we must have for some $i \in [k]$, $\pi(i) = \pi(i+1) = \mathcal{V}_s$. However, since $\pi \in \mathcal{P}_1(k+1)$, $|\mathcal{V}_s| \leq 2$, and hence, $\mathcal{V}_s = \{i, i+1\}$. This means that $W_{ss}(\boldsymbol{\ell}_{k+1}, \pi) = 1 = W_{ss}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)$. Consequently, since $\mathbb{E}Z_{ss} = 0$, we have:

$$\prod_{\substack{s,t \in [|\pi|] \\ s \leq t}} \mathbb{E}\left[Z_{st}^{W_{st}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)} \right] = 0,$$

or

$$|\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) \right]| = \frac{C_k \log^K(m)}{m^{\frac{1+|\mathcal{S}(\pi)|+k}{2} + \frac{1}{4}}},$$

where C_k, K are constants that depend only on k. Recalling Lemma 30,

$$|\mathcal{C}(\pi)| \leq m^{|\pi|} \leq m^{\frac{k+3+|\mathcal{S}(\pi)|}{2}},$$

we obtain:

$$|\mathsf{IV}| \leq \frac{C(\mathcal{A})}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} |\mathcal{C}(\pi)| \cdot \frac{C_k \log^K(m)}{m^{\frac{1+|\mathcal{S}(\pi)|+k}{2}+\frac{1}{4}}} = O\left(\frac{\operatorname{polylog}(m)}{m^{\frac{1}{4}}}\right) \to 0.$$

Step 2: $\mathsf{V} \to 0$. Using Lemma 21, we know that

$$|\mathcal{C}(\pi) \backslash \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)| \leq (k+1)^4 m^{|\pi|-1}.$$

In Lemma 30, we showed that for any $\pi \in \mathcal{P}_1([k+1])$,

$$|\pi| = \frac{k+3+|\mathcal{S}(\pi)|}{2}.$$

Hence,

$$|\mathcal{C}(\pi) \backslash \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)| \leq (k+1)^4 \cdot m^{\frac{k+1+|\mathcal{S}(\pi)|}{2}}.$$

We already know from Lemma 19 that:

$$|\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right]| \leq \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{\|\boldsymbol{w}\|+\|\ell_{k+1}\|}{2}} \leq \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}}.$$

This gives us:

$$\begin{aligned}|\mathsf{V}| &\leq \frac{C}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} |\mathcal{C}(\pi) \backslash \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}, \pi)| \cdot \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \\ &= O\left(\frac{\text{polylog}(m)}{m}\right)\end{aligned}$$

which goes to zero as claimed.

□

To conclude, we have shown that:

$$\begin{aligned}&\lim_{m \to \infty} \frac{\mathbb{E} \boldsymbol{z}^{\mathsf{T}} \mathcal{A} \boldsymbol{z}}{m} = \\ &\lim_{m \to \infty} \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}, \pi)} g(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}+\boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right],\end{aligned}$$

provided the limit on the RHS exists. In the following lemma we explicitly evaluate the limit on the RHS, and in particular, show it exists and is identical for the two sensing models.

Lemma 33. *For both the subsampled Haar sensing and Hadamard sensing model, we have:*

$$\lim_{m \to \infty} \frac{\mathbb{E} \boldsymbol{z}^{\mathsf{T}} \mathcal{A} \boldsymbol{z}}{m} = \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} g(\boldsymbol{w}, \pi) \cdot \mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}, \pi),$$

where,

$$\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}, \pi) \overset{def}{=} \prod_{\substack{s,t \in [|\pi|] \\ s<t}} \mathbb{E}\left[Z^{W_{st}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}, \pi)}\right], \; Z \sim \mathcal{N}\left(0, \kappa(1-\kappa)\right).$$

Proof. By Propositions 11 (for the subsampled Hadamard model) and 10 (for the subsampled Haar model) we know that, if $\boldsymbol{w} + \boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)$ and $\boldsymbol{a} \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)$, we have:

$$\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a}) = \mu(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi) + \epsilon(\boldsymbol{w}, \pi, \boldsymbol{a}),$$

where

$$|\epsilon(\boldsymbol{w}, \pi, \boldsymbol{a})| \leq \frac{K_1 \log^{K_2}(m)}{m^{\frac{1}{4}}}, \ \forall\ m \geq K_3,$$

for some constants K_1, K_2, K_3 depending only on k. Hence, we can consider the decomposition:

$$\frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w} + \boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)} g(\boldsymbol{w}, \pi) \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})\right] = \mathsf{VI} + \mathsf{VII},$$

where:

$$\mathsf{VI} \overset{\text{def}}{=} \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w} + \boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)} g(\boldsymbol{w}, \pi) \cdot \frac{\mu(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)}{m^{\frac{1+\mathcal{S}(\pi)+k}{2}}},$$

$$\mathsf{VII} \overset{\text{def}}{=} \frac{1}{m} \sum_{\pi \in \mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_1(\pi) \\ \boldsymbol{w} + \boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)} g(\boldsymbol{w}, \pi) \cdot \frac{\epsilon(\boldsymbol{w}, \pi, \boldsymbol{a})}{m^{\frac{1+\mathcal{S}(\pi)+k}{2}}}.$$

We can upper bound $|\mathsf{VII}|$ as follows:

$$|\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)| \leq |C(\pi)| \leq m^{\frac{k+3+|\mathcal{S}(\pi)|}{2}}.$$

Thus:

$$\begin{aligned} |\mathsf{VII}| &\leq \frac{C(\mathcal{A})}{m} \cdot C_k \cdot |\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi)| \cdot \frac{1}{m^{\frac{1+|\mathcal{S}(\pi)|+k}{2}}} \cdot \frac{K_1 \log^{K_2}(m)}{m^{\frac{1}{4}}} \\ &= O\left(\frac{\text{polylog}(m)}{m^{\frac{1}{4}}}\right) \to 0. \end{aligned}$$

Moreover, can compute:

$$\begin{aligned}
\lim_{m\to\infty} (\mathsf{VI}) &= \lim_{m\to\infty} \frac{1}{m} \sum_{\pi\in\mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a\in\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)} g(\boldsymbol{w},\pi)\cdot \frac{\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)}{m^{\frac{1+\mathcal{S}(\pi)+k}{2}}} \\
&= \lim_{m\to\infty} \frac{1}{m} \sum_{\pi\in\mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} g(\boldsymbol{w},\pi)\cdot \frac{\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)}{m^{\frac{1+|\mathcal{S}(\pi)|+k}{2}}} \cdot |\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)| \\
&\overset{\text{(a)}}{=} \lim_{m\to\infty} \sum_{\pi\in\mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} g(\boldsymbol{w},\pi)\cdot \mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)\cdot \frac{|\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)|}{m^{|\pi|}} \\
&\overset{\text{(b)}}{=} \sum_{\pi\in\mathcal{P}_1([k+1])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_1(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} g(\boldsymbol{w},\pi)\cdot \mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi).
\end{aligned}$$

In the step marked (a) we used the fact that $|\pi| = (3 + |\mathcal{S}(\pi)| + k)/2$ for any $\pi \in \mathcal{P}_1([k+1])$ (Lemma 30), and in step (b) we used Lemma 21 ($|\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)|/m^{|\pi|} \to 1$). This proves the claim of the lemma. □

In the following lemma, we show that the combinatorial sum obtained in Lemma 33 can be significantly simplified.

Lemma 34. *For both the subsampled Haar sensing and Hadamard sensing models, we have:*

$$\lim_{m\to\infty} \frac{\mathbb{E}\boldsymbol{z}^\mathsf{T}\mathcal{A}\boldsymbol{z}}{m} = (1-\kappa)^k \cdot \prod_{i=1}^{k-1} \hat{q}_i(2).$$

In particular, Proposition 12 holds.

Proof. We claim that the only partition with a non-zero contribution is:

$$\pi = \bigsqcup_{i=1}^{k+1} \{i\}.$$

In order to see this, suppose π is not entirely composed of singleton blocks. Define:

$$i_\star \overset{\text{def}}{=} \min\{i \in [k+1] : |\pi(i)| > 1\}.$$

Note that $i_\star > 1$ since we know that $|\pi(1)| = |\mathscr{F}(\pi)| = 1$ for any $\pi \in \mathcal{P}_1(k+1)$. Since $\pi \in \mathcal{P}_1([k+1])$, we must have $|\pi(i_\star)| = 2$, hence, denote:

$$\pi(i_\star) = \{i_\star, j_\star\},$$

for some $j_\star > i_\star + 1$ ($i_\star \leq j_\star$ since it is the first index which is not in a singleton block, and $j_\star \neq i_\star + 1$ since otherwise $\boldsymbol{w} + \boldsymbol{\ell}_{k+1}$ will not be disassortative). Let us label the first few blocks of π as:

$$\mathcal{V}_1 = \{1\},\ \mathcal{V}_2 = \{2\}, \ldots, \mathcal{V}_{i_\star - 1} = \{i_\star - 1\},\ \mathcal{V}_{i_\star} = \{i_\star, j_\star\}.$$

Next, we compute:

$$\begin{aligned} W_{i_\star-1,i_\star}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi) &= W_{i_\star-1,i_\star}(\boldsymbol{\ell}_{k+1}, \pi) + W_{i_\star-1,i_\star}(\boldsymbol{w}, \pi) \\ &\overset{\text{(a)}}{=} W_{i_\star-1,i_\star}(\boldsymbol{\ell}_{k+1}, \pi) \\ &\overset{\text{(b)}}{=} \mathbf{1}_{i_\star-1 \in \mathcal{V}_{i_\star-1}} + \mathbf{1}_{i_\star+1 \in \mathcal{V}_{i_\star-1}} + \mathbf{1}_{j_\star-1 \in \mathcal{V}_{i_\star-1}} + \mathbf{1}_{j_\star+1 \in \mathcal{V}_{i_\star-1}} \\ &\overset{\text{(c)}}{=} \mathbf{1}_{i_\star-1 = i_\star-1} + \mathbf{1}_{i_\star+1 = i_\star-1} + \mathbf{1}_{j_\star-1 = i_\star-1} + \mathbf{1}_{j_\star+1 = i_\star-1} \\ &\overset{\text{(d)}}{=} 1. \end{aligned}$$

In the step marked (a), we used the fact that since $\boldsymbol{w} \in \mathcal{G}_1(\pi)$ and $|\pi(i_\star)| = |\pi(j_\star)| = 2$, we must have $d_{i_\star}(\boldsymbol{w}) = d_{j_\star}(\boldsymbol{w}) = 0$ and $W_{i_\star-1,i_\star}(\boldsymbol{w}, \pi) = 0$. In the step marked (b), we used the definition of $\boldsymbol{\ell}_{k+1}$ (that it is the line graph). In the step marked (c), we used the fact that $\mathcal{V}_{i_\star-1} = \{i_{\star-1}\}$. In the step marked (d), we used the fact that $j_\star > i_\star + 1$.

Hence, we have shown that for any $\pi \neq \sqcup_{i=1}^{k+1}\{i\}$, we have:

$$\mu(\boldsymbol{w}, \pi) = 0\ \forall\ \boldsymbol{w} \text{ such that } \boldsymbol{w} \in \mathcal{G}_1(\pi),\ \boldsymbol{w} + \boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi).$$

Next, let $\pi = \sqcup_{i=1}^{k+1} \{i\}$. We observe for any $\boldsymbol{w}$ such that $\boldsymbol{w} \in \mathcal{G}_{1}(\pi),\ \boldsymbol{w} + \boldsymbol{\ell}_{k+1} \in \mathcal{G}_{\mathsf{DA}}(\pi)$, we have:

$$\begin{aligned}\mu(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi) &= \prod_{\substack{s,t \in [|\pi|] \\ s<t}} \mathbb{E}\left[Z^{W_{st}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1},\pi)}\right],\ Z \sim \mathcal{N}\left(0, \kappa(1-\kappa)\right) \\ &= \prod_{\substack{i,j \in [k+1] \\ i<j}} \mathbb{E}\left[Z^{w_{ij}+(\ell_{k+1})_{ij},\pi)}\right],\ Z \sim \mathcal{N}\left(0, \kappa(1-\kappa)\right).\end{aligned}$$

Note that since $\mathbb{E}Z = 0$, for $\mu(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi) \neq 0$, we must have:

$$w_{ij} \geq (\ell_{k+1})_{ij},\ \forall\, i, j \ \in \ [k].$$

However, since $\boldsymbol{w} \in \mathcal{G}_{1}(\pi)$ we have:

$$\mathsf{d}_1(\boldsymbol{w}) = \mathsf{d}_{k+1}(\boldsymbol{w}) = 1,\ \mathsf{d}_i(\boldsymbol{w}) = 2\ \forall\, i \ \in \ [2:k],$$

so, $\boldsymbol{w} = \boldsymbol{\ell}_{k+1}$. Hence, recalling the formula for $g(\boldsymbol{w}, \pi)$ from Lemma 29, we obtain:

$$\lim_{m\to\infty} \frac{\mathbb{E}\boldsymbol{z}^{\mathsf{T}}\mathcal{A}\boldsymbol{z}}{m} = (1-\kappa)^{k} \cdot \prod_{i=1}^{k-1} \hat{q}_i(2).$$

This proves the statement of the lemma and also Proposition 12 (see Remark 21 regarding how the analysis extends to other types). □

Throughout this section, we assumed that the alternating product $\mathcal{A}$ was of Type I. The following remark outlines how the analysis of this section extends to other types.

Remark 21. *The analysis of the other cases can be reduced to Type 1 as follows: Consider an alternating form $\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})$ of Type 1:*

$$\mathcal{A} = p_1(\boldsymbol{\Psi})q_1(\boldsymbol{Z})p_1(\boldsymbol{\Psi}) \cdots q_{k-1}(\boldsymbol{Z})p_k(\boldsymbol{\Psi}),$$

but the more general quadratic form:

$$\frac{1}{m}\mathbb{E}\alpha(\boldsymbol{z})^{\mathsf{T}}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})\beta(\boldsymbol{z}), \tag{4.31}$$

where $\alpha, \beta : \mathbb{R} \to \mathbb{R}$ are <u>odd functions</u> whose absolute values can be upper bounded by a polynomial. They act on the vector $\boldsymbol{z}$ entry-wise. This covers all the types in a unified way:

1. *For Type 1 case: We take $\alpha(z) = \beta(z) = z$.*

2. *For the Type 2 case, we write:*

   $$\boldsymbol{z}^{\mathsf{T}} p_1(\boldsymbol{\Psi}) q_1(\boldsymbol{Z}) p_1(\boldsymbol{\Psi}) \cdots q_k(\boldsymbol{Z}) p_k(\boldsymbol{\Psi}) q_k(\boldsymbol{Z}) \boldsymbol{z} = \alpha(\boldsymbol{z})^{\mathsf{T}}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})\beta(\boldsymbol{z}),$$

   where $\alpha(z) = z, \beta(z) = zq_k(z)$.

3. *For the Type 3 case:*

   $$\boldsymbol{z}^{\mathsf{T}} q_0(\boldsymbol{Z}) p_1(\boldsymbol{\Psi}) q_1(\boldsymbol{Z}) p_1(\boldsymbol{\Psi}) \cdots q_{k-1}(\boldsymbol{Z}) p_k(\boldsymbol{\Psi}) q_k(\boldsymbol{Z}) \boldsymbol{z} = \alpha(\boldsymbol{z})^{\mathsf{T}}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})\beta(\boldsymbol{z}),$$

   where $\alpha(z) = zq_0(z), \beta(z) = zq_k(z)$.

4. *For the Type 4 case:*

   $$\boldsymbol{z}^{\mathsf{T}} q_0(\boldsymbol{Z}) p_1(\boldsymbol{\Psi}) q_1(\boldsymbol{Z}) p_2(\boldsymbol{\Psi}) \cdots q_{k-1}(\boldsymbol{Z}) p_k(\boldsymbol{\Psi}) \boldsymbol{z} = \alpha(\boldsymbol{z})^{\mathsf{T}}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})\beta(\boldsymbol{z}),$$

   where $\alpha(z) = zq_0(z), \beta(z) = z$.

The analysis of the more general quadratic form in (4.31) *is analogous to the analysis outlined in this section. Lemmas 27 and 28 extend straightforwardly. Inspecting the proof of Lemma 29 shows that the same error bound continues to hold (after suitably redefining $c(\boldsymbol{w},\pi)$), since α, β are odd (as in the case $\alpha(z) = \beta(z) = z$). The subsequent lemmas after that hold verbatim for the more general quadratic form* (4.31).

4.9 Conclusion and Future Work

In this work we analyzed the dynamics of linearized Approximate message passing algorithms for phase retrieval when the sensing matrix is generated by sub-sampling n columns of a $m \times m$ Hadamard-Walsh matrix under an average-case Gaussian prior assumption on the signal. We showed that the dynamics of linearized AMP algorithms for these sensing matrices are asymptotically indistinguishable from the dynamics in the case when the sensing matrix is generated by sampling n columns of a uniformly random $m \times m$ orthogonal matrix. This provides a theoretical justification for an empirically observed universality phenomena in a particular case. It would be interesting to extend our results in the following ways:

Other structured ensembles: In this chapter, while we focused on the sub-sampled Hadamard sensing model, we believe our results should extend to other popular structured matrices with orthogonal columns such as randomly sub-sampled Fourier, Discrete Cosine Transform matrices, and CDP matrices. For these ensembles, there exist analogues of Lemma 20 which would make it possible to prove counterparts of Proposition 11.

Non-linear AMP Algorithms: Our results hold for linearized AMP algorithms which are not the state-of-the-art message passing algorithms for phase retrieval. It would be interesting to extend our results to include general non-linear AMP algorithms.

Non-Gaussian Priors: Simulations show that the universality of the dynamics of linearized AMP algorithms continues to hold even if the signal is not drawn from a Gaussian prior, but is an actual image. Hence it would be interesting to extend our results to general i.i.d. priors and more realistic models for signals.

4.10 Supplementary materials

4.10.1 Proof of Lemmas 22 and 23

Proof of Lemma 22

Proof of Lemma 22. Recall that, $\boldsymbol{A}\boldsymbol{A}^\mathsf{T} = \boldsymbol{U}\boldsymbol{B}\boldsymbol{U}^\mathsf{T},\ \boldsymbol{\Psi} = \boldsymbol{A}\boldsymbol{A}^\mathsf{T} - \mathbb{E}[\boldsymbol{A}\boldsymbol{A}^\mathsf{T}|\boldsymbol{U}] = \boldsymbol{U}(\boldsymbol{B} - \kappa \boldsymbol{I}_m)\boldsymbol{U}^\mathsf{T}$ where $\boldsymbol{B}$ is a uniformly random $m \times m$ diagonal matrix with exactly n entries set to 1 and the remaining entries set to 0. Using the concentration inequality of Lemma 18:

$$\mathbb{P}\left(|(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{ij} - \mathbb{E}(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{ij}| > \epsilon \;\Big|\; \boldsymbol{U}\right) \leq 4\exp\left(-\frac{\epsilon^2}{8m\|\boldsymbol{U}\|_\infty^4}\right),. \tag{4.32}$$

Setting $\epsilon = \sqrt{32 \cdot m \cdot \|\boldsymbol{U}\|_\infty^4 \cdot \log(m)}$ in (4.32) we obtain,

$$\mathbb{P}\left(|(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{ij} - \mathbb{E}(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{ij}| > \sqrt{32 \cdot m \cdot \|\boldsymbol{U}\|_\infty^4 \cdot \log(m)} \;\Big|\; \boldsymbol{U}\right) \leq \frac{4}{m^4}.$$

By a union bound, $\mathbb{P}(\mathcal{E}^c|\boldsymbol{U}) \leq 4/m^2 \to 0$. In order to prove the claim of the lemma for the subsampled Haar model, we first note that by Fact 5 we have,

$$\mathbb{P}\left(|O_{ij}| > \sqrt{\frac{8\log(m)}{m}}\right) \leq \frac{2}{m^4}.$$

By a union bound $\mathbb{P}(\|\boldsymbol{O}\|_\infty > \sqrt{8\log(m)/m}) \leq 2m^{-2}$. This gives us:

$$\begin{aligned}\mathbb{P}\left(\left\{\|\boldsymbol{O}\|_\infty \leq \sqrt{\frac{8\log(m)}{m}}\right\} \cap \mathcal{E}\right) &\geq 1 - \mathbb{P}\left(\|\boldsymbol{O}\|_\infty > \sqrt{\frac{8\log(m)}{m}}\right) - \mathbb{P}(\mathcal{E}^c)\\ &\geq 1 - \frac{2}{m^2} - \mathbb{E}\mathbb{P}(\mathcal{E}^c|\boldsymbol{U})\\ &\geq 1 - \frac{6}{m^2}.\end{aligned}$$

This concludes the proof of the lemma. □

Proof of Lemma 23

Proof of Lemma 23. Consider any alternating product $\mathcal{A}$ (see Definition 7):

$$\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z}) = (\boldsymbol{\Psi})q_1(\mathbf{Z})(\boldsymbol{\Psi}) \cdots q_k(\mathbf{Z}).$$

Note that in the above expression, we have assumed the alternating product is of Type 2 but the following argument applies to all the other types too. We define:

$$\mathcal{A}_i = (\boldsymbol{\Psi})q_1(\mathbf{Z})(\boldsymbol{\Psi})q_2(\mathbf{Z}) \cdots (\boldsymbol{\Psi})q_i(\mathbf{Z})(\boldsymbol{\Psi})q_{i+1}(\mathbf{Z}')(\boldsymbol{\Psi})q_{i+2}(\mathbf{Z}') \cdots (\boldsymbol{\Psi})q_k(\mathbf{Z}').$$

Then we can express $\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z}') - \mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z})$ as a telescoping sum:

$$\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z}) - \mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z}') = \sum_{i=1}^{k} (\mathcal{A}_i - \mathcal{A}_{i-1}).$$

Hence,

$$\left| \frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z})}{m} - \frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \mathbf{Z}')}{m} \right| \leq \frac{1}{m} \sum_{i=1}^{k} |\mathsf{Tr}(\mathcal{A}_i - \mathcal{A}_{i-1})|.$$

Next we observe that:

$$\begin{aligned}
&|\mathsf{Tr}(\mathcal{A}_i - \mathcal{A}_{i-1})| \\
&= |\mathsf{Tr}((\boldsymbol{\Psi})q_1(\boldsymbol{Z}) \cdots (\boldsymbol{\Psi})q_{i-1}(\boldsymbol{Z}) \cdot (q_i(\boldsymbol{Z}) - q_i(\boldsymbol{Z}')) \cdot (\boldsymbol{\Psi})q_{i+1}(\boldsymbol{Z}') \cdots (\boldsymbol{\Psi})q_k(\boldsymbol{Z}'))| \\
&\leq \left\|(\boldsymbol{\Psi})q_1(\boldsymbol{Z}) \cdots (\boldsymbol{\Psi})q_{i-1}(\boldsymbol{Z}) \cdot (\boldsymbol{\Psi})q_{i+1}(\boldsymbol{Z}') \cdots (\boldsymbol{\Psi})q_k(\boldsymbol{Z}')\right\|_{\mathsf{op}} \cdot \left(\sum_{j=1}^m |q_i(z_j) - q_i(z_j')| \right) \\
&\leq \|(\boldsymbol{\Psi})\|_{\mathsf{op}} \|q_1(\boldsymbol{Z})\|_{\mathsf{op}} \cdots \|(\boldsymbol{\Psi})\|_{\mathsf{op}} \|q_k(\boldsymbol{Z}')\|_{\mathsf{op}} \cdot \left(\sum_{j=1}^m |q_i(z_j) - q_i(z_j')| \right) \\
&\overset{(a)}{\leq} \left(\prod_{j=1}^k \|q_j\|_\infty \right) \cdot \|q_i\|_{\mathsf{Lip}} \cdot \left(\sum_{j=1}^m |z_j - z_j'| \right) \\
&\leq \sqrt{m} \cdot C(\mathcal{A}) \cdot \|\boldsymbol{Z} - \boldsymbol{Z}'\|_{\mathsf{Fr}}.
\end{aligned}$$

In the step marked (a), we observed that: $\|(\boldsymbol{\Psi})\|_{\mathsf{op}} = \|\boldsymbol{U}(\overline{\boldsymbol{B}})\boldsymbol{U}^\mathsf{T}\|_{\mathsf{op}} \leq \max(|\kappa)|, |1-\kappa|) \leq 1$. Similarly, $\|q_j(\boldsymbol{Z})\|_{\mathsf{op}} \leq \|q_j\|_\infty \overset{\text{def}}{=} \sup_{\xi \in \mathbb{R}} |q_j(\xi)|$. We also recalled the functions q_i are assumed to be Lipchitz and denoted the Lipchitz constant of q_i by $\|q_i\|_{\mathsf{Lip}}$. Hence we obtain:

$$\left| \frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})}{m} - \frac{\mathsf{Tr}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}')}{m} \right| \leq \frac{k \cdot C(\mathcal{A})}{\sqrt{m}} \cdot \|\boldsymbol{Z} - \boldsymbol{Z}'\|_{\mathsf{Fr}}.$$

This concludes the proof of the lemma. □

4.10.2 Proof of Proposition 13

The proof of Proposition 13 is very similar to the proof of Proposition 12 and hence we will be brief in our arguments.

As discussed in the proof of Proposition 12, we will assume that alternating form is of Type 1. The other types are handled as outlined in Remark 21. Furthermore, in light of Lemma 17 we can further assume that all polynomials $p_i(\psi) = \psi$. Hence we assume that $\mathcal{A}$ is of the form:

$$\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}) = \boldsymbol{\Psi} q_1(\boldsymbol{Z}) \boldsymbol{\Psi} \cdots q_{k-1}(\boldsymbol{Z}) \boldsymbol{\Psi}.$$

The proof of Proposition 13 consists of various steps which will be organized as separate lemmas. We begin by recall that

$$z \sim \mathcal{N}\left(0, \frac{\boldsymbol{A}\boldsymbol{A}^\mathsf{T}}{\kappa}\right).$$

Define the event:

$$\mathcal{E} = \left\{ \max_{i \neq j} |(\boldsymbol{A}\boldsymbol{A}^\mathsf{T}|)_{ij} \leq \sqrt{\frac{2048 \cdot \log^3(m)}{m}}, \ \max_{i \in [m]} |(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{ii} - \kappa| \leq \sqrt{\frac{2048 \cdot \log^3(m)}{m}} \right\} \tag{4.33}$$

By Lemma 22 we know that $\mathbb{P}(\mathcal{E}^c) \to 0$ for both the subsampled Haar sensing and the subsampled Hadamard model. We define the normalized random vector $\widetilde{z}$ as:

$$\widetilde{z}_i = \frac{z_i}{\sigma_i}, \ \sigma_i^2 = \frac{(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{ii}}{\kappa}$$

Note that conditional on $\boldsymbol{A}$, $\widetilde{z}$ is a zero mean Gaussian vector with:

$$\mathbb{E}[\widetilde{z}_i^2|\boldsymbol{A}] = 1, \ \mathbb{E}[\widetilde{z}_i\widetilde{z}_j|\boldsymbol{A}] = \frac{(\boldsymbol{A}\boldsymbol{A}^\mathsf{T})_{ij}/\kappa}{\sigma_i\sigma_j}.$$

We define the diagonal matrix $\widetilde{\boldsymbol{Z}} = \mathrm{Diag}\left(\widetilde{z}\right)$.

Lemma 35. *We have,*

$$\lim_{m\to\infty} \frac{\mathbb{E}(z^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})z)^2}{m^2} = \lim_{m\to\infty} \frac{\mathbb{E}(\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi}, \widetilde{\boldsymbol{Z}})\widetilde{z})^2}{m^2}\mathbb{I}_{\mathcal{E}},$$

provided the latter limit exists.

The proof of this lemma is analogous the proof of Lemma 27 and is omitted. The advantage of Lemma 35 is that $\widetilde{z}_i \sim \mathcal{N}(0, 1)$ and on the event $\mathcal{E}$ the coordinates of $\widetilde{z}$ have weak correlations. Consequently, Mehler's Formula (Proposition 9) can be used to analyze the leading order term in $\mathbb{E}[\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi}, \widetilde{\boldsymbol{Z}})\widetilde{z}\,\mathbb{I}_{\mathcal{E}}]$. Before we do so, we do one additional preprocessing step.

Lemma 36. *We have,*

$$\lim_{m\to\infty} \frac{\mathbb{E}(\widetilde{z}^{\mathsf{T}}\mathcal{A}(\mathbf{\Psi}, \widetilde{\boldsymbol{Z}})\widetilde{z})^2}{m^2}\mathbb{I}_{\mathcal{E}} = \lim_{m\to\infty} \frac{\mathbb{E}\,\mathsf{Tr}(\mathcal{A}\cdot(\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2)\cdot\mathcal{A}\cdot(\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2))\mathbb{I}_{\mathcal{E}}}{m^2},$$

provided the latter limit exists.

Proof Sketch. Observe that we can write:

$$\begin{aligned}
(\widetilde{z}^{\mathsf{T}}\mathcal{A}\widetilde{z})^2 &= \mathsf{Tr}(\mathcal{A}\cdot\widetilde{z}\widetilde{z}^{\mathsf{T}}\cdot\mathcal{A}\cdot\widetilde{z}\widetilde{z}^{\mathsf{T}}) \\
&= \mathsf{Tr}(\mathcal{A}\cdot(\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2 + \widetilde{\boldsymbol{Z}}^2)\cdot\mathcal{A}\cdot(\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2 + \widetilde{\boldsymbol{Z}}^2)) \\
&= \mathsf{Tr}(\mathcal{A}\cdot(\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2)\cdot\mathcal{A}\cdot(\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2)) + \mathsf{Tr}(\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}\cdot\widetilde{z}\widetilde{z}^{\mathsf{T}}) + \mathsf{Tr}(\mathcal{A}\cdot\widetilde{z}\widetilde{z}^{\mathsf{T}}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}) \\
&\qquad - \mathsf{Tr}(\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2) \\
&= \mathsf{Tr}(\mathcal{A}\cdot(\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2)\cdot\mathcal{A}\cdot(\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2)) + 2\widetilde{z}^{\mathsf{T}}\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}\cdot\widetilde{z} - \mathsf{Tr}(\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2).
\end{aligned}$$

Next we note that:

$$|\widetilde{z}^{\mathsf{T}}\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}\cdot\widetilde{z}| \leq \|\widetilde{z}\|^2\cdot\|\mathcal{A}\|_{\mathsf{op}}^2\cdot\left(\max_{i\in[m]}|\widetilde{z}_i|^2\right) \leq O_P(m)\cdot O(1)\cdot O_P(\text{polylog}(m)),$$

Hence it can be shown that,

$$\frac{\mathbb{E}|\widetilde{z}^{\mathsf{T}}\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}\cdot\widetilde{z}|}{m^2} \to 0.$$

Similarly,

$$\begin{aligned}
|\mathsf{Tr}(\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2)| \leq m\|\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\cdot\mathcal{A}\cdot\widetilde{\boldsymbol{Z}}^2\|_{\mathsf{op}} &\leq m\|\mathcal{A}\|_{\mathsf{op}}^2\cdot\left(\max_{i\in[m]}|\widetilde{z}_i|^4\right) \\
&\leq O(m)\cdot O(1)\cdot O_P(\text{polylog}(m)),
\end{aligned}$$

and hence one expects that,

$$\frac{\mathbb{E}|\mathsf{Tr}(\mathcal{A} \cdot \widetilde{\boldsymbol{Z}}^2 \cdot \mathcal{A} \cdot \widetilde{\boldsymbol{Z}}^2)|}{m^2} \rightarrow 0.$$

We omit the detailed arguments. This concludes the proof of the lemma. □

Note that, so far, we have shown that:

$$\lim_{m\rightarrow\infty} \frac{\mathbb{E}(z^\mathsf{T} \mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})z)^2}{m^2} = \lim_{m\rightarrow\infty} \frac{\mathbb{E}\ \mathsf{Tr}(\mathcal{A} \cdot (\widetilde{z}\widetilde{z}^\mathsf{T} - \widetilde{\boldsymbol{Z}}^2) \cdot \mathcal{A} \cdot (\widetilde{z}\widetilde{z}^\mathsf{T} - \widetilde{\boldsymbol{Z}}^2))\mathbb{I}_{\mathcal{E}}}{m^2},$$

provided the latter limit exists. We now focus on analyzing the RHS. We expand

$$\mathsf{Tr}(\mathcal{A} \cdot (\widetilde{z}\widetilde{z}^\mathsf{T} - \widetilde{\boldsymbol{Z}}^2) \cdot \mathcal{A} \cdot (\widetilde{z}\widetilde{z}^\mathsf{T} - \widetilde{\boldsymbol{Z}}^2)) =$$
$$\sum_{\substack{a_{1:2k+2} \in [m] \\ a_1 \neq a_{2k+2} \\ a_{k+1} \neq a_{k+2}}} (\boldsymbol{\Psi})_{a_1,a_2} q_1(\widetilde{z}_{a_2}) \cdots (\boldsymbol{\Psi})_{a_k,a_{k+1}} \widetilde{z}_{a_{k+1}} \widetilde{z}_{a_{k+2}} (\boldsymbol{\Psi})_{a_{k+2},a_{k+3}} q_1(\widetilde{z}_{a_{k+3}}) \cdots (\boldsymbol{\Psi})_{a_{2k+1},a_{2k+2}} \widetilde{z}_{a_{2k+2}} \widetilde{z}_{a_1}.$$

This can be written compactly in terms of matrix moments (Definition 8) as follows: Let $\boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}(2k+2)$ denote the graph formed by combining two disconnected copies of the simple line graph on vertices $[1:k+1]$ and $[k+2:2k+2]$:

$$(\boldsymbol{\ell}_{k+1}^{\otimes 2})_{ij} = \begin{cases} 1: & |i-j| = 1,\ \{i,j\} \neq \{k+1, k+2\}, \\ 0: & \text{otherwise} \end{cases}.$$

Recall the notation for partitions introduced in Section 4.6.1. Observe that:

$$\{(a_1 \dots a_{2k+2}) \in [m]^{2k+2} :\ a_1 \neq a_{2k+2},\ a_{k+1} \neq a_{k+2}\} = \bigsqcup_{\pi \in \mathcal{P}_0([2k+2])} C(\pi),$$

where,

$$\mathcal{P}_0([2k+2]) \overset{\text{def}}{=} \{\pi \in \mathcal{P}(2k+2) : \pi(1) \neq \pi(2k+2),\ \pi(k+1) \neq \pi(k+2)\}.$$

Recalling Definition 8, we have,

$$(\boldsymbol{\Psi})_{a_1,a_2} \cdots (\boldsymbol{\Psi})_{a_k,a_{k+1}} (\boldsymbol{\Psi})_{a_{k+2},a_{k+3}} \cdots (\boldsymbol{\Psi})_{a_{2k+1},a_{2k+2}} = \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})$$

Hence,

$$\frac{\mathbb{E}\ \mathsf{Tr}(\mathcal{A} \cdot (\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2) \cdot \mathcal{A} \cdot (\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2))\mathbb{I}_{\mathcal{E}}}{m^2} = \\ \frac{1}{m^2} \sum_{\substack{\pi \in \mathcal{P}_0(2k+2) \\ \boldsymbol{a} \in C(\pi)}} \mathbb{E}\ \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a}) \cdot (\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) \cdots \widetilde{z}_{a_{k+1}} \widetilde{z}_{a_{k+2}} q_1(\widetilde{z}_{a_{k+3}}) \cdots \widetilde{z}_{a_{2k+2}}) \cdot \mathbb{I}_{\mathcal{E}}.$$

By the tower property,

$$\mathbb{E}\ \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a}) \cdot (\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) \cdots \widetilde{z}_{a_{k+1}} \widetilde{z}_{a_{k+2}} q_1(\widetilde{z}_{a_{k+3}}) \cdots \widetilde{z}_{a_{2k+2}}) \cdot \mathbb{I}_{\mathcal{E}} = \\ \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a}) \cdot \mathbb{E}[\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) \cdots \widetilde{z}_{a_{k+1}} \widetilde{z}_{a_{k+2}} q_1(\widetilde{z}_{a_{k+3}}) \cdots \widetilde{z}_{a_{2k+2}} | \boldsymbol{A}]\mathbb{I}_{\mathcal{E}}\right].$$

We will now use Mehler's formula (Proposition 9) to evaluate $\mathbb{E}[\cdots|\boldsymbol{A}]$ upto leading order. Note that some of the random variables $\widetilde{z}_{a_{1:2k+2}}$ are equal (as given by the partition π). Hence we group them together and recenter the resulting functions. The blocks corresponding to $a_1, a_{k+1}, a_{k+2}, a_{2k+2}$ need to be treated specially due to the presence of $\widetilde{z}_{a_1}, \widetilde{z}_{a_{k+1}}, \widetilde{z}_{a_{k+2}}, \widetilde{z}_{a_{2k+2}}$ in the above expectations. Hence, we introduce the following notations: We introduce the following notations:

$$\mathscr{F}_1(\pi) = \pi(1),\ \mathscr{L}_1(\pi) = \pi(k+1),\ \mathscr{F}_2(\pi) = \pi(k+2),\ \mathscr{L}_2(\pi) = \pi(2k+2)$$
$$\mathcal{S}(\pi) = \{i \in [1:2k+2] \backslash \{1, k+1, k+2, 2k+2\} : |\pi(i)| = 1\}.$$

We label all the remaining blocks of π as $\mathcal{V}_1, \mathcal{V}_2 \dots \mathcal{V}_{|\pi|-|\mathcal{S}(\pi)|-4}$. Hence the partition π is given by:

$$\pi = \mathscr{F}_1(\pi) \sqcup \mathscr{L}_1(\pi) \sqcup \mathscr{F}_2(\pi) \sqcup \mathscr{L}_2(\pi) \sqcup \left(\bigsqcup_{i \in \mathcal{S}(\pi)} \{i\} \right) \sqcup \left(\bigsqcup_{t=1}^{|\pi|-|\mathcal{S}(\pi)|-4} \mathcal{V}_i \right).$$

To simplify notation, we additionally define:

$$q_{k+1+i}(\xi) \overset{\text{def}}{=} q_i(\xi),\ i = 1, 2 \dots k-1.$$

Note that:

$$\widetilde{z}_{a_1} \widetilde{z}_{a_{k+1}} \widetilde{z}_{a_{k+2}} \widetilde{z}_{a_{2k+2}} \prod_{\substack{i=1 \\ i \neq k, k+1}}^{2k} q_i(\widetilde{z}_{a_{i+1}}) =$$

$$Q_{\mathscr{F}_1}(\widetilde{z}_{a_1}) Q_{\mathscr{L}_1}(\widetilde{z}_{a_{k+1}}) Q_{\mathscr{F}_2}(\widetilde{z}_{a_{k+2}}) Q_{\mathscr{L}_2}(\widetilde{z}_{a_{2k+2}}) \left(\prod_{i \in \mathcal{S}(\pi)} q_{i-1}(\widetilde{z}_{a_i}) \right) \prod_{i=1}^{|\pi|-|\mathcal{S}(\pi)|-4} (Q_{\mathcal{V}_i}(z_{a_{\mathcal{V}_i}}) + \mu_{\mathcal{V}_i}),$$

where,

$$\begin{aligned}
Q_{\mathscr{F}_1}(\xi) &= \xi \cdot \prod_{i \in \mathscr{F}_1(\pi), i \neq 1} q_{i-1}(\xi), \\
Q_{\mathscr{L}_1}(\xi) &= \xi \cdot \prod_{i \in \mathscr{L}_1(\pi), i \neq k+1} q_{i-1}(\xi), \\
Q_{\mathscr{F}_2}(\xi) &= \xi \cdot \prod_{i \in \mathscr{F}_2(\pi), i \neq k+2} q_{i-1}(\xi), \\
Q_{\mathscr{L}_2}(\xi) &= \xi \cdot \prod_{i \in \mathscr{L}_2(\pi), i \neq 2k+2} q_{i-1}(\xi), \\
\mu_{\mathcal{V}_i} &= \mathbb{E}_{\xi \sim \mathcal{N}(0,1)} \left[\prod_{j \in \mathcal{V}_i} q_{j-1}(\xi) \right], \\
Q_{\mathcal{V}_i}(\xi) &= \prod_{j \in \mathcal{V}_i} q_{j-1}(\xi) - \mu_{\mathcal{V}_i},
\end{aligned}$$

With this notation in place we can apply Mehler's formula. The result is summarized in the following lemma.

Lemma 37. *For any $\pi \in \mathcal{P}_0([2k+2])$ and any $\boldsymbol{a} \in C(\pi)$ we have,*

$$\mathbb{I}_{\mathcal{E}} \left| \mathbb{E}[\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) \cdots \widetilde{z}_{a_{k+1}} \widetilde{z}_{a_{k+2}} q_1(\widetilde{z}_{a_{k+3}}) \cdots \widetilde{z}_{a_{2k+2}} | \boldsymbol{A}] - \sum_{\boldsymbol{w} \in \mathcal{G}_2(\pi)} G(\boldsymbol{w}, \pi) \cdot \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) \right| \\ \leq C(\mathcal{A}) \cdot \left(\frac{\log^3(m)}{m\kappa^2} \right)^{\frac{3+|\mathcal{S}(\pi)|}{2}},$$

where, $\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ is the matrix moment as defined in Definition 8,

$$G(\boldsymbol{w}, \pi) = \frac{1}{\kappa^{\|\boldsymbol{w}\|} \boldsymbol{w}!} \left(\hat{Q}_{\mathscr{F}_1}(1) \hat{Q}_{\mathscr{L}_1}(1) \hat{Q}_{\mathscr{F}_2}(1) \hat{Q}_{\mathscr{L}_2}(1) \prod_{i \in \mathcal{S}(\pi)} \hat{q}_{i-1}(2) \right) \left(\prod_{i \in [|\pi| - |\mathcal{S}(\pi)| - 4]} \mu_{\mathcal{V}_i} \right)$$

$$\mathcal{G}_2(\pi) \overset{def}{=} \big\{ \boldsymbol{w} \in \mathcal{G}(2k+2) : \mathsf{d}_i(\boldsymbol{w}) = 1 \; \forall \; i \; \in \; \{1, k+1, k+2, 2k+2\}, \\ \mathsf{d}_i(\boldsymbol{w}) = 2 \; \forall \; i \; \in \; \mathcal{S}(\pi), \mathsf{d}_i(\boldsymbol{w}) = 0 \; \forall \; i \; \notin \; \{1, k+1, k+2, 2k+2\} \cup \mathcal{S}(\pi) \big\},$$

The proof of the lemma involves instantiating Mehler's formula for this situation and identifying the leading order term. Since the proof is analogous to the proof of Lemma 29 provided in Appendix 4.10.4, we omit it.

We return to our analysis of:

$$\frac{\mathbb{E}\, \mathsf{Tr}(\mathcal{A} \cdot (\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2) \cdot \mathcal{A} \cdot (\widetilde{z}\widetilde{z}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2)) \mathbb{I}_{\mathcal{E}}}{m^2} = \\ \frac{1}{m^2} \sum_{\substack{\pi \in \mathcal{P}_0(2k+2) \\ \boldsymbol{a} \in C(\pi)}} \mathbb{E}\, \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a}) \cdot (\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) \cdots \widetilde{z}_{a_{k+1}} \widetilde{z}_{a_{k+2}} q_1(\widetilde{z}_{a_{k+3}}) \cdots \widetilde{z}_{a_{2k+2}}) \cdot \mathbb{I}_{\mathcal{E}}.$$

We define the following subsets of $\mathcal{P}_0(2k+2)$ as:

$$\mathcal{P}_1([2k+2]) \overset{\text{def}}{=} \big\{\pi \in \mathcal{P}_0(2k+2) : \ |\pi(i)| = 1, \ \forall\, i \ \in \ \{1, k+1, k+2, 2k+2\}, \qquad |\pi(j)| \leq 2 \ \forall\, j \ \in \ [k+1]\big\}, \tag{4.35a}$$

$$\mathcal{P}_2([2k+2]) \overset{\text{def}}{=} \mathcal{P}_0([2k+2]) \backslash \mathcal{P}_1([2k+2]), \tag{4.35b}$$

and the error term which was controlled in Lemma 29:

$$\epsilon(\boldsymbol{\Psi}, \boldsymbol{a}) \overset{\text{def}}{=} \mathbb{I}_{\mathcal{E}}\left(\mathbb{E}[\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) \cdots \widetilde{z}_{a_{k+1}} \widetilde{z}_{a_{k+2}} q_1(\widetilde{z}_{a_{k+3}}) \cdots \widetilde{z}_{a_{2k+2}} |\boldsymbol{A}] - \sum_{\boldsymbol{w} \in \mathcal{G}_2(\pi)} G(\boldsymbol{w}, \pi) \cdot \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})\right)$$

.

With these definitions we consider the decomposition:

$$\frac{\mathbb{E}\ \mathsf{Tr}(\mathcal{A} \cdot (\widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2) \cdot \mathcal{A} \cdot (\widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^{\mathsf{T}} - \widetilde{\boldsymbol{Z}}^2))\mathbb{I}_{\mathcal{E}}}{m^2} = \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} \sum_{a \in \mathcal{C}(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_2(\pi)} G(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right] - \mathsf{I} + \mathsf{II} + \mathsf{III},$$

where,

$$\mathsf{I} = \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_0([2k+2])} \sum_{a \in \mathcal{C}(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_2(\pi)} G(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\mathbb{I}_{\mathcal{E}^c}\right],$$

$$\mathsf{II} = \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_0(2k+2])} \sum_{a \in \mathcal{C}(\pi)} \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\epsilon(\boldsymbol{\Psi}, \boldsymbol{a})\mathbb{I}_{\mathcal{E}}\right],$$

$$\mathsf{III} = \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_2([2k+2])} \sum_{a \in \mathcal{C}(\pi)} \sum_{\boldsymbol{w} \in \mathcal{G}_2(\pi)} G(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right].$$

We will show that $\mathsf{I}, \mathsf{II}, \mathsf{III} \rightarrow 0$. Showing this involves the following components:

1. Bounds on matrix moments $\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right]$ which have been developed in Lemma 19.

2. Controlling the size of the set $|C(\pi)|$ (since we sum over $\boldsymbol{a} \in C(\pi)$ in the above terms). Since,

$$|C(\pi)| = m(m-1)\cdots(m-|\pi|+1) \asymp m^{|\pi|},$$

we need to develop bounds on $|\pi|$. This is done in the following lemma. In contrast, the sums over $\pi \in \mathcal{P}_0([2k+2])$ and $\boldsymbol{w} \in \mathcal{G}_1(\pi)$ are not a cause of concern since $|\mathcal{P}_0([2k+2])|, |\mathcal{G}_1(\pi)|$ depend only on k (which is held fixed) and not on m.

Lemma 38. *For any $\pi \in \mathcal{P}_1([2k+2])$ we have,*

$$|\pi| = \frac{2k+6+|\mathcal{S}(\pi)|}{2} \implies |C(\pi)| \leq m^{\frac{2k+6+|\mathcal{S}(\pi)|}{2}}.$$

For any $\pi \in \mathcal{P}_2([2k+2])$, we have,

$$|\pi| \leq \frac{2k+5+|\mathcal{S}(\pi)|}{2} \implies |C(\pi)| \leq m^{\frac{2k+5+|\mathcal{S}(\pi)|}{2}}.$$

Proof. Consider any $\pi \in \mathcal{P}_0([2k+2])$. Recall that the disjoint blocks of $|\pi|$ were given by:

$$\pi = \mathscr{F}_1(\pi) \sqcup \mathscr{L}_1(\pi) \sqcup \mathscr{F}_2(\pi) \sqcup \mathscr{L}_2(\pi) \sqcup \left(\bigsqcup_{i \in \mathcal{S}(\pi)} \{i\} \right) \sqcup \left(\bigsqcup_{t=1}^{|\pi|-|\mathcal{S}(\pi)|-4} \mathcal{V}_i \right).$$

Hence,

$$2k+2 = |\mathscr{F}_1(\pi)| + |\mathscr{F}_2(\pi)| + |\mathscr{L}_1(\pi)| + |\mathscr{L}_2(\pi)| + |\mathcal{S}(\pi)| + \sum_{t=1}^{|\pi|-|\mathcal{S}(\pi)|-4} |\mathcal{V}_i|.$$

Note that:

$$|\mathscr{F}_1(\pi)| \geq 1 \qquad \text{(Since } 1 \in \mathscr{F}_1(\pi)\text{)} \tag{4.36a}$$

$$|\mathscr{F}_2(\pi)| \geq 1 \qquad \text{(Since } k+2 \in \mathscr{F}_2(\pi)\text{)} \tag{4.36b}$$

$$|\mathscr{L}_1(\pi)| \geq 1 \qquad \text{(Since } k+1 \in \mathscr{L}_1(\pi)\text{)} \tag{4.36c}$$

$$|\mathscr{L}_2(\pi)| \geq 1 \qquad \text{(Since } 2k+2 \in \mathscr{L}_1(\pi)\text{)} \tag{4.36d}$$

$$|\mathcal{V}_i| \geq 2 \qquad \text{(Since } \mathcal{V}_i \text{ are not singletons).} \tag{4.36e}$$

Hence,

$$2k+2 \geq 4 + 2|\pi| - |\mathcal{S}(\pi)| - 8,$$

which implies,

$$|\pi| \leq \frac{2k+6+|\mathcal{S}(\pi)|}{2}, \tag{4.37}$$

and hence,

$$|C(\pi)| \leq m^{|\pi|} \leq m^{\frac{2k+6+|\mathcal{S}(\pi)|}{2}}.$$

Finally observe that:

1. For any $\pi \in \mathcal{P}_2([2k+2])$ each of the inequalities in (4.36) are exactly tight by the definition of $\mathcal{P}_1([k+1])$ in (4.35), and hence,

$$|\pi| = \frac{2k+6+|\mathcal{S}(\pi)|}{2}.$$

2. For any $\pi \in \mathcal{P}_2([2k+2])$, one of the inequalities in (4.36) must be strict (see (4.35)). Hence,

when $\pi \in \mathcal{P}_2([k+1])$ we have the improved bound:

$$|\pi| \leq \frac{2k+5+|\mathcal{S}(\pi)|}{2}.$$

This proves the claims of the lemma. □

We will now show that $\mathsf{I}, \mathsf{II}, \mathsf{III} \to 0$.

Lemma 39. *We have,*

$$\mathsf{I} \to 0,\ \mathsf{II} \to 0,\ \mathsf{III} \to 0 \ \textit{as } m \to \infty,$$

and hence,

$$\lim_{m\to\infty} \frac{\mathbb{E}(\boldsymbol{z}^{\mathsf{T}}\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z})\boldsymbol{z})^2}{m^2} = \lim_{m\to\infty} \frac{1}{m^2} \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{a\in\mathcal{C}(\pi)} \sum_{\boldsymbol{w}\in\mathcal{G}_2(\pi)} G(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right],$$

provided the latter limit exists.

Proof. First note that for any $\boldsymbol{w} \in \mathcal{G}_1(\pi)$, we have,

$$\|\boldsymbol{w}\| = \frac{1}{2}\sum_{i=1}^{2k+2} \mathsf{d}_i(\boldsymbol{w}) = \frac{1+1+1+1+2|\mathcal{S}(\pi)|}{2} = 2 + |\mathcal{S}(\pi)| \ \text{(See Lemma 37).}$$

Furthermore recalling the definition of $\boldsymbol{\ell}_{k+1}^{\otimes 2}$, $\|\boldsymbol{\ell}_{k+1}^{\otimes 2}\| = 2k$. Now we apply Lemma 19 to obtain:

$$\begin{aligned}
|\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\mathbb{I}_{\mathcal{E}^c}\right]| &\le \sqrt{\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, 2\boldsymbol{w} + 2\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right]}\sqrt{\mathbb{P}(\mathcal{E}^c)} \\
&\le \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+2+2k}{2}} \cdot \sqrt{\mathbb{P}(\mathcal{E}^c)}, \\
&\overset{(a)}{\le} \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+2+2k}{2}} \cdot \frac{C_k}{m}. \\
\mathbb{E}|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})| &\le \left(\frac{C_k \log^2(m)}{m}\right)^{k}, \\
\mathbb{E}\left[|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}, \pi, \boldsymbol{a})|\right] &\le \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+2+2k}{2}}
\end{aligned}$$

In the step marked (a) we used Lemma 22. Further recall that by Lemma 29 we have,

$$|\epsilon(\boldsymbol{\Psi}, \boldsymbol{a})| \le C(\mathcal{A}) \cdot \left(\frac{\log^3(m)}{m\kappa^2}\right)^{\frac{3+|\mathcal{S}(\pi)|}{2}}.$$

Using these estimates, we obtain,

$$
\begin{aligned}
|\mathsf{I}| &\leq \frac{C(\mathcal{A})\cdot}{m^2} \cdot \sum_{\pi:\mathcal{P}_0([2k+2])} |C(\pi)| \cdot \left(\frac{C_k \log^2(m)}{m} \right)^{\frac{|\mathcal{S}(\pi)|+2+2k}{2}} \cdot \frac{C_k}{m} \\
&\leq \frac{C(\mathcal{A})\cdot}{m^2} \cdot \sum_{\pi:\mathcal{P}_0([2k+2])} m^{\frac{2k+6+|\mathcal{S}(\pi)|}{2}} \cdot \left(\frac{C_k \log^2(m)}{m} \right)^{\frac{|\mathcal{S}(\pi)|+2+2k}{2}} \cdot \frac{C_k}{m} \\
&= O\left(\frac{\text{polylog}(m)}{m} \right) \\
|\mathsf{II}| &\leq \frac{C(\mathcal{A})}{m^2} \cdot \left(\frac{C_k \log^2(m)}{m} \right)^{k} \cdot \sum_{\pi:\mathcal{P}_0([2k+2])} |C(\pi)| \cdot \left(\frac{\log^3(m)}{m\kappa^2} \right)^{\frac{3+|\mathcal{S}(\pi)|}{2}} \\
&\leq \frac{C(\mathcal{A})}{m^2} \cdot \left(\frac{C_k \log^2(m)}{m} \right)^{k} \cdot \sum_{\pi:\mathcal{P}_0([2k+2])} m^{\frac{2k+6+|\mathcal{S}(\pi)|}{2}} \cdot \left(\frac{\log^3(m)}{m\kappa^2} \right)^{\frac{3+|\mathcal{S}(\pi)|}{2}} \\
&= O\left(\frac{\text{polylog}(m)}{\sqrt{m}} \right) \\
|\mathsf{III}| &\leq \frac{C(\mathcal{A})\cdot}{m^2} \cdot \sum_{\pi:\mathcal{P}_2([2k+2])} |C(\pi)| \cdot \left(\frac{C_k \log^2(m)}{m} \right)^{\frac{|\mathcal{S}(\pi)|+1+k}{2}} \\
&\leq \frac{C(\mathcal{A})\cdot}{m^2} \cdot \sum_{\pi:\mathcal{P}_2([2k+2])} m^{\frac{2k+5+|\mathcal{S}(\pi)|}{2}} \cdot \left(\frac{C_k \log^2(m)}{m} \right)^{\frac{|\mathcal{S}(\pi)|+2+2k}{2}} \\
&= O\left(\frac{\text{polylog}(m)}{\sqrt{m}} \right).
\end{aligned}
$$

This concludes the proof of this lemma. □

Next, we consider the decomposition:

$$
\begin{aligned}
&\frac{1}{m^2} \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{a\in C(\pi)} \sum_{\boldsymbol{w}\in\mathcal{G}_2(\pi)} G(\boldsymbol{w},\pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a}) \right] = \\
&\frac{1}{m^2} \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_2(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a\in\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)} G(\boldsymbol{w},\pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a}) \right] + \mathsf{IV} + \mathsf{V},
\end{aligned}
$$

where,

$$\mathsf{IV} \overset{\text{def}}{=} \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} \sum_{a \in \mathcal{C}(\pi)} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_2(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2} \notin \mathcal{G}_{\mathsf{DA}}(\pi)}} G(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right],$$

$$\mathsf{V} \overset{\text{def}}{=} \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_2(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{C}(\pi) \backslash \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)} G(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right].$$

Lemma 40. *We have,* $\mathsf{IV} \rightarrow 0, \mathsf{V} \rightarrow 0$ *as* $m \rightarrow \infty$*, and hence,*

$$\lim_{m \to \infty} \frac{\mathbb{E}(\boldsymbol{z}^{\mathsf{T}} \mathcal{A} \boldsymbol{z})^2}{m^2} =$$
$$\lim_{m \to \infty} \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_2(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)} G(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right],$$

provided the latter limit exists.

Proof. We will prove this in two steps.

Step 1: $\mathsf{IV} \rightarrow 0$**.** We consider the two sensing models separately:

1. Subsampled Hadamard Sensing: In this case, Proposition 11 tells us that if $\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2} \notin \mathcal{G}_{\mathsf{DA}}(\pi)$, then,
$$\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right] = 0$$
and hence $\mathsf{IV} = 0$.

2. Subsampled Haar Sensing: Observe that, since $\|\boldsymbol{w}\| + \|\boldsymbol{\ell}_{k+1}^{\otimes 2}\| = 2 + |\mathcal{S}(\pi)| + 2k$, we have,
$$\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right] = \frac{\mathbb{E}\left[\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right]}{m^{\frac{2+|\mathcal{S}(\pi)|+2k}{2}}}.$$

By Proposition 10 we know that,

$$\left| \mathbb{E}\left[\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a}) \right] - \prod_{\substack{s,t \in [|\pi|] \\ s \leq t}} \mathbb{E}\left[Z_{st}^{W_{st}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)} \right] \right| \leq \frac{K_1 \log^{K_2}(m)}{m^{\frac{1}{4}}},$$

$\forall\ m \geq K_3$, where K_1, K_2, K_3 are universal constants depending only on k. Note that since $\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2} \notin \mathcal{G}_{\mathsf{DA}}(\pi)$, must have some $s \in [|\pi|]$ such that:

$$W_{ss}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi) \geq 1.$$

Recall that, $\mathsf{d}_i(\boldsymbol{w}) = 0$ for any $i \notin \{1, k+1, k+2, 2k+2\} \cup \mathcal{S}(\pi)$ (since $\boldsymbol{w} \in \mathcal{G}_2(\pi)$) and furthermore, $|\pi(i)| = 1 \forall\ i \in \{1, k+1, k+2, 2k+2\} \cup \mathcal{S}(\pi)$ (since $\pi \in \mathcal{P}_1(2k+2)$). Hence, we have $\boldsymbol{w} \in \mathcal{G}_{\mathsf{DA}}(\pi)$ and in particular, $W_{ss}(\boldsymbol{w}, \pi) = 0$. Consequently, we must have $W_{ss}(\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi) \geq 1$. Recall the definition of $\boldsymbol{\ell}_{k+1}^{\otimes 2}$, since $W_{ss}(\boldsymbol{\ell}_{k+1}, \pi) \geq 1$ we must have that for some $i \in [2k+2]$, we have, $\pi(i) = \pi(i+1) = \mathcal{V}_s$. However, since $\pi \in \mathcal{P}_1(2k+2)$, $|\mathcal{V}_s| \leq 2$, and hence $\mathcal{V}_s = \{i, i+1\}$. This means that $W_{ss}(\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi) = 1 = W_{ss}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)$. Consequently since $\mathbb{E}Z_{ss} = 0$, we have,

$$\prod_{\substack{s,t \in [|\pi|] \\ s \leq t}} \mathbb{E}\left[Z_{st}^{W_{st}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)} \right] = 0,$$

or,

$$\left| \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a}) \right] \right| = \frac{\operatorname{polylog}(m)}{m^{\frac{2+|\mathcal{S}(\pi)|+2k}{2} + \frac{1}{4}}}.$$

Recalling Lemma 38,

$$|C(\pi)| \leq m^{|\pi|} \leq m^{\frac{2k+6+|\mathcal{S}(\pi)|}{2}},$$

we obtain,

$$|\mathsf{IV}| \leq \frac{C(\mathcal{A})}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} |C(\pi)| \cdot \frac{\text{polylog}(m)}{m^{\frac{2+|\mathcal{S}(\pi)|+2k}{2}+\frac{1}{4}}} = O\left(\frac{\text{polylog}(m)}{m^{\frac{1}{4}}}\right) \to 0.$$

Step 2: $\mathsf{V} \to 0$. Using Lemma 21, we know that

$$|C(\pi) \backslash \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)| \leq O(m^{|\pi|-1})$$

In Lemma 38, we showed that for any $\pi \in \mathcal{P}_1([k+1])$,

$$|\pi| = \frac{2k+6+|\mathcal{S}(\pi)|}{2}.$$

Hence,

$$|C(\pi) \backslash \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)| \leq O(m^{\frac{2k+4+|\mathcal{S}(\pi)|}{2}}).$$

We already know from Lemma 19 that,

$$|\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right]| \leq \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{\|\boldsymbol{w}\|+\|\boldsymbol{\ell}_{k+1}^{\otimes 2}\|}{2}} \leq \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+2+2k}{2}},$$

This gives us:

$$\begin{aligned} |\mathsf{V}| &\leq \frac{C}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_2(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} |C(\pi) \backslash \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)| \left(\frac{C_k \log^2(m)}{m}\right)^{\frac{|\mathcal{S}(\pi)|+2+2k}{2}} \\ &= O\left(\frac{\text{polylog}(m)}{m}\right) \end{aligned}$$

which goes to zero as claimed.

This conclude the proof of the lemma. □

So far we have shown that:

$$\lim_{m\to\infty} \frac{\mathbb{E}(\boldsymbol{z}^{\mathsf{T}}\mathcal{A}\boldsymbol{z})^2}{m^2} =$$

$$\lim_{m\to\infty} \frac{1}{m^2} \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_2(\pi)\\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{\boldsymbol{a}\in\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)} G(\boldsymbol{w},\pi)\cdot\mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi},\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi,\boldsymbol{a})\right].$$

provided the latter limit exists. In the following lemma we explicitly calculate the limit on the RHS and hence show that it exists and is same for the subsampled Haar and subsampled Hadamard sensing models.

Lemma 41. *For both the subsampled Haar sensing and Hadamard sensing model, we have,*

$$\lim_{m\to\infty} \frac{\mathbb{E}(\boldsymbol{z}^{\mathsf{T}}\mathcal{A}\boldsymbol{z})^2}{m^2} = \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_2(\pi)\\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} G(\boldsymbol{w},\pi)\cdot\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi),$$

where,

$$\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi) \overset{def}{=} \prod_{\substack{s,t\in[|\pi|]\\ s<t}} \mathbb{E}\left[Z^{W_{st}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)}\right],\ Z\sim\mathcal{N}\left(0,\kappa(1-\kappa)\right).$$

Proof. By Propositions 11 (for the subsampled Hadamard model) and 10 (for the subsampled Haar model) we know that, if $\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi),\ \boldsymbol{a}\in\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)$, we have,

$$\mathcal{M}(\sqrt{m}\boldsymbol{\Psi},\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi,\boldsymbol{a}) = \mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)+\epsilon(\boldsymbol{w},\pi,\boldsymbol{a}),$$

where

$$|\epsilon(\boldsymbol{w},\pi,\boldsymbol{a})| \leq \frac{K_1\log^{K_2}(m)}{m^{\frac{1}{4}}},\ \forall\, m\geq K_3,$$

for some constants K_1, K_2, K_3 depending only on k. Hence, we can consider the decomposition:

$$\frac{1}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_2(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)} G(\boldsymbol{w}, \pi) \cdot \mathbb{E}\left[\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi, \boldsymbol{a})\right]$$

$$= \mathsf{VI} + \mathsf{VII},$$

where,

$$\mathsf{VI} \overset{\text{def}}{=} \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_2(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)} G(\boldsymbol{w}, \pi) \cdot \frac{\mu(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)}{m^{\frac{2+\mathcal{S}(\pi)+2k}{2}}},$$

$$\mathsf{VII} \overset{\text{def}}{=} \frac{1}{m^2} \sum_{\pi \in \mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w} \in \mathcal{G}_2(\pi) \\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a \in \mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)} G(\boldsymbol{w}, \pi) \cdot \frac{\epsilon(\boldsymbol{w}, \pi, \boldsymbol{a})}{m^{\frac{2+\mathcal{S}(\pi)+2k}{2}}}$$

We can upper bound $|\mathsf{VII}|$ as follows:

$$|\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)| \leq |\mathcal{C}(\pi)| \leq m^{\frac{2k+6+|\mathcal{S}(\pi)|}{2}},$$

$$|\mathsf{VII}| \leq \frac{C(\mathscr{A})}{m^2} \cdot C_k \cdot |\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)| \cdot \frac{1}{m^{\frac{2+|\mathcal{S}(\pi)|+2k}{2}}} \cdot \frac{K_1 \log^{K_2}(m)}{m^{\frac{1}{4}}}$$

$$= O\left(\frac{\text{polylog}(m)}{m^{\frac{1}{4}}}\right) \to 0.$$

We can compute:

$$\begin{aligned}
\lim_{m\to\infty} (\mathsf{VI}) &= \lim_{m\to\infty} \frac{1}{m^2} \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_2(\pi)\\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} \sum_{a\in\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)} G(\boldsymbol{w},\pi)\cdot\frac{\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)}{m^{\frac{2+\mathcal{S}(\pi)+2k}{2}}} \\
&= \lim_{m\to\infty} \frac{1}{m^2} \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_2(\pi)\\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} G(\boldsymbol{w},\pi)\cdot\frac{\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)}{m^{\frac{2+\mathcal{S}(\pi)+2k}{2}}}\cdot|\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)| \\
&= \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_2(\pi)\\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} G(\boldsymbol{w},\pi)\cdot\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)\cdot\frac{m^{|\pi|}}{m^{\frac{6+\mathcal{S}(\pi)+2k}{2}}}\cdot\frac{|\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)|}{m^{|\pi|}} \\
&\overset{\text{(a)}}{=} \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_2(\pi)\\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} G(\boldsymbol{w},\pi)\cdot\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)\cdot\frac{|\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)|}{m^{|\pi|}} \\
&\overset{\text{(b)}}{=} \sum_{\pi\in\mathcal{P}_1([2k+2])} \sum_{\substack{\boldsymbol{w}\in\mathcal{G}_2(\pi)\\ \boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}\in\mathcal{G}_{\mathsf{DA}}(\pi)}} G(\boldsymbol{w},\pi)\cdot\mu(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi).
\end{aligned}$$

In the step marked (a) we used the fact that $|\pi| = (6+|\mathcal{S}(\pi)|+2k)/2$ for any $\pi\in\mathcal{P}_1([2k+2])$ (Lemma 38) and in step (b) we used Lemma 21 ($|\mathcal{L}_{\mathsf{CF}}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2},\pi)|/m^{|\pi|}\to 1$). This proves the claim of the lemma and Proposition 13. □

We can actually significantly simply the combinatorial sum obtained in Lemma 41 which we do so in the following lemma.

Lemma 42. *For both the subsampled Haar sensing and Hadamard sensing models, we have,*

$$\lim_{m\to\infty}\frac{\mathbb{E}(\boldsymbol{z}^{\mathsf{T}}\mathcal{A}\boldsymbol{z})^2}{m^2} = (1-\kappa)^{2k}\cdot\prod_{i=1}^{k-1}\hat{q}_i^2(2).$$

In particular, Proposition 13 holds.

Proof. We claim that the only partition with a non-zero contribution is:

$$\pi = \bigsqcup_{i=1}^{2k+2}\{i\}.$$

In order to see this suppose π is not entirely composed of singleton blocks. Define:

$$i_\star \overset{\text{def}}{=} \min\{i \in [2k+2] : |\pi(i)| > 1\}.$$

Note $i_\star > 1$ since we know that $|\pi(1)| = |\mathscr{F}_1(\pi)| = 1$ for any $\pi \in \mathcal{P}_1(2k+2)$. Since $\pi \in \mathcal{P}_1([2k+2])$ we must have $|\pi(i_\star)| = 2$, hence denote:

$$\pi(i_\star) = \{i_\star, j_\star\}.$$

for some $j_\star > i_\star+1$ ($i_\star \leq j_\star$ since it is the first index which is not in a singleton block, and $j_\star \neq i_\star+1$ since otherwise $\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}$ will not be disassortative. Similarly we know that $i_\star, j_\star \neq k+1, k+2, 2k+2$ because $|\pi(k+1)| = |\pi(k+2)| = |\pi(2k+2)| = 1$ since $\pi \in \mathcal{P}_1([2k+2])$. Let us label the first few blocks of π as:

$$\mathcal{V}_1 = \{1\},\ \mathcal{V}_2 = \{2\}, \dots \mathcal{V}_{i_\star-1} = \{i_\star - 1\},\ \mathcal{V}_{i_\star} = \{i_\star, j_\star\}.$$

Next we compute:

$$\begin{aligned}
W_{i_\star-1,i_\star}(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi) &= W_{i_\star-1,i_\star}(\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi) + W_{i_\star-1,i_\star}(\boldsymbol{w}, \pi) \\
&\overset{(a)}{=} W_{i_\star-1,i_\star}(\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi) \\
&\overset{(b)}{=} \mathbf{1}_{i_\star-1\in\mathcal{V}_{i_\star-1}} + \mathbf{1}_{i_\star+1\in\mathcal{V}_{i_\star-1}} + \mathbf{1}_{j_\star-1\in\mathcal{V}_{i_\star-1}} + \mathbf{1}_{j_\star+1\in\mathcal{V}_{i_\star-1}} \\
&\overset{(c)}{=} \mathbf{1}_{i_\star-1=i_\star-1} + \mathbf{1}_{i_\star+1=i_\star-1} + \mathbf{1}_{j_\star-1=i_\star-1} + \mathbf{1}_{j_\star+1=i_\star-1} \\
&\overset{(d)}{=} 1.
\end{aligned}$$

In the step marked (a), we used the fact that since $\boldsymbol{w} \in \mathcal{G}_2(\pi)$ and $|\pi(i_\star)| = |\pi(j_\star)| = 2$, we must have $d_{i_\star}(\boldsymbol{w}) = d_{j_\star}(\boldsymbol{w}) = 0$ and $W_{i_\star-1,i_\star}(\boldsymbol{w}, \pi) = 0$. In the step marked (b) we used the definition of $\boldsymbol{\ell}_{k+1}^{\otimes 2}$. In the step marked (c) we used the fact that $\mathcal{V}_{i_\star-1} = \{i_{\star-1}\}$. In the step marked (d) we used the fact that $j_\star > i_\star + 1$.

Hence we have shown that for any $\pi \neq \sqcup_{i=1}^{2k+2}\{i\}$, we have

$$\mu(\boldsymbol{w}, \pi) = 0 \; \forall \; \boldsymbol{w} \text{ such that } \boldsymbol{w} \in \mathcal{G}_2(\pi), \; \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}_{\mathsf{DA}}(\pi).$$

Next, let $\pi = \sqcup_{i=1}^{2k+2}\{i\}$. We observe for any $\boldsymbol{w}$ such that $\boldsymbol{w} \in \mathcal{G}_2(\pi), \; \boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2} \in \mathcal{G}_{\mathsf{DA}}(\pi)$, we have,

$$\begin{aligned}\mu(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi) &= \prod_{\substack{s,t \in [|\pi|] \\ s<t}} \mathbb{E}\left[Z^{W_{st}(\boldsymbol{w}+\boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi)}\right], \; Z \sim \mathcal{N}\left(0, \kappa(1-\kappa)\right) \\ &= \prod_{\substack{i,j \in [2k+2] \\ i<j}} \mathbb{E}\left[Z^{w_{ij} + (\ell_{k+1})_{ij}, \pi)}\right], \; Z \sim \mathcal{N}\left(0, \kappa(1-\kappa)\right)\end{aligned}$$

Note that since $\mathbb{E}Z = 0$, for $\mu(\boldsymbol{w} + \boldsymbol{\ell}_{k+1}^{\otimes 2}, \pi) \neq 0$ we must have:

$$w_{ij} \geq (\boldsymbol{\ell}_{k+1}^{\otimes 2})_{ij}, \; \forall \; i, j \; \in \; [2k+2].$$

However since $\boldsymbol{w} \in \mathcal{G}_2(\pi)$ we have,

$$\begin{aligned}&\mathsf{d}_1(\boldsymbol{w}) = \mathsf{d}_{k+1}(\boldsymbol{w}) = \mathsf{d}_{k+2}(\boldsymbol{w}) = \mathsf{d}_{2k+2}(\boldsymbol{w}) = 1, \\ &\mathsf{d}_i(\boldsymbol{w}) = 2 \; \forall \; i \; \in \; [2k+2] \backslash \{1, k+1, k+2, 2k+2\},\end{aligned}$$

hence $\boldsymbol{w} = \boldsymbol{\ell}_{k+1}^{\otimes 2}$. Hence, recalling the formula for $g(\boldsymbol{w}, \pi)$ from Lemma 29 we obtain:

$$\lim_{m \to \infty} \frac{\mathbb{E}(\boldsymbol{z}^{\mathsf{T}} \mathcal{A} \boldsymbol{z})^2}{m^2} = (1-\kappa)^{2k} \cdot \prod_{i=1}^{k-1} \hat{q}_i^2(2).$$

This proves the statement of the lemma and also Proposition 12 (see Remark 21 regarding how the analysis extends to other types). □

4.10.3 Proofs from Section 4.6.4

Proof of Lemma 19

Proof of Lemma 19. Recall that,

$$\begin{aligned}
\mathbb{E}|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})| &= \mathbb{E} \prod_{\substack{i,j \in [k] \\ i<j}} |\Psi_{a_i,a_j}^{w_{ij}}| \\
&\overset{(a)}{\leq} \sum_{\substack{i,j \in [k] \\ i<j}} \frac{w_{ij}}{\|\boldsymbol{w}\|} \mathbb{E}|\Psi_{a_i,a_j}^{\|\boldsymbol{w}\|_1}| \\
&\leq \max_{i,j \in [m]} \mathbb{E}|\Psi_{ij}|^{\|\boldsymbol{w}\|},
\end{aligned}$$

where step (a) follows from the AM-GM inequality. We now consider the subsampled Haar and Hadamard cases separately.

Hadamard Case: By Lemma 18, Ψ_{ij} is subgaussian with with variance proxy bounded by C/m for some universal constant C. Hence,

$$\mathbb{E}|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})| \leq \left(\frac{C\|\boldsymbol{w}\|}{m}\right)^{\frac{\|\boldsymbol{w}\|}{2}}.$$

Haar Case: By Lemma 18, conditional on $\boldsymbol{O}$, Ψ_{ij} is subgaussian with variance proxy $Cm\|\boldsymbol{o}_i\|_\infty^2 \|\boldsymbol{o}_j\|_\infty^2$. Hence,

$$\begin{aligned}
\mathbb{E}|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})| &\leq \max_{i,j \in [m]} \mathbb{E}|\Psi_{ij}|^{\|\boldsymbol{w}\|} \\
&= \max_{i,j \in [m]} \mathbb{E}[\mathbb{E}[|\Psi_{ij}|^{\|\boldsymbol{w}\|}|\boldsymbol{O}]] \\
&\leq \max_{i,j \in [m]} (C\|\boldsymbol{w}\|m)^{\frac{\|\boldsymbol{w}\|}{2}} \mathbb{E}\left[\|\boldsymbol{o}_i\|_\infty^{\|\boldsymbol{w}\|} \|\boldsymbol{o}_j\|_\infty^{\|\boldsymbol{w}\|}\right] \\
&\leq \max_{i,j \in [m]} (C\|\boldsymbol{w}\|m)^{\frac{\|\boldsymbol{w}\|}{2}} \left(\mathbb{E}\|\boldsymbol{o}_i\|_\infty^{2\|\boldsymbol{w}\|} + \mathbb{E}\|\boldsymbol{o}_j\|_\infty^{2\|\boldsymbol{w}\|}\right).
\end{aligned}$$

Note that $\boldsymbol{o}_i \overset{\mathrm{d}}{=} \boldsymbol{o}_j \overset{\mathrm{d}}{=} \boldsymbol{u} \sim \mathrm{Unif}\left(\mathbb{S}_{m-1}\right)$. Applying Fact 6 gives us,

$$\mathbb{E}|\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})| \leq \left(\sqrt{\frac{C\|\boldsymbol{w}\| \log^2(m)}{m}}\right)^{\|\boldsymbol{w}\|}.$$

□

Proofs of Propositions 10 and 11

This section is dedicated to the proof of Propositions 10 and 11. We consider the following general setup. Let $\boldsymbol{v}_1, \boldsymbol{v}_2 \cdots, \boldsymbol{v}_m$ be fixed vectors in $\mathbb{R}^d$ for a fixed $d \in \mathbb{N}$. Define the statistic:

$$\boldsymbol{T} = \sqrt{m} \sum_{i=1}^{m} \overline{B}_{ii} \boldsymbol{v}_i,$$

where $\overline{\boldsymbol{B}}$ denotes a diagonal matrix whose n diagonal entries are set to $1-\kappa$ uniformly at random and the remaining $m-n$ are set to $-\kappa$.

Analogously, we define the statistic:

$$\hat{\boldsymbol{T}} = \sqrt{m} \sum_{i=1}^{m} \hat{B}_{ii} \boldsymbol{v}_i,$$

where,

$$\hat{B}_{ii} \overset{\text{i.i.d.}}{\sim} \begin{cases} 1-\kappa : & \text{with prob. } \kappa \\ -\kappa : & \text{with prob. } 1-\kappa \end{cases}.$$

As in the proof of Lemma 18 we $\overline{\boldsymbol{B}}$ and $\hat{\boldsymbol{B}}$ in the same probability space as follows:

1. We first sample $\overline{\boldsymbol{B}}$. Let $S = \{i \in [m] : \overline{B}_{ii} = 1-\kappa\}$
2. Next sample $N \sim \mathsf{Binom}(m, \kappa)$.
3. Sample a subset $\hat{S} \subset [m]$ with $|\hat{S}| = N$ as follows:

- If $N \leq n$, then set $\hat{S}$ to be a uniformly random subset of S of size N.
- If $N > n$ first sample a uniformly random subset A of S^c of size $N - n$ and set $\hat{S} = S \cup A$

4. Set $\hat{\boldsymbol{B}}$ as follows:

$$\hat{B}_{ii} = \begin{cases} -\kappa & : i \notin \hat{S} \\ 1 - \kappa & : i \in \hat{S}. \end{cases}.$$

We stack the vectors $\boldsymbol{v}_{1:m}$ along the rows of a matrix $\boldsymbol{V} \in \mathbb{R}^{m \times d}$ and refer to the columns of $\boldsymbol{V}$ as $\boldsymbol{V}_1, \boldsymbol{V}_2 \cdots \boldsymbol{V}_d$:

$$\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2 \cdots \boldsymbol{V}_d] = \begin{bmatrix} \boldsymbol{v}_1^\mathsf{T} \\ \boldsymbol{v}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{v}_m^\mathsf{T} \end{bmatrix}.$$

Lastly we introduce the matrix $\hat{\boldsymbol{\Sigma}} \in \mathbb{R}^{d \times d}$:

$$\hat{\boldsymbol{\Sigma}} \overset{\text{def}}{=} \mathbb{E}[\hat{\boldsymbol{T}}\hat{\boldsymbol{T}}^\mathsf{T} | V] = m\kappa(1 - \kappa)\boldsymbol{V}^\mathsf{T}\boldsymbol{V}.$$

These definitions are intended to capture the matrix moments $\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ as follows: Consider any $k \in \mathbb{N}, \pi \in \mathcal{P}([k]), \boldsymbol{w} \in \mathcal{G}(k)$ and any $\boldsymbol{a} \in \mathcal{C}(\pi)$. Let the disjoint blocks of π be given by $\pi = \mathcal{V}_1 \sqcup \mathcal{V}_2 \cdots \sqcup \mathcal{V}_{|\pi|}$.

In order to capture $\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ in the subsampled Hadamard case $\boldsymbol{\Psi} = \boldsymbol{H}\overline{\boldsymbol{B}}\boldsymbol{H}^\mathsf{T}$ and the subsampled Haar case $\boldsymbol{\Psi} = \boldsymbol{O}\overline{\boldsymbol{B}}\boldsymbol{O}^\mathsf{T}$ we will set $\boldsymbol{V}_{1:d}$ as follows:

1. In the subsampled Haar case, we set:

$$\{\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots \boldsymbol{V}_d\} = \{(\boldsymbol{o}_{a_{\mathcal{V}_s}} \odot \boldsymbol{o}_{a_{\mathcal{V}_t}}) - \delta(s,t)\hat{\boldsymbol{e}} : s, t \in [|\pi|],\ s \leq t,\ W_{st}(\boldsymbol{w}, \pi) > 0\},$$

where,

$$\boldsymbol{e}^{\mathsf{T}} = \left(\frac{1}{m}, \frac{1}{m} \cdots \frac{1}{m}\right), \ \delta(s,t) = \begin{cases} 1: & s = t \\ 0: & s \neq t \end{cases}.$$

If for some $i \in [d]$ and some $s, t \in [|\pi|]$ we have $\boldsymbol{V}_i = \boldsymbol{o}_{a_{\mathcal{V}_s}} \odot \boldsymbol{o}_{a_{\mathcal{V}_t}} - \delta(s,t)\hat{\boldsymbol{e}}$, we will abuse notation and often refer to $\boldsymbol{V}_i$ as $\boldsymbol{V}_{st}$. Likewise the corresponding entries of $\boldsymbol{T}, \hat{\boldsymbol{T}}, T_i, \hat{T}_i$ will be referred to as $T_{st}, \hat{T}_{st}$.

2. In the subsampled Hadamard case, we set:

$$\{\boldsymbol{V}_1, \boldsymbol{V}_2, \cdots \boldsymbol{V}_d\} = \{\boldsymbol{h}_{a_{\mathcal{V}_s}} \odot \boldsymbol{h}_{a_{\mathcal{V}_t}} - \delta(s,t)\hat{\boldsymbol{e}} : s, t \in [|\pi|], \ s \leq t, \ W_{st}(\boldsymbol{w}, \pi) > 0\}.$$

If for some $i \in [d]$ and some $s, t \in [|\pi|]$ we have $\boldsymbol{V}_i = \boldsymbol{h}_{a_{\mathcal{V}_s}} \odot \boldsymbol{h}_{a_{\mathcal{V}_t}} - \delta(s,t)\hat{\boldsymbol{e}}$, we will abuse notation and often refer to $\boldsymbol{V}_i$ as $\boldsymbol{V}_{st}$. Likewise the corresponding entries of $\boldsymbol{T}, \hat{\boldsymbol{T}}$: $T_i, \hat{T}_i$ will be referred to as $T_{st}, \hat{T}_{st}$.

With the above conventions and the observation that $\sum_{i=1}^{m} \overline{B}_{ii} = 0$ we have:

$$\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) = \prod_{\substack{s,t \in [|\pi|] \\ s \leq t \\ W_{st}(\boldsymbol{w},\pi) > 0}} T_{st}^{W_{st}(\boldsymbol{w},\pi)}.$$

The remainder of this section is organized as follows:

1. First, in Lemma 43 we show that $\hat{\boldsymbol{\Sigma}}$ converges to a fixed deterministic matrix $\boldsymbol{\Sigma}$ and bound the rate of convergence in terms of $\mathbb{E}\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{Fr}}^2$.

2. In Lemma 44 we upper bound $\mathbb{E}\|\hat{\boldsymbol{T}} - \boldsymbol{T}\|_2^2$. Consequently a Gaussian approximation result for $\hat{\boldsymbol{T}}$ implies a Gaussian approximation result for $\boldsymbol{T}$.

3. In Lemma 45, we use a standard Berry Eseen bound of [132] to derive a Gaussian approximation result for $\hat{\boldsymbol{T}}$ since it is a weighted sum of i.i.d. centered random variables.

4. Finally we conclude by using the above lemmas to provide a proof for Propositions 11 and 10.

Lemma 43. *1. For the Hadamard case suppose $\boldsymbol{w}$ is disassortative with respect to π and $\boldsymbol{a}$ is a conflict free labelling of $(\boldsymbol{w}, \pi)$. Then,*

$$\hat{\boldsymbol{\Sigma}} = \kappa(1-\kappa)\boldsymbol{I}_d.$$

2. For the Haar case there exists a universal constant $C < \infty$ such that for any partition $\pi \in \mathcal{P}([k])$, any weight matrix $\boldsymbol{w} \in \mathcal{G}(k)$ and any labelling $\boldsymbol{a} \in \mathcal{C}(\pi)$ we have,

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{Fr}}^2 \leq \frac{C \cdot k^4 \cdot (\kappa^2(1-\kappa)^2)}{m}.$$

where the matrix $\boldsymbol{\Sigma}$ is a diagonal matrix whose diagonal entries are given by:

$$\Sigma_{st,st} = \begin{cases} \kappa(1-\kappa): & s \neq t \\ 2\kappa(1-\kappa): & s = t \end{cases}.$$

Proof. Recall that,

$$\hat{\boldsymbol{\Sigma}} = m\kappa(1-\kappa)\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}.$$

We consider the Hadamard and the Haar case separately.

Hadamard Case: Consider two pairs (s, t) and (s', t') such that:

$$s \leq t, \ W_{st}(\boldsymbol{w}, \pi) > 0, \ s, t \ \in \ [|\pi|].$$

and the analogous assumptions on the pair (s', t'). Then the entry $\hat{\Sigma}_{st,s't'}$ is given by:

$$\begin{aligned}
\hat{\Sigma}_{st,s't'} &= m\kappa(1-\kappa)\langle \boldsymbol{V}_{st}, \boldsymbol{V}_{s't'}\rangle \\
&= m\kappa(1-\kappa)\langle \boldsymbol{h}_{a_{\mathcal{V}_s}} \odot \boldsymbol{h}_{a_{\mathcal{V}_t}} - \delta(s,t)\hat{\boldsymbol{e}}, \boldsymbol{h}_{a_{\mathcal{V}'_s}} \odot \boldsymbol{h}_{a_{\mathcal{V}'_t}} - \delta(s',t')\hat{\boldsymbol{e}}\rangle \\
&\overset{\text{(a)}}{=} \kappa(1-\kappa)\langle \boldsymbol{h}_{a_{\mathcal{V}_s} \oplus a_{\mathcal{V}_t}} - \sqrt{m}\delta(s,t)\hat{\boldsymbol{e}}, \boldsymbol{h}_{a_{\mathcal{V}'_s} \oplus a_{\mathcal{V}'_t}} - \sqrt{m}\delta(s',t')\hat{\boldsymbol{e}}\rangle \\
&\overset{\text{(b)}}{=} \kappa(1-\kappa)\langle \boldsymbol{h}_{a_{\mathcal{V}_s} \oplus a_{\mathcal{V}_t}}, \boldsymbol{h}_{a_{\mathcal{V}'_s} \oplus a_{\mathcal{V}'_t}}\rangle \\
&\overset{\text{(c)}}{=} \kappa(1-\kappa)\delta(s,s')\delta(t,t').
\end{aligned}$$

In the step marked (a) we appealed to Lemma 20. In the step marked (b), we noted that $\hat{\boldsymbol{e}} = \boldsymbol{h}_1/\sqrt{m}$ and $\hat{\boldsymbol{e}} \perp \boldsymbol{h}_{a_{\mathcal{V}_s} \oplus a_{\mathcal{V}_t}}$ unless $s = t$ which is ruled out by the fact that $\boldsymbol{w}$ is disassortative with respect to π i.e. $W_{ss}(\boldsymbol{w}, \pi) = 0$. In the step marked (c) we used the fact that $\boldsymbol{a}$ is a conflict free labelling. Consequently, we have shown that $\hat{\boldsymbol{\Sigma}} = \kappa(1-\kappa)\boldsymbol{I}_d$.

Haar case: By the bias-variance decomposition:

$$\mathbb{E}\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{Fr}}^2 = \mathbb{E}\|\hat{\boldsymbol{\Sigma}} - \mathbb{E}\hat{\boldsymbol{\Sigma}}\|_{\mathsf{Fr}}^2 + \|\mathbb{E}\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{Fr}}^2.$$

We will first compute $\mathbb{E}\hat{\boldsymbol{\Sigma}}$. Consider the $(st, s't')$ entry of $\hat{\boldsymbol{\Sigma}}$:

$$\begin{aligned}
\hat{\Sigma}_{st,s't'} &= m\kappa(1-\kappa)\langle \boldsymbol{V}_{st}, \boldsymbol{V}_{s't'}\rangle \\
&= m\kappa(1-\kappa)\langle \boldsymbol{o}_{a_{\mathcal{V}_s}} \odot \boldsymbol{o}_{a_{\mathcal{V}_t}} - \delta(s,t)\hat{\boldsymbol{e}}, \boldsymbol{o}_{a_{\mathcal{V}'_s}} \odot \boldsymbol{o}_{a_{\mathcal{V}'_t}} - \delta(s',t')\hat{\boldsymbol{e}}\rangle \\
&= m\kappa(1-\kappa)\left[\sum_{i=1}^{m}\left((\boldsymbol{o}_{a_{\mathcal{V}_s}})_i(\boldsymbol{o}_{a_{\mathcal{V}_t}})_i - \frac{\delta(s,t)}{m}\right)\left((\boldsymbol{o}_{a_{\mathcal{V}'_s}})_i(\boldsymbol{o}_{a_{\mathcal{V}'_t}})_i - \frac{\delta(s',t')}{m}\right)\right].
\end{aligned}$$

Note that $\boldsymbol{O}_i$ is a uniformly random unit vector. Hence we can compute $\mathbb{E}\hat{\boldsymbol{\Sigma}}$ using Fact 4. We obtain:

$$\frac{\mathbb{E}\hat{\Sigma}_{st,s't'}}{\kappa(1-\kappa)} = \begin{cases} 2 - \frac{6}{m+2} : & s = s' = t = t' \\ \frac{2}{(m-1)(m+2)} : & s = t, s' = t', s \neq s' \\ 1 + \frac{2}{(m-1)(m+2)} : & s = s', t = t', s \neq t \\ 0 : & \text{otherwise} \end{cases}.$$

Hence, the bias term can be bounded by:

$$\|\mathbb{E}\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathsf{Fr}}^2 \leq \frac{36 \cdot k^4 \cdot \kappa^2(1-\kappa)^2}{(m+2)^2}.$$

On the other hand, applying the Poincare Inequality (Fact 7) and a tedious calculation involving 6th moments of a random unit vector (see for example Proposition 2.5 of [133]) shows that,

$$\mathrm{Var}(\hat{\Sigma}_{st,s't'}) \leq \frac{C \cdot \kappa^2(1-\kappa)^2}{m},$$

for some universal constant C. Hence,

$$\mathbb{E}\|\hat{\mathbf{\Sigma}} - \mathbb{E}\hat{\mathbf{\Sigma}}\|_{\mathsf{Fr}}^2 \leq \frac{C \cdot k^4 \cdot \kappa^2(1-\kappa)^2}{m},$$

for some universal constant C, and consequently the claim of the lemma holds.

□

Lemma 44. *We have,*

$$\mathbb{E}\left[\|\boldsymbol{T} - \hat{\boldsymbol{T}}\|_2^2\right] \leq \frac{Ck^3}{\sqrt{m}},$$

for a universal constant C.

Proof. Let $\overline{\boldsymbol{b}}, \hat{\boldsymbol{b}} \in \mathbb{R}^m$ be the vectors formed by the diagonals of $\overline{\boldsymbol{B}}, \hat{\boldsymbol{B}}$, respectively. Define:

$$p_1 = \mathbb{P}(\overline{b}_1 \neq \hat{b}_1), \;\; p_2 = \mathbb{P}(\overline{b}_1 \neq \hat{b}_1, \; \overline{b}_2 \neq \hat{b}_2).$$

We have,

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{T} - \hat{\boldsymbol{T}}\|_2^2 \mid \boldsymbol{V}\right] &= m\mathbb{E}\left[(\overline{\boldsymbol{b}} - \hat{\boldsymbol{b}})^\mathsf{T}\boldsymbol{V}\boldsymbol{V}^\mathsf{T}(\overline{\boldsymbol{b}} - \hat{\boldsymbol{b}})\right] \\
&= m\mathsf{Tr}\left(\boldsymbol{V}\boldsymbol{V}^\mathsf{T}\mathbb{E}\left[(\overline{\boldsymbol{b}} - \hat{\boldsymbol{b}})(\overline{\boldsymbol{b}} - \hat{\boldsymbol{b}})^\mathsf{T}\right]\right) \\
&= m\mathsf{Tr}\left(\boldsymbol{V}\boldsymbol{V}^\mathsf{T}(1-2\kappa)^2\left(p_2\mathbf{1}\mathbf{1}^\mathsf{T} + (p_1 - p_2)\boldsymbol{I}_m\right)\right) \\
&= m(1-2\kappa)^2\left(p_2\left\|\boldsymbol{V}^\mathsf{T}\mathbf{1}\right\|_2^2 + (p_1 - p_2)\mathsf{Tr}\left(\boldsymbol{V}\boldsymbol{V}^\mathsf{T}\right)\right).
\end{aligned}$$

Now, since $\boldsymbol{V}^\mathsf{T}$ has centered coordinate-wise product of columns of an orthogonal matrix we have $\boldsymbol{V}^\mathsf{T}\mathbf{1} = 0$. Hence,

$$\mathbb{E}\left[\|\boldsymbol{T} - \hat{\boldsymbol{T}}\|_2^2 \mid \boldsymbol{V}\right] = (p_1 - p_2)\mathsf{Tr}\left(\boldsymbol{V}\boldsymbol{V}^\mathsf{T}\right).$$

Next we compute $p_1 = \mathbb{P}(\overline{b}_1 \neq \hat{b}_1)$. Observe that conditional on N, the symmetric difference $S \triangle \hat{S}$ is a uniformly random set of size $|N - n|$. Hence,

$$\mathbb{P}(\overline{b}_1 \neq \hat{b}_1 | N) = \mathbb{P}(1 \in S \triangle \hat{S} | N) = \frac{|n - N|}{m}.$$

Therefore

$$p_1 = \frac{\mathbb{E}\left[N - n\right]}{m} \leq \frac{\sqrt{\mathrm{Var}(N))}}{m} = \frac{\sqrt{\kappa(1-\kappa)}}{\sqrt{m}}.$$

Hence, we obtain

$$\mathbb{E}\left[\|\boldsymbol{T} - \hat{\boldsymbol{T}}\|_2^2 | \boldsymbol{V}\right] \leq \frac{(1-2\kappa)^2}{\sqrt{m \cdot \kappa(1-\kappa)}} \cdot \mathsf{Tr}(\hat{\boldsymbol{\Sigma}}). \tag{4.38}$$

By Lemma 43 we have,

$$\begin{aligned}\mathbb{E}\mathsf{Tr}(\hat{\mathbf{\Sigma}}) &\leq \mathbb{E}\mathsf{Tr}(\mathbf{\Sigma}) + \sqrt{d \cdot \mathbb{E}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\mathsf{Fr}}^2} \\ &\leq C\kappa(1-\kappa)k^3.\end{aligned}$$

where constant $C_{\kappa,d}$ depends only on κ, d. And hence,

$$\mathbb{E}\left[\|\boldsymbol{T} - \hat{\boldsymbol{T}}\|_2^2\right] \leq \frac{Ck^3}{\sqrt{m}},$$

for a universal constant C. □

Lemma 45. *Under the assumptions and notations of Lemma 43 for both the subsampled Haar sensing and the subsampled Hadamard sensing models, we have, for any bounded Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$:*

$$\mathbb{E}\left|\mathbb{E}[f(\hat{\boldsymbol{T}})|\boldsymbol{V}] - \mathbb{E}f(\hat{\mathbf{\Sigma}}^{1/2}\mathbf{Z})\right| \leq \frac{C_k \cdot (\|f\|_\infty + \|f\|_{\mathsf{Lip}})}{\sqrt{m}}. \tag{4.39}$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$, C_k is a constant depending only on k.

Proof. Note that $\hat{\boldsymbol{T}} = \sqrt{m}\boldsymbol{V}^\mathsf{T}\hat{\boldsymbol{b}}$ and $\sqrt{m}\hat{\mathbf{\Sigma}}^{\frac{-1}{2}}\boldsymbol{V}^\mathsf{T}\hat{\boldsymbol{b}}$ has the identity covariance matrix. Hence, by the Berry Eseen bound of [132] for any bounded and Lipschitz function g we have

$$\left|\mathbb{E}\left[g\left(\hat{\mathbf{\Sigma}}^{\frac{-1}{2}}\hat{\boldsymbol{T}}\right)\right] - \mathbb{E}\left[\boldsymbol{Z}\right]\right| \leq \frac{C_d \cdot \rho_3' \cdot \left(\left\|g\right\|_\infty + \left\|g\right\|_{Lip}\right)}{\sqrt{m}}, \tag{4.40}$$

where C_d is a constant only dependent on d and

$$\begin{aligned}\rho_3' &= m^2 \sum_{i=1}^{m} \mathbb{E}\left[\hat{b}_i \|\hat{\boldsymbol{\Sigma}}^{\frac{-1}{2}} \boldsymbol{v}_i\|_2^3 | \boldsymbol{V}\right] \\ &= m^2 \left(\kappa(1-\kappa)^3 + (1-\kappa)\kappa^3\right) \sum_{i=1}^{m} \|\hat{\boldsymbol{\Sigma}}^{\frac{-1}{2}} \boldsymbol{v}_i\|_2^3 \\ &\leq m^2 \cdot \sqrt{d} \cdot \|\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\|_{\mathsf{op}}^3 \cdot (\kappa(1-\kappa)) \cdot \sum_{i=1}^{m} \|\boldsymbol{v}_i\|_3^3\end{aligned}$$

.

Define $g(X) \triangleq f\left(\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{X}\right)$, hence, $g\left(\hat{\boldsymbol{\Sigma}}^{\frac{-1}{2}} \boldsymbol{V}^{\top} \hat{\boldsymbol{b}}\right) = f\left(\hat{T}\right)$. Moreover, $\|g\|_{\infty} \leq \|f\|_{\infty}$ and $\|g\|_{Lip} \leq \|\boldsymbol{\Sigma}\|_{op}^{\frac{1}{2}} \|f\|_{Lip}$. Hence we obtain:

$$\left|\mathbb{E}[f(\hat{T})|\boldsymbol{V}] - \mathbb{E} f(\hat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{Z})\right| \leq \\ C_d(\kappa(1-\kappa)) \cdot m^{\frac{3}{2}} \cdot (\|f\|_{\infty} + \|\hat{\boldsymbol{\Sigma}}\|_{\mathsf{op}}^{\frac{1}{2}} \|f\|_{\mathsf{Lip}}) \cdot \|\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\|_{\mathsf{op}}^3 \cdot \sum_{i=1}^{m} \|\boldsymbol{v}_i\|_3^3. \tag{4.41}$$

We define the event:

$$\mathcal{E} \overset{\text{def}}{=} \left\{\boldsymbol{V} : \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{Fr}}^2 \leq \frac{\kappa^2(1-\kappa)^2}{4}\right\}.$$

By Markov Inequality and Lemma 43, we know that, $\mathbb{P}(\mathcal{E}^c) \leq Ck^4/m$ for some universal constant C. Hence,

$$\mathbb{E}\left|\mathbb{E}[f(\hat{T})|\boldsymbol{V}] - \mathbb{E} f(\hat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{Z})\right| \leq \frac{2C \cdot \|f\|_{\infty} \cdot k^4}{m} + \mathbb{E}\left|\mathbb{E}[f(\hat{T})|\boldsymbol{V}] - \mathbb{E} f(\hat{\boldsymbol{\Sigma}}^{1/2} \boldsymbol{Z})\right| \mathbb{I}_{\mathcal{E}}.$$

On the event $\mathcal{E}$ we have,

$$\begin{aligned}
\|\hat{\mathbf{\Sigma}}\|_{\mathsf{op}} &\leq \|\mathbf{\Sigma}\|_{\mathsf{op}} + \frac{\kappa(1-\kappa)}{2} \leq \frac{5\kappa(1-\kappa)}{2}, \\
\|\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\|_{\mathsf{op}} &\leq \|\mathbf{\Sigma}^{-\frac{1}{2}}\|_{\mathsf{op}} + \|\hat{\mathbf{\Sigma}}^{-\frac{1}{2}} - \mathbf{\Sigma}^{-\frac{1}{2}}\|_{\mathsf{op}} \overset{\text{(a)}}{\leq} \frac{1}{\kappa(1-\kappa)} + \frac{1}{2} \leq \frac{9}{8(\kappa(1-\kappa))}, \\
\mathbb{E}\|\boldsymbol{v}_i\|^3 &= \sum_{j=1}^{d} \mathbb{E}|v_{ij}|^3 \overset{\text{(b)}}{\leq} \frac{Cd}{m^3}.
\end{aligned}$$

In the step marked (a) we used the continuity estimate for matrix square root in Fact 8. In the step marked (b), we recalled the definition of $\boldsymbol{v}_i$ and used the moment bounds for a coordinate of a random unit vector from Fact 4. Substituting these estimates in (4.41) we obtain:

$$\mathbb{E}\left|\mathbb{E}[f(\hat{\boldsymbol{T}})|\boldsymbol{V}] - \mathbb{E}f(\hat{\mathbf{\Sigma}}^{1/2}\boldsymbol{Z})\right| \leq \frac{2C \cdot \|f\|_\infty \cdot k^4}{m} + \frac{C_k \cdot (\|f\|_\infty + \|f\|_{\mathsf{Lip}})}{\sqrt{m}}.$$

□

Using the above lemmas, we can now provide a proof of Propositions 11 and 10.

Proof of Propositions 11 and 10. Define the polynomial $p(\boldsymbol{z})$ as:

$$p(\boldsymbol{z}) \overset{\text{def}}{=} \prod_{\substack{s,t \in [|\pi|] \\ s \leq t \\ W_{st}(\boldsymbol{w},\pi) > 0}} z_{st}^{W_{st}(\boldsymbol{w},\pi)},$$

and the indicator function:

$$\mathbb{I}_{\mathcal{E}}(\boldsymbol{z}) \overset{\text{def}}{=} \begin{cases} 1: & \boldsymbol{z} \in \mathcal{E} \\ 0: & \boldsymbol{z} \notin \mathcal{E} \end{cases},$$

where:

$$\mathcal{E} \overset{\text{def}}{=} \left\{ \max_{s,t} |z_{st}| \leq \left(2048 \log^3(m)\right)^{\frac{1}{2}} \right\}.$$

Recall that we had,

$$\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) = \prod_{\substack{s,t\in[|\pi|] \\ s\leq t \\ W_{st}(\boldsymbol{w},\pi)>0}} T_{st}^{W_{st}(\boldsymbol{w},\pi)} = p(\boldsymbol{T}),$$

and in Lemma 22 we showed that,

$$\mathbb{P}(\boldsymbol{T} \notin \mathcal{E}) \leq \frac{C}{m^2}.$$

We additionally define the function $\widetilde{p}(\boldsymbol{z}) \overset{\text{def}}{=} p(\boldsymbol{z})\mathbb{I}_{\mathcal{E}}(\boldsymbol{z})$. observe that:

$$\|\widetilde{p}\|_\infty \leq \left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}}, \ \|\widetilde{p}\|_{\mathsf{Lip}} \leq \|\boldsymbol{w}\|\left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}}.$$

Let $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Then, we can write:

$$\begin{aligned}
\left|\mathbb{E}p(\boldsymbol{T}) - \mathbb{E}p(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z})\right| &\leq \left|\mathbb{E}\widetilde{p}(\boldsymbol{T}) - \mathbb{E}\widetilde{p}(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z})\right| + |\mathbb{E}p(\boldsymbol{T})\mathbb{I}_{\mathcal{E}^c}(\boldsymbol{T})| + |\mathbb{E}p(\boldsymbol{T})\mathbb{I}_{\mathcal{E}^c}(\boldsymbol{\Sigma}^{\frac{1}{2}}Z)| \\
&\leq \underbrace{\left|\mathbb{E}\widetilde{p}(\boldsymbol{T}) - \mathbb{E}\widetilde{p}(\hat{\boldsymbol{T}})\right|}_{(\mathsf{I})} + \underbrace{\left|\mathbb{E}\widetilde{p}(\boldsymbol{T}) - \mathbb{E}\widetilde{p}(\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{Z})\right|}_{(\mathsf{II})} + \underbrace{\left|\mathbb{E}\widetilde{p}(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}) - \mathbb{E}\widetilde{p}(\widehat{\boldsymbol{\Sigma}^{\frac{1}{2}}}\boldsymbol{Z})\right|}_{(\mathsf{III})} \\
&\quad + \underbrace{|\mathbb{E}p(\boldsymbol{T})\mathbb{I}_{\mathcal{E}^c}(\boldsymbol{T})|}_{(\mathsf{IV})} + \underbrace{|\mathbb{E}p(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z})\mathbb{I}_{\mathcal{E}^c}(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z})|}_{(\mathsf{V})}.
\end{aligned}$$

We control each of these terms separately.

Analysis of (I)**:** In order to control I observe that:

$$\begin{aligned}
(\mathsf{I}) &\leq \|\widetilde{p}\|_{\mathsf{Lip}}\mathbb{E}\|\boldsymbol{T} - \hat{\boldsymbol{T}}\|_2 \\
&\leq \|\widetilde{p}\|_{\mathsf{Lip}} \cdot (\mathbb{E}\|\boldsymbol{T} - \hat{\boldsymbol{T}}\|_2^2)^{\frac{1}{2}} \\
&\leq C \cdot \|\boldsymbol{w}\| \cdot \left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}} \cdot \frac{\sqrt{k^3}}{m^{\frac{1}{4}}}.
\end{aligned}$$

In the last step, we appealed to Lemma 44.

Analysis of (II)**:** In order to control I, recall that:

$$\|\widetilde{p}\|_\infty \leq \left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}}, \ \|\widetilde{p}\|_{\mathsf{Lip}} \leq \|\boldsymbol{w}\| \left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}}.$$

Hence, by Lemma 45 we have,

$$(\mathsf{II}) \leq \frac{C_k \cdot (2048\log^3(m))^{\frac{\|\boldsymbol{w}\|}{2}}(1 + \|\boldsymbol{w}\|)}{\sqrt{m}}.$$

Analysis of (III)**:** Again using the Lipchitz bound on $\widetilde{p}$ we have,

$$\begin{aligned}
(\mathsf{III}) &\leq \mathbb{E}|\widetilde{p}(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z}) - \widetilde{p}(\widehat{\boldsymbol{\Sigma}^{\frac{1}{2}}}\boldsymbol{Z})| \\
&\leq \|\boldsymbol{w}\| \left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}} \cdot \mathbb{E}\|(\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}} - \boldsymbol{\Sigma}^{\frac{1}{2}})\boldsymbol{Z}\|_2 \\
&\leq \|\boldsymbol{w}\| \left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}} \cdot \sqrt{\mathbb{E}\|(\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}} - \boldsymbol{\Sigma}^{\frac{1}{2}})\boldsymbol{Z}\|_2^2} \\
&\leq \|\boldsymbol{w}\| \left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}} \cdot \sqrt{\mathbb{E}\|\hat{\boldsymbol{\Sigma}}^{\frac{1}{2}} - \boldsymbol{\Sigma}^{\frac{1}{2}}\|_{\mathsf{Fr}}^2} \\
&\overset{(a)}{\leq} \|\boldsymbol{w}\| \left(2048\log^3(m)\right)^{\frac{\|\boldsymbol{w}\|}{2}} \cdot \frac{k^2}{\lambda_{\max}(\boldsymbol{\Sigma})} \cdot \mathbb{E}\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{Fr}}^2 \\
&\overset{(b)}{\leq} \frac{C \cdot k^6 \cdot \|\boldsymbol{w}\|(2048\log^3(m))^{\frac{\|\boldsymbol{w}\|}{2}}}{m}.
\end{aligned}$$

In the step marked (a) we used the fact that the continuity estimate for matrix square roots given in Fact 8. In the step marked (b) we recalled the definition of $\boldsymbol{\Sigma}$ and observed that $\lambda_{\max}(\boldsymbol{\Sigma}) \geq \kappa(1-\kappa)$ for the subsampled Haar and the Hadamard sensing model. We also used the bound on $\mathbb{E}\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathsf{Fr}}^2$ obtained in Lemma 43.

Analysis of (IV)**:** We can control (III) as follows:

$$\begin{aligned}(\mathsf{IV}) &\leq \sqrt{\mathbb{E}p^2(\boldsymbol{T})} \cdot \sqrt{\mathbb{P}(\boldsymbol{T} \notin \mathcal{E})} \\ &\overset{(c)}{\leq} \frac{C\sqrt{\mathbb{E}\mathcal{M}(\sqrt{m}\boldsymbol{\Psi}, 2\boldsymbol{w}, \pi, \boldsymbol{a})}}{m} \\ &\overset{(d)}{\leq} \frac{(C\|\boldsymbol{w}\| \log^2(m))^{\frac{\|\boldsymbol{w}\|}{2}}}{m}\end{aligned}$$

In the step marked (c) we recalled that $\mathbb{P}(\boldsymbol{T} \notin \mathcal{E}) \leq C/m^2$ and expressed $p^2(\boldsymbol{T})$ as a matrix moment. In the step marked (d) we used the bounds on matrix moments obtained in Lemma 19.

Analysis of (IV)**:** We recall that $\boldsymbol{\Sigma}$ was a diagonal matrix with $|\Sigma_{ii}| \leq 2\kappa(1-\kappa) \leq 1$. Hence,

$$\begin{aligned}(\mathsf{V}) &\leq \sqrt{\mathbb{E}p^2(\boldsymbol{\Sigma}^{\frac{1}{2}})} \cdot \sqrt{\mathbb{P}(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z} \notin \mathcal{E})} \\ &\overset{(e)}{\leq} \frac{k\|\boldsymbol{w}\|^{\frac{\|\boldsymbol{w}\|}{2}}}{m}.\end{aligned}$$

In the step marked (e) we used standard moment and tail bounds on Gaussian random variables.

Combining the bounds on I – V immediately yields the claims of Proposition 11 and 10. □

4.10.4 Missing Proofs from Section 4.8

Proof of Lemma 26

Proof of Lemma 26. We will assume that $\mathcal{A}$ is of Type 1 (the proof of the other types is analogous):

$$\mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}) = p_1(\boldsymbol{\Psi})q_1(\boldsymbol{Z})p_2(\boldsymbol{\Psi}) \cdots q_{k-1}(\boldsymbol{Z})p_k(\boldsymbol{\Psi}).$$

Define for any $i \in [k]$:

$$\mathcal{A}_0 \overset{\text{def}}{=} p_1(\boldsymbol{\Psi})q_1(\mathrm{Diag}\,(z))p_2(\boldsymbol{\Psi})\cdots q_{k-1}(\mathrm{Diag}\,(z))p_k(\boldsymbol{\Psi}),$$
$$\mathcal{A}_i \overset{\text{def}}{=} p_1(\boldsymbol{\Psi})q_1(\mathrm{Diag}\,(\widetilde{z}))\cdots q_i(\mathrm{Diag}\,(\widetilde{z}))p_{i+1}(\boldsymbol{\Psi})q_{i+1}(\mathrm{Diag}\,(z))\cdots q_{k-1}(\mathrm{Diag}\,(z))p_k(\boldsymbol{\Psi}).$$

where $\boldsymbol{\Psi} = \boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^\mathsf{T}$. Observe that we can write:

$$\begin{aligned}
&z^\mathsf{T}\mathcal{A}(\boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^\mathsf{T}, \mathrm{Diag}\,(z))z - \widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^\mathsf{T}, \mathrm{Diag}\,(\widetilde{z}))\widetilde{z} = z^\mathsf{T}\mathcal{A}_0 z - \widetilde{z}^\mathsf{T}\mathcal{A}_{k-1}\widetilde{z} \\
&= z^\mathsf{T}\mathcal{A}_0 z - z^\mathsf{T}\mathcal{A}_{k-1}z + z^\mathsf{T}\mathcal{A}_{k-1}z + \widetilde{z}^\mathsf{T}\mathcal{A}_{k-1}\widetilde{z} \\
&= \left(\sum_{i=0}^{k-2} z^\mathsf{T}(\mathcal{A}_i - \mathcal{A}_{i+1})z\right) + \langle \mathcal{A}_{k-1}, zz^\mathsf{T} - \widetilde{z}\widetilde{z}^\mathsf{T}\rangle.
\end{aligned}$$

We bound each of these terms separately. First observe that:

$$\begin{aligned}
|z^\mathsf{T}(\mathcal{A}_i - \mathcal{A}_{i+1})z| &\leq \|z\|_2^2 \cdot \|\mathcal{A}_i - \mathcal{A}_{i+1}\|_{\mathsf{op}} \\
&\leq C(\mathcal{A}) \cdot \|z\|_2^2 \cdot \|z - \widetilde{z}\|_\infty.
\end{aligned}$$

Next we note that,

$$\begin{aligned}
|\langle \mathcal{A}_{k-1}, zz^\mathsf{T} - \widetilde{z}\widetilde{z}^\mathsf{T}\rangle| &\leq 2\|\mathcal{A}_{k-1}\|_{\mathsf{op}} \cdot \|zz^\mathsf{T} - \widetilde{z}\widetilde{z}^\mathsf{T}\|_{\mathsf{op}} \\
&= C(\mathcal{A}) \cdot \|z - \widetilde{z}\|_2 \cdot (\|z\|_2 + \|\widetilde{z}\|_2).
\end{aligned}$$

This gives is the estimate:

$$\left|\frac{z^\mathsf{T}\mathcal{A}(\boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^\mathsf{T}, \mathrm{Diag}\,(z))z}{m} - \frac{\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{U}\overline{\boldsymbol{B}}\boldsymbol{U}^\mathsf{T}, \mathrm{Diag}\,(\widetilde{z}))\widetilde{z}}{m}\right| \leq$$
$$\frac{C(\mathcal{A})}{m} \cdot \Big(\|z\|_2^2 \cdot \|z - \widetilde{z}\|_\infty + \|z - \widetilde{z}\|_2 \cdot (\|z\|_2 + \|\widetilde{z}\|_2)\Big),$$

where $C(\mathcal{A})$ denotes a finite constant depending only on the $\|\|_\infty$ norms and Lipchitz constants of the functions appearing in $\mathcal{A}$. □

Proof of Lemma 27

Proof of Lemma 27. Using the continuity estimate from Lemma 26 we know that on the event $\mathcal{E}$,

$$\left|\frac{z^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})z}{m}-\frac{\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}})\widetilde{z}}{m}\right|\leq\frac{C(\mathcal{A})}{m}\cdot\left(\|z\|_2^2\cdot\|z-\widetilde{z}\|_\infty+\|z-\widetilde{z}\|_2\cdot(\|z\|_2+\|\widetilde{z}\|_2)\right)$$
$$\leq\frac{C(\mathcal{A})}{m}\cdot\left(\|z\|_2^2\cdot\|z\|_\infty+\|z\|_2\cdot(\|z\|_2+\|\widetilde{z}\|_2)\right)\cdot\left(\max_{i\in[m]}\left|\frac{1}{\sigma_i}-1\right|\right)$$
$$\leq\frac{C(\mathcal{A})}{m\kappa}\cdot\left(\|z\|_2^2\cdot\|z\|_\infty+\|z\|_2\cdot(\|z\|_2+\|\widetilde{z}\|_2)\right)\cdot\sqrt{\frac{\log^3(m)}{m}}$$

Hence,

$$\left|\mathbb{E}\frac{z^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})z}{m}-\mathbb{E}\frac{\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}})\widetilde{z}}{m}\mathbb{I}_\mathcal{E}\right|\leq\left|\mathbb{E}\frac{z^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})z}{m}\mathbb{I}_{\mathcal{E}^c}\right|$$
$$+\frac{C(\mathcal{A})\log^{\frac{3}{2}}(m)}{m\sqrt{m}\kappa}\cdot\left(\mathbb{E}\|z\|_2^2\cdot\|z\|_\infty+\mathbb{E}\|z\|_2\cdot(\|z\|_2+\|\widetilde{z}\|_2)\right).$$

Observe that $z^\mathsf{T}\mathcal{A}z\leq\|\mathcal{A}\|_{\mathsf{op}}\|z\|^2\leq C(\mathcal{A})\|z\|_2^2\leq C(\mathcal{A})\|\boldsymbol{x}\|_2^2$. Hence,

$$\left|\mathbb{E}\frac{z^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})z}{m}\mathbb{I}_{\mathcal{E}^c}\right|\leq C(\mathcal{A})\frac{\sqrt{\mathbb{E}\|\boldsymbol{x}\|_2^4\cdot\mathbb{P}(\mathcal{E}^c)}}{m}\leq\frac{C(\mathcal{A})\sqrt{\mathbb{P}(\mathcal{E}^c)}}{\kappa^2}\to 0,$$
$$\mathbb{E}\|z\|_2^2+\mathbb{E}\|z\|_2\|\widetilde{z}\|_2\leq 2\mathbb{E}\|z\|_2^2+\mathbb{E}\|\widetilde{z}\|_2^2\leq 2\mathbb{E}\|\boldsymbol{x}\|_2^2+\mathbb{E}\|\widetilde{z}\|_2^2=\frac{2m}{\kappa}+m,$$
$$\mathbb{E}\|z\|_2^2\cdot\|z\|_\infty\leq m\mathbb{E}\|z\|_\infty^3\leq m\left(\mathbb{E}\|z\|_9^9\right)^{\frac{1}{3}}\leq Cm^{\frac{4}{3}}.$$

This gives us,

$$\left|\mathbb{E}\frac{z^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\boldsymbol{Z})z}{m}-\mathbb{E}\frac{\widetilde{z}^\mathsf{T}\mathcal{A}(\boldsymbol{\Psi},\widetilde{\boldsymbol{Z}})\widetilde{z}}{m}\mathbb{I}_\mathcal{E}\right|\to 0,$$

and hence we have shown,

$$\lim_{m\to\infty} \frac{\mathbb{E} z^{\mathsf{T}} \mathcal{A}(\boldsymbol{\Psi}, \boldsymbol{Z}) z}{m} = \lim_{m\to\infty} \mathbb{E} \frac{\widetilde{z}^{\mathsf{T}} \mathcal{A}(\boldsymbol{\Psi}, \widetilde{\boldsymbol{Z}}) \widetilde{z}}{m} \mathbb{I}_{\mathcal{E}},$$

provided the latter limit exists. □

Proof of Lemma 29

Proof of Lemma 29. Recall that:

$$\widetilde{z}_{a_1} \widetilde{z}_{a_{k+1}} \prod_{i=1}^{k} q_i(\widetilde{z}_{a_i}) = Q_{\mathscr{F}}(\widetilde{z}_{a_1}) \cdot Q_{\mathscr{L}}(\widetilde{z}_{a_{k+1}}) \left(\prod_{i \in \mathcal{S}(\pi)} q_{i-1}(\widetilde{z}_{a_i}) \right) \prod_{i=1}^{|\pi|-|\mathcal{S}(\pi)|-2} (Q_{\mathcal{V}_i}(z_{a_{\mathcal{V}_i}}) + \mu_{\mathcal{V}_i})$$

Hence,

$$\begin{aligned} &\mathbb{E}[\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) q_2(\widetilde{z}_{a_3}) \cdots q_{k-1}(\widetilde{z}_{a_k}) \widetilde{z}_{a_{k+1}} | \boldsymbol{A}] = \\ &\sum_{V \subset [|\pi|-|\mathcal{S}(\pi)|-2]} \mathbb{E} \left[Q_{\mathscr{F}}(\widetilde{z}_{a_1}) Q_{\mathscr{L}}(\widetilde{z}_{a_{k+1}}) \left(\prod_{i \in \mathcal{S}(\pi)} q_{i-1}(\widetilde{z}_{a_i}) \right) \prod_{i \in V} (Q_{\mathcal{V}_i}(\widetilde{z}_{a_{\mathcal{V}_i}})) \middle| \boldsymbol{A} \right] \left(\prod_{i \notin V} \mu_{\mathcal{V}_i} \right) \end{aligned} \tag{4.42}$$

We now apply Mehler's formula to estimate the above conditional expectations. We first check the conditions for Mehler's formula:

1. The random variables $\widetilde{z}$ are marginally $\mathcal{N}(0,1)$. Define $\boldsymbol{\Sigma} = \mathbb{E}[\widetilde{z}\widetilde{z}^{\mathsf{T}} | \boldsymbol{A}]$. $\widetilde{z}$ and are weakly correlated on the event $\mathcal{E}$ since:

   $$\begin{aligned} \max_{i \neq j} |\Sigma_{ij}| &= \left| \frac{(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})_{ij}/\kappa}{\sigma_i \sigma_j} \right| \\ &= \left| \frac{(\boldsymbol{\Psi})_{ij}/\kappa}{\sigma_i \sigma_j} \right| \\ &\leq C \sqrt{\frac{\log^3(m)}{m\kappa^2}}, \text{ for } m \text{ large enough,} \end{aligned}$$

   where C denotes a universal constant.

2. Let $S \subset [m]$ with $|S| \leq k+2$. Let $\boldsymbol{\Sigma}_{S,S}$ denote the principal submatrix of $\boldsymbol{\Sigma}$ formed by picking rows and columns in S. Then by Gershgorin's Circle theorem, on the event $\mathcal{E}$,

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Sigma}) &\geq 1-(k+1)\max_{i\neq j}|\Sigma_{ij}| \\ &\geq 1-C(k+1)\sqrt{\frac{\log^3(m)}{m\kappa^2}} \\ &\geq \frac{1}{2}, \quad \text{for } m \text{ large enough.}\end{aligned}$$

3. Note that for $\xi \sim \mathcal{N}(0,1)$, we have,

$$\begin{aligned}&\mathbb{E}Q_{\mathscr{F}}(\xi)=0,\ \mathbb{E}Q_{\mathscr{L}}(\xi)=0 \ \text{ (Since they are odd functions, see (4.23), (4.24))},\\ &\mathbb{E}q_{i-1}(\xi)=\mathbb{E}\xi q_{i-1}(\xi)=0\ \forall\ i\in\mathcal{S}(\pi) \ \text{ (They are centered, even functions, see Def. 7)},\\ &\mathbb{E}Q_{\mathcal{V}_i}(\xi)=\mathbb{E}\xi Q_{\mathcal{V}_i}(\xi)=0\ \forall\ i\ \in\ [|\pi|-|\mathcal{S}(\pi)|-2] \ \text{ (See (4.26))}\end{aligned}$$

Hence applying the first non-zero term in Mehler's Expansion (Proposition 9) of the conditional expectation:

$$\mathbb{E}\left[Q_{\mathscr{F}}(\widetilde{z}_{a_1})\cdot Q_{\mathscr{L}}(\widetilde{z}_{a_{k+1}})\cdot\left(\prod_{i\in\mathcal{S}(\pi)}q_{i-1}(\widetilde{z}_{a_i})\right)\cdot\prod_{i\in V}(Q_{\mathcal{V}_i}(\widetilde{z}_{a_{\mathcal{V}_i}}))\middle|\boldsymbol{A}\right]$$

has total weight $\|\boldsymbol{w}\|$ given by:

$$\|\boldsymbol{w}\| \geq \frac{1+1+2|\mathcal{S}(\pi)|+2|V|}{2}=1+|\mathcal{S}(\pi)|+|V|.$$

Hence, by Proposition 9 we have,

$$\mathbb{I}_{\mathcal{E}} \cdot \left| \mathbb{E}\left[Q_{\mathscr{F}}(\widetilde{z}_{a_1}) \cdot Q_{\mathscr{L}}(\widetilde{z}_{a_{k+1}}) \cdot \left(\prod_{i \in \mathcal{S}(\pi)} q_{i-1}(\widetilde{z}_{a_i}) \right) \cdot \prod_{i \in V} (Q_{\mathcal{V}_i}(\widetilde{z}_{a_{\mathcal{V}_i}})) \middle| \boldsymbol{A} \right] \right|$$

$$\leq C(\mathscr{A})(\max_{i \neq j} |\Sigma_{i,j}|)^{1+|\mathcal{S}(\pi)|+|V|} \leq C(\mathscr{A}) \cdot \left(\frac{\log^2(m)}{m\kappa^2} \right)^{\frac{1+|\mathcal{S}(\pi)|+|V|}{2}}, \qquad (4.43)$$

where $C(\mathscr{A})$ denotes a finite constant depending only on the functions $q_{1:k}$. When $V = \emptyset$ we will also need to estimate the leading order term more accurately. Define,

$$\mathcal{G}_1(\pi) \overset{\text{def}}{=} \left\{ \boldsymbol{w} \in \mathcal{G}(k+1) : \mathsf{d}_1(\boldsymbol{w}) = 1,\ \mathsf{d}_{k+1}(\boldsymbol{w}) = 1,\ \mathsf{d}_i(\boldsymbol{w}) = 2\ \forall\ i\ \in\ \mathcal{S}(\pi), \right.$$
$$\left. \mathsf{d}_i(\boldsymbol{w}) = 0\ \forall\ i\ \notin\ \{1, k+1\} \cup \mathcal{S}(\pi) \right\}.$$

By Mehler's formula, on the event $\mathcal{E}$, we have:

$$\left| \mathbb{E}\left[Q_{\mathscr{F}}(\widetilde{z}_{a_1}) \cdot Q_{\mathscr{L}}(\widetilde{z}_{a_{k+1}}) \cdot \left(\prod_{i \in \mathcal{S}(\pi)} q_{i-1}(\widetilde{z}_{a_i}) \right) \middle| \boldsymbol{A} \right] - \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} \hat{g}(\boldsymbol{w}, \boldsymbol{\Psi}) \cdot \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) \right|$$

$$\leq C(\mathscr{A}) \cdot \left(\frac{\log^3(m)}{m\kappa^2} \right)^{\frac{2+|\mathcal{S}(\pi)|}{2}},$$

where,

$$\hat{g}(\boldsymbol{w}, \boldsymbol{\Psi}) = \frac{1}{\boldsymbol{w}!} \cdot \left(\prod_{i=1}^{k+1} \frac{1}{\sigma_{a_i}^{\mathsf{d}_i(\boldsymbol{w})}} \right) \cdot \left(\hat{Q}_{\mathscr{F}}(1) \hat{Q}_{\mathscr{L}}(1) \prod_{i \in \mathcal{S}(\pi)} \hat{q}_{i-1}(2) \right) \frac{1}{\kappa^{\|\boldsymbol{w}\|}},$$

and $\mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a})$ are matrix moments as defined in Definition 8. Note that the coefficients $\hat{g}(\boldsymbol{w}, \boldsymbol{\Psi})$ depend on $\boldsymbol{\Psi}$ since,

$$\sigma_i^2 = 1 + \frac{\boldsymbol{\Psi}_{ii}}{\kappa},$$

but we can remove this dependence. On the event $\mathcal{E}$, note that,

$$\max_{i\in[m]} |\sigma_{ii}^2 - 1| \leq C\sqrt{\frac{\log^3(m)}{m\kappa^2}}.$$

Hence defining:

$$\hat{g}(\boldsymbol{w},\pi) = \frac{1}{\boldsymbol{w}!}\cdot\left(\hat{Q}_{\mathscr{F}}(1)\hat{Q}_{\mathscr{L}}(1)\prod_{i\in\mathscr{S}(\pi)}\hat{q}_{i-1}(2)\right)\frac{1}{\kappa^{\|\boldsymbol{w}\|}},$$

we have, for m large enough and on the event $\mathcal{E}$,

$$|\hat{g}(\boldsymbol{w},\pi) - \hat{g}(\boldsymbol{w},\boldsymbol{\Psi})| \leq C_k\sqrt{\frac{\log^3(m)}{m\kappa^2}}.$$

Furthermore, we have the estimate,

$$\begin{aligned}|\mathcal{M}(\boldsymbol{\Psi},\boldsymbol{w},\pi,\boldsymbol{a})| &\leq (\max_{i,j}|\Psi_{ij}|)^{\|\boldsymbol{w}\|_1} \\ &\overset{\text{(a)}}{\leq} C\left(\frac{\log^3(m)}{m\kappa^2}\right)^{\frac{1+|\mathscr{S}(\pi)|}{2}},\end{aligned}$$

where in the step (a), we used the definition of the event $\mathcal{E}$ in (4.22) and the fact that $\|\boldsymbol{w}\| = 1+|\mathscr{S}(\pi)|$ for any $\boldsymbol{w}\in\mathcal{G}_1(\pi)$. Hence we obtain, on the event $\mathcal{E}$,

$$\begin{aligned}&\left|\mathbb{E}\left[Q_{\mathscr{F}}(\widetilde{z}_{a_1})\cdot Q_{\mathscr{L}}(\widetilde{z}_{a_{k+1}})\cdot\left(\prod_{i\in\mathscr{S}(\pi)} q_{i-1}(\widetilde{z}_{a_i})\right)\middle|\boldsymbol{A}\right] - \sum_{\boldsymbol{w}\in\mathcal{G}_1(\pi)}\hat{g}(\boldsymbol{w},\pi)\cdot\mathcal{M}(\boldsymbol{\Psi},\boldsymbol{w},\pi,\boldsymbol{a})\right| \\ &\qquad\qquad\leq C(\mathcal{A})\cdot\left(\frac{\log^3(m)}{m\kappa^2}\right)^{\frac{2+|\mathscr{S}(\pi)|}{2}},\end{aligned}$$

Combining this estimate with (4.42) and (4.43) gives us:

$$\mathbb{I}_{\mathcal{E}} \cdot \left| \mathbb{E}[\widetilde{z}_{a_1} q_1(\widetilde{z}_{a_2}) q_2(\widetilde{z}_{a_3}) \cdots q_{k-1}(\widetilde{z}_{a_k}) \widetilde{z}_{a_{k+1}} | \boldsymbol{A}] - \sum_{\boldsymbol{w} \in \mathcal{G}_1(\pi)} g(\boldsymbol{w}, \pi) \cdot \mathcal{M}(\boldsymbol{\Psi}, \boldsymbol{w}, \pi, \boldsymbol{a}) \right|$$
$$\leq C(\mathscr{A}) \cdot \left(\frac{\log^3(m)}{m\kappa^2} \right)^{\frac{2+|\mathcal{S}(\pi)|}{2}},$$

where,

$$g(\boldsymbol{w}, \pi) = \frac{1}{\kappa^{\|\boldsymbol{w}\|} \boldsymbol{w}!} \cdot \left(\hat{Q}_{\mathscr{F}}(1) \hat{Q}_{\mathscr{L}}(1) \prod_{i \in \mathcal{S}(\pi)} \hat{q}_{i-1}(2) \right) \cdot \left(\prod_{i \in [|\pi| - |\mathcal{S}(\pi)| - 2]} \mu_{\mathcal{V}_i} \right)$$

$$\mathcal{G}_1(\pi) \overset{\text{def}}{=} \left\{ \boldsymbol{w} \in \mathcal{G}(k+1) : \mathsf{d}_1(\boldsymbol{w}) = 1, \ \mathsf{d}_{k+1}(\boldsymbol{w}) = 1, \ \mathsf{d}_i(\boldsymbol{w}) = 2 \ \forall \ i \ \in \ \mathcal{S}(\pi), \right.$$
$$\left. \mathsf{d}_i(\boldsymbol{w}) = 0 \ \forall \ i \ \notin \ \{1, k+1\} \cup \mathcal{S}(\pi) \right\},$$

and $C(\mathscr{A})$ denotes a constant depending only on the functions appearing in $\mathscr{A}$ and k. This was precisely the claim of Lemma 29. □

4.10.5 Proof of Proposition 9

Proof of Proposition 9. Let $\psi(z; \boldsymbol{\Sigma})$ denote the density of a k dimensional zero mean Gaussian vector with positive definite covariance matrix $\boldsymbol{\Sigma}$ i.e. $z \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Suppose that $\Sigma_{ii} = 1 \ \forall \ i \in [k]$. In this situation [130] has found an explicit expression for the Taylor series expansion of $\psi(z; \boldsymbol{\Sigma})$ around $\boldsymbol{\Sigma} = \boldsymbol{I}_k$ given by:

$$\psi(z; \boldsymbol{\Sigma}) = \sum_{\boldsymbol{w} \in \mathcal{G}(k)} \frac{D_{\boldsymbol{\Sigma}}^{\boldsymbol{w}} \ \psi(z; \boldsymbol{I}_k)}{\boldsymbol{w}!} \cdot \left(\prod_{i<j} \Sigma_{ij}^{w_{ij}} \right),$$

where $D_{\mathbf{\Sigma}}^{\boldsymbol{w}}\ \psi(z;\boldsymbol{I}_k)$ denotes the derivative:

$$\begin{aligned} D_{\mathbf{\Sigma}}^{\boldsymbol{w}}\ \psi(z;\boldsymbol{I}_k) &\overset{\text{def}}{=} \frac{\partial^{\|\boldsymbol{w}\|}}{\partial\Sigma_{12}^{w_{12}}\ \partial\Sigma_{13}^{w_{13}}\cdots\partial\Sigma_{23}^{w_{23}}\ \partial\Sigma_{24}^{w_{24}}\cdots\partial\Sigma_{k-1,k}^{w_{k-1,k}}}\ \psi(z;\mathbf{\Sigma})\bigg|_{\mathbf{\Sigma}=\boldsymbol{I}_k} \\ &= \left(\prod_{i=1}^{k} H_{\mathsf{d}_i(\boldsymbol{w})}(z_i)\right)\cdot\psi(z;\boldsymbol{I}_k). \end{aligned}$$

We intend to integrate the Taylor series for $\psi(z;\mathbf{\Sigma})$ to obtain the expansion for the expectation in Proposition 9. In order to do so we need to understand the truncation error in the Taylor Series. By Taylors Theorem, we know that:

$$\psi(z;\mathbf{\Sigma}) - \sum_{\boldsymbol{w}\in\mathcal{G}(k):\|\boldsymbol{w}\|\leq t} \frac{D_{\mathbf{\Sigma}}^{\boldsymbol{w}}\ \psi(z;\boldsymbol{I}_k)}{\boldsymbol{w}!}\cdot\left(\prod_{i<j}\Sigma_{ij}^{w_{ij}}\right) = \sum_{\boldsymbol{w}\in\mathcal{G}(k):\|\boldsymbol{w}\|=t+1} \frac{D_{\mathbf{\Sigma}}^{\boldsymbol{w}}\ \psi(z;\mathbf{\Sigma}_\gamma)}{\boldsymbol{w}!}\cdot\mathbf{\Sigma}^{\boldsymbol{w}}, \tag{4.45}$$

where $\mathbf{\Sigma}_\gamma = \gamma\mathbf{\Sigma} + (1-\gamma)\boldsymbol{I}_k$ for some $\gamma \in (0,1)$. [130] has further showed the following remarkable identity:

$$D_{\mathbf{\Sigma}}^{\boldsymbol{w}}\ \psi(z;\mathbf{\Sigma}) = \frac{\partial^{2\|\boldsymbol{w}\|}}{\partial z_1^{\mathsf{d}_1(\boldsymbol{w})}\ \partial z_2^{\mathsf{d}_2(\boldsymbol{w})}\cdots\partial z_k^{\mathsf{d}_k(\boldsymbol{w})}}\ \psi(z;\mathbf{\Sigma}).$$

An inductive calculation shows that the ratio:

$$\frac{1}{\psi(z;\mathbf{\Sigma})}\ \frac{\partial^{2\|\boldsymbol{w}\|}}{\partial z_1^{\mathsf{d}_1(\boldsymbol{w})}\ \partial z_2^{\mathsf{d}_2(\boldsymbol{w})}\cdots\partial z_k^{\mathsf{d}_k(\boldsymbol{w})}}\ \psi(z;\mathbf{\Sigma}),$$

is a polynomial of degree $4\|\boldsymbol{w}\|$ in the variables $z_1, z_2\dots z_k, \{(\mathbf{\Sigma}^{-1})_{ij}\}_{i<j}$. Hence:

$$\begin{aligned} &\left|\frac{1}{\psi(z;\mathbf{\Sigma})}\ \frac{\partial^{2\|\boldsymbol{w}\|}}{\partial z_1^{\mathsf{d}_1(\boldsymbol{w})}\ \partial z_2^{\mathsf{d}_2(\boldsymbol{w})}\cdots\partial z_k^{\mathsf{d}_k(\boldsymbol{w})}}\ \psi(z;\mathbf{\Sigma})\right| \leq \\ &\qquad\qquad C_{\|\boldsymbol{w}\|}\cdot(1+\sum_{i<j}|(\mathbf{\Sigma}^{-1})_{ij}|^{4\|\boldsymbol{w}\|}+\sum_{i=1}^{k}|z_i|^{4\|\boldsymbol{w}\|}), \end{aligned}$$

where $C_{\|\boldsymbol{w}\|}$ denotes a constant depending only on $\|\boldsymbol{w}\|$. Observing that:

$$(\boldsymbol{\Sigma}^{-1})_{ij} \leq \|\boldsymbol{\Sigma}^{-1}\|_{\mathsf{op}} = \frac{1}{\lambda_{\min}(\boldsymbol{\Sigma})} < \infty.$$

This gives us:

$$\left| \frac{1}{\psi(z;\boldsymbol{\Sigma})} \, \frac{\partial^{2\|\boldsymbol{w}\|}}{\partial z_1^{\mathsf{d}_1(\boldsymbol{w})} \, \partial z_2^{\mathsf{d}_2(\boldsymbol{w})} \cdots \partial z_k^{\mathsf{d}_k(\boldsymbol{w})}} \, \psi(z;\boldsymbol{\Sigma}) \right| \leq C_{\|\boldsymbol{w}\|} \left(1 + \frac{k^2}{\lambda_{\min}^{4\|\boldsymbol{w}\|}(\boldsymbol{\Sigma})} + \sum_{i=1}^{k} |z_i|^{4\|\boldsymbol{w}\|} \right).$$

Substituting this estimate in (4.45) gives us:

$$\begin{aligned} &\left| \psi(z;\boldsymbol{\Sigma}) - \sum_{\boldsymbol{w} \in \mathcal{G}(k):\|\boldsymbol{w}\| \leq t} \frac{D_{\boldsymbol{\Sigma}}^{\boldsymbol{w}} \, \psi(z;\boldsymbol{I}_k)}{\boldsymbol{w}!} \cdot \boldsymbol{\Sigma}^{\boldsymbol{w}} \right| \\ &\qquad \leq C_{t,k} \cdot \left(1 + \frac{k^2}{\lambda_{\min}^{4t+4}(\boldsymbol{\Sigma}_\gamma)} + \sum_{i=1}^{k} |z_i|^{4t+4} \right) \cdot \left(\max_{i \neq j} |\Sigma_{ij}| \right)^{t+1} \cdot \psi(z;\boldsymbol{\Sigma}_\gamma). \end{aligned}$$

Note that $\lambda_{\min}(\boldsymbol{\Sigma}_\gamma) = \gamma + (1-\gamma)\lambda_{\min}(\boldsymbol{\Sigma}) \geq \min(1, \lambda_{\min}(\boldsymbol{\Sigma}))$. Hence,

$$\begin{aligned} &\left| \psi(z;\boldsymbol{\Sigma}) - \sum_{\boldsymbol{w} \in \mathcal{G}(k):\|\boldsymbol{w}\| \leq t} \frac{D_{\boldsymbol{\Sigma}}^{\boldsymbol{w}} \, \psi(z;\boldsymbol{I}_k)}{\boldsymbol{w}!} \cdot \boldsymbol{\Sigma}^{\boldsymbol{w}} \right| \\ &\qquad \leq C_{t,k} \cdot \left(1 + \frac{k^2}{\min(\lambda_{\min}^{4t+4}(\boldsymbol{\Sigma}), 1)} + \sum_{i=1}^{k} |z_i|^{4t+4} \right) \cdot \left(\max_{i \neq j} |\Sigma_{ij}| \right)^{t+1} \cdot \psi(z;\boldsymbol{\Sigma}_\gamma). \end{aligned}$$

Using this expansion to compute the expectation of $\prod_{i=1}^{k} f_i(z_i)$ we obtain:

$$\left| \mathbb{E}\left[\prod_{i=1}^{k} f_i(z_i) \right] - \sum_{\substack{\boldsymbol{w} \in \mathcal{G}(k) \\ \|\boldsymbol{w}\| \leq t}} \left(\prod_{i=1}^{k} \hat{f}_i(\mathsf{d}_i(\boldsymbol{w})) \right) \cdot \frac{\boldsymbol{\Sigma}^{\boldsymbol{w}}}{\boldsymbol{w}!} \right| \leq C \left(1 + \frac{1}{\lambda_{\min}^{4t+4}(\boldsymbol{\Sigma})} \right) \left(\max_{i \neq j} |\Sigma_{ij}| \right)^{t+1},$$

where $C = C_{t,k,f_{1:k}}$ denotes a constant depending only on t, k and the functions $f_{1:k}$. In obtaining the above estimate we use the fact that since the functions f_i have polynomial growth and marginally

$z_i \sim \mathcal{N}(0,1)$ under the measure $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\gamma\right)$ (since $(\Sigma_\gamma)_{ii} = 1$) we have,

$$\mathbb{E}_{z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\gamma)} \left[|z_i|^{4t+4} \prod_{j=1}^{k} |f_j(z_j)| \right] \leq \sum_{j=1}^{k} \mathbb{E}_{z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\gamma)} \left[|z_i|^{4t+4} |f_j(z_j)|^k \right] = C_{t,k,f_{1:k}} < \infty.$$

□

4.10.6 Some Miscellaneous Facts

Fact 2 (Hanson-Wright Inequality [65]). *Let $\boldsymbol{x} = (x_1, x_2 \ldots, x_n) \in \mathbb{R}^n$ be a random vector with independent 1-subgaussian, zero mean components. Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\mathbb{P}\left(|\boldsymbol{x}^\mathsf{T} \boldsymbol{A} \boldsymbol{x} - \mathbb{E} \boldsymbol{x}^\mathsf{T} \boldsymbol{A} \boldsymbol{x}| > t \right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{\|\boldsymbol{A}\|_{\mathsf{Fr}}^2}, \frac{t}{\|\boldsymbol{A}\|_{\mathsf{op}}} \right) \right).$$

Fact 3 (Gaussian Poincare Inequality). *Let $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_n)$. Then, for any L-Lipchitz function $f : \mathbb{R}^n \to \mathbb{R}$ we have,*

$$\mathrm{Var}(f(\boldsymbol{x})) \leq L^2.$$

Fact 4 (Moments of a Random Unit vector, Lemma 2.22 & Proposition 2.5 of [133]). *Let $\boldsymbol{x} \sim$ Unif$(\mathbb{S}_{n-1})$. Let i, j, k, ℓ be distinct indices. Then:*

$$\mathbb{E} x_i^4 = \frac{3}{n(n+2)}, \ \mathbb{E} x_i^2 x_j^2 = \frac{n+1}{n(n-1)(n+2)} \ \mathbb{E} x_i^3 x_j = 0 \ \mathbb{E} x_i x_j x_k^2 = 0, \ \mathbb{E} x_i x_j x_k x_l = 0.$$

Furthermore, there exists a universal constant C such that, for any $t \in \mathbb{N}$:

$$\mathbb{E}|x_i|^t \leq \left(\frac{Ct}{m} \right)^{\frac{t}{2}}.$$

Fact 5 (Concentration on the Sphere, [134])**.** *Let* $\boldsymbol{x} \sim \mathit{Unif}(\mathbb{S}_{n-1})$*. Then*

$$\mathbb{P}\left(|x_1| \geq \epsilon\right) \leq 2e^{-n\epsilon^2/2}.$$

Fact 6 (ℓ_∞ norm of a random unit vector)**.** $\boldsymbol{x} \sim \mathit{Unif}(\mathbb{S}_{n-1})$*. Then*

$$\mathbb{E}\|\boldsymbol{x}\|_\infty^t \leq \left(\frac{C\log(n)}{n}\right)^{\frac{t}{2}},$$

for a universal constant C*.*

Proof. For a random unit vector we can control $\mathbb{E}\|\boldsymbol{x}\|_\infty^t$ as follows. Let $q \in \mathbb{N}$ be a parameter to be set suitably. Then,

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}\|_\infty^t &= \left(\mathbb{E}\|\boldsymbol{x}\|_\infty^{qt}\right)^{\frac{1}{q}} \\
&\leq \left(\sum_{i=1}^n \mathbb{E}|x_i|^{qt}\right)^{\frac{1}{q}} \\
&\overset{\text{(a)}}{=} \left(n\mathbb{E}|x_1|^{qt}\right)^{\frac{1}{q}} \\
&\overset{\text{(b)}}{=} n^{\frac{1}{q}} \cdot q^{\frac{t}{2}} \cdot \left(\frac{Ct}{n}\right)^{\frac{t}{2}} \\
&\overset{\text{(c)}}{\leq} e^t \cdot (2\log(n))^{\frac{t}{2}} \cdot \left(\frac{C}{n}\right)^{\frac{t}{2}}.
\end{aligned}$$

In the step marked (a) we used the fact that the coordinates of a random unit vector are exchangeable, in (b) we used the fact that u_1 is C/m-subgaussian (see Fact 5) and in (c) we set $q = \lfloor\frac{2\log(n)}{t}\rfloor$. □

Fact 7 (Poincare Inequality for Haar Measure, [135])**.** *Consider the following setups:*

1. *Let* $\boldsymbol{O} \sim \mathit{Unif}\big(\mathbb{O}(m)\big)$ *and* $f : \mathbb{R}^{m\times m} \rightarrow \mathbb{R}$ *be a function such that:*

$$f(\boldsymbol{O}) = f(\boldsymbol{O}\boldsymbol{D}),\ \boldsymbol{D} = \mathit{Diag}\left(1, 1, 1, \ldots, 1, \mathsf{sign}(\det(\boldsymbol{O}))\right), \tag{4.46}$$

then,

$$\mathrm{Var}(f(\boldsymbol{O})) \leq \frac{8}{m} \cdot \mathbb{E}\|\nabla f(\boldsymbol{O})\|_{\mathsf{Fr}}^2.$$

for any $m \geq 4$.

2. *Let* $\boldsymbol{O} \sim$ *Unif*$\big(\mathbb{U}(m)\big)$ *and* $f : \mathbb{C}^{m \times m} \to \mathbb{R}$. *Then,*

$$\mathrm{Var}(f(\boldsymbol{O})) \leq \frac{8}{m} \cdot \mathbb{E}\|\nabla f(\boldsymbol{O})\|_{\mathsf{Fr}}^2.$$

Proof. This result is due to [135]. Our reference for these inequalities was the book of [133]. Theorem 5.16 of [133] shows that Haar measures on $\mathbb{SO}(m), \mathbb{U}(m)$ satisfy Log-sobolev inequality with constant $8/m$. It is well known that Log-Sobolev Inequality implies the Poincare Inequality (see for e.g. Lemma 8.12 in [136]). Note that, in the real case we only obtain the Poincare inequality for the Haar measure on $\mathbb{SO}(m)$, condition (4.46) ensures the result still holds for $\boldsymbol{O} \sim \mathrm{Unif}\left(\mathbb{O}(m)\right)$. □

Fact 8 (Continuity of Matrix Square Root [137, Lemma 2.2])**.** *For any two symmetric positive semi-definite matrices* $\boldsymbol{M}_1, \boldsymbol{M}_2$ *we have,*

$$\|\boldsymbol{M}_1^{\frac{1}{2}} - \boldsymbol{M}_2^{\frac{1}{2}}\|_{\mathsf{op}} \leq \frac{\|\boldsymbol{M}_1 - \boldsymbol{M}_2\|_{\mathsf{op}}}{\sqrt{\lambda_{\min}(\boldsymbol{M}_1)}}.$$

Chapter 5: Compressive Phase Retrieval

5.1 Introduction

5.1.1 Motivation

Consider the problem of recovering $\boldsymbol{x} \in Q$ from m noisy phase-less linear measurements

$$\boldsymbol{y} = |A\boldsymbol{x}| + \boldsymbol{\epsilon},$$

where $A \in \mathbb{C}^{m\times n}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^m$ denote the sensing matrix and the measurement noise, respectively. Here Q denotes a compact subset of $\mathbb{C}^n$ and $|\cdot|$ denotes the element-wise absolute value operator. Assume that the class of signals denotes by Q is "structured", but instead of the set Q, or its underlying structure, for recovering $\boldsymbol{x}$ from $\boldsymbol{y}$, we have access to a compression code that takes advantage of the structure of signals in Q to compress them efficiently. For instance, consider the class of images or videos for which compression algorithms, such as JPEG2000 or MPEG4, take advantage of complicated structures within such signals and encode them efficiently. Employing such structures in a phase retrieval algorithm can reduce the number of measurements or equivalently increase the quality of the recovered signals. This raises the following questions:

1. Is it possible to use a given compression algorithm for the recovery of $\boldsymbol{x}$ from its undersampled set of phaseless observations?
2. What is the required number of observations (in terms of the rate-distortion performance of the code), for almost zero-distortion recovery of $\boldsymbol{x}$?
3. Can we find polynomial time algorithms to use a given compression algorithm to recover $\boldsymbol{x}$ from its undersampled set of phaseless observations? If so, how does the answer to the

second question change if we want to approximate the solution in the polynomial time?

Affirmative answers to these questions enable us to use the structures that are employed by the state-of-the-art compression algorithms, such as JPEG2000 or MPEG4, and build phase retrieval recovery algorithms with higher reconstruction quality or equivalently lower number of required measurements for a given desired quality. Furthermore, whenever the image or video compression communities design new compression algorithms that are capable of employing more complex structures, the framework we develop in this chapter, by integrating those codes, automatically, with no extra effort, leads to new phase retrieval algorithms that take advantage to those new structures

In the remainder of this section, we first review the formal definitions of compression algorithms and their rate-distortion performance measures. We will then briefly sates our responses to the above three questions. Finally, we compare our contribution with the existing work in the literature.

5.1.2 Background on compression algorithms

A rate-r compression code is composed of an encoder mapping $\mathcal{E}$ and a decoder mapping $\mathcal{D}$, where

$$\mathcal{E} : \mathbb{C}^n \to \{0,1\}^r, \text{ and } \qquad \mathcal{D} : \{0,1\}^r \to \mathbb{C}^n.$$

The distortion performance of the compression code defined by mappings $(\mathcal{E}, \mathcal{D})$ on set $\mathcal{Q}$ is measured as

$$\delta \triangleq \sup_{\boldsymbol{x} \in \mathcal{Q}} \left\| \boldsymbol{x} - \mathcal{D}(\mathcal{E}(\boldsymbol{x})) \right\|.$$

Throughout the chapter sometimes we use subscript r for the encoder and decoder mappings as $(\mathcal{E}_r, \mathcal{D}_r)$ to highlight the rate of the code. The codebook of compression code $(\mathcal{E}_r, \mathcal{D}_r)$ operating at rate r is defined as

$$\mathcal{C}_r \triangleq \mathcal{D}_r(\mathcal{E}_r(\mathcal{Q})) = \{\mathcal{D}_r(\mathcal{E}_r(\boldsymbol{x})) : \ \boldsymbol{x} \in \mathcal{Q}\}.$$

It is straightforward to confirm that $|\mathcal{C}_r| \leq 2^r$.

In many application areas, the user has access to a family of compression codes. For instance, in image processing, a user can tune the rate in JPEG2000. Given a family of compression codes $\mathcal{F} = \left\{(\mathcal{E}_r, \mathcal{D}_r)\right\}_{r\in\mathbb{N}}$ for set Q indexed by their rate r, let $\delta(r)$ denote the distortion performance of the code operating at rate r, i.e., $(\mathcal{E}_r, \mathcal{D}_r)$. Then, the rate-distortion function of this family of codes is defined as

$$r(\delta) \triangleq \inf\{r : \delta(r) < \delta\}.$$

Define the α-dimension of this family of codes as

$$\dim_\alpha(\mathcal{F}) \triangleq \limsup_{\delta\to 0} \frac{r(\delta)}{\log_2 \frac{1}{\delta}}. \tag{5.1}$$

We will later show that this quantity is closely connected to the number of measurements our proposed recovery methods require for accurate phase retrieval. To offer some insight on this quantity and what it measures consider the following well-known example. Let

$$\mathcal{B}_n = \left\{\boldsymbol{x} \in \mathbb{R}^n \middle| \|\boldsymbol{x}\| \leq 1\right\},$$

and

$$\mathcal{S}_{n,k} = \left\{\boldsymbol{x} \in \mathcal{B}_n \middle| \|\boldsymbol{x}\|_0 \leq k\right\}$$

denote the unit n-dimensional ball and the set of k-sparse signals in the unit ball, respectively. It is straightforward to show that the α-dimension of any family of compression codes for $\mathcal{B}_n$ and $\mathcal{S}_{n,k}$ are lower-bounded by n and k, respectively. As shown in [138], there exist compression codes that achieve these lower bounds in both cases. A straightforward generalization of this result implies that for k-sparse signals in the unit ball in $\mathbb{C}^n$, the α-dimension of any family of compression codes is lower-bounded by $2k$, and this bound is achievable.

5.1.3 Summary of our contributions

Consider the problem of noiseless phase retrieval, i.e., recovering $\boldsymbol{x}$ up to its phase from $\boldsymbol{y} = |A\boldsymbol{x}|$. To answer the first two questions we raised in Section 5.1.1, we propose COmpressible PhasE Retrieval (COPER) that employs a given compression code to solve the described phase retrieval problem. Given measurement matrix $A \in \mathbb{C}^{m\times n}$, define the distortion measure $d_A : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{R}^+$ as follows

$$\begin{aligned} d_A(\boldsymbol{x}, \boldsymbol{c}) &\triangleq \sum_{k=1}^{m} \left(\left|\boldsymbol{a}_k{}^*\boldsymbol{x}\right|^2 - \left|\boldsymbol{a}_k{}^*\boldsymbol{c}\right|^2\right)^2 \\ &= \sum_{k=1}^{m} \left(\boldsymbol{a}_k{}^*(\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*)\boldsymbol{a}_k\right)^2, \end{aligned} \tag{5.2}$$

where $\boldsymbol{a}_k{}^*$ denotes the k^{th} row of A. When there is no ambiguity about the signal of interest $\boldsymbol{x}$, we use $d_A(\boldsymbol{c})$ instead of $d_A(\boldsymbol{x}, \boldsymbol{c})$. Throughout the chapter, for complex matrix A, A^* and $\overline{A}$ denote its transposed-conjugate, and conjugate, respectively. Based on the defined distance measure, we define COPER, a non-convex optimization problem for recovering $\boldsymbol{x}$ from measurements $\boldsymbol{y}$, as follows:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{c}\in\mathcal{C}_r} d_A(\boldsymbol{x}, \boldsymbol{c}). \tag{5.3}$$

In other words, among all elements of the codebook, COPER finds the one for which $|A\boldsymbol{c}|$ is closest to measurements $\boldsymbol{y}$. Note that since $y_k = |\boldsymbol{a}_k{}^*\boldsymbol{x}|$, to calculate $d_A(\boldsymbol{x}, \boldsymbol{c})$, we do not need to know $\boldsymbol{x}$.

In phase retrieval, since the measurements are phaseless, the recovery of $\boldsymbol{x}$ can never be exact; if $\boldsymbol{x}$ satisfies $\boldsymbol{y} = |A\boldsymbol{x}|$, then so does $\mathrm{e}^{i\theta}\boldsymbol{x}$, for any $\theta \in \mathbb{R}$. Hence, following the standard procedure in the phase retrieval literature, we measure the quality of our estimate $\hat{\boldsymbol{x}}$ as

$$\inf_{\theta\in[0,2\pi)} \left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \hat{\boldsymbol{x}}\right\|^2.$$

In Section 5.2, we will bound $\inf_{\theta}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \hat{\boldsymbol{x}}\right\|^2$ in terms of the number of measurements and the rate-distortion function of the code. We show that $m > \dim_\alpha(\mathcal{F})$ observations suffice for an accurate recovery of $\boldsymbol{x}$ by COPER. For the aforementioned set of k-sparse signals that lie in the unit ball in $\mathbb{C}^n$, using a family of compression codes with an α-dimension of $2k$, our results imply that COPER requires slightly more than $2k$ noise-free phase-less measurements for an accurate recovery.

Despite the nice theoretical properties of COPER, it is not directly useful in practice as it is based on an exhaustive search over the set of all codewords, which is exponentially large. This leads us to the third question asked in Section 5.1.1. In response to this question, we introduce an iterative algorithm called gradient descent for COPER (GD-COPER). Let z_0 denote some selected initial point, and define gradient of real-valued function d as $\nabla d_A(z) \triangleq \left(\frac{\partial d_A}{\partial z}\right)^*$, where

$$\frac{\partial d_A}{\partial z} \triangleq \frac{\partial d_A(z, \overline{z})}{\partial z}\bigg|_{\overline{z}=\text{constant}},$$

is the Wirtinger derivative [25]. The iterations of GD-COPER proceed as follows:

$$\begin{aligned} \boldsymbol{s}_{t+1} &\triangleq z_t - \mu \nabla d_A(z_t), \\ z_{t+1} &\triangleq \mathcal{P}_{C_r}(\boldsymbol{s}_{t+1}), \end{aligned} \tag{5.4}$$

where t represents the iteration index. Moreover, here, for $z \in \mathbb{C}^n$,

$$\begin{aligned} d_A(z) = d_A(\boldsymbol{x}, z) &= \sum_{k=1}^{m} \left(\left|\boldsymbol{a}_k^* z\right|^2 - \left|\boldsymbol{a}_k^* \boldsymbol{x}\right|^2\right)^2 \\ &= \sum_{k=1}^{m} \left(\boldsymbol{a}_k{}^* (\boldsymbol{x}\boldsymbol{x}^* - z z^*) \boldsymbol{a}_k\right)^2, \end{aligned}$$

and therefore,

$$\nabla d_A(z) = 2 \sum_{k=1}^{m} \left(\left|\boldsymbol{a}_k^* z\right|^2 - \left|\boldsymbol{a}_k^* \boldsymbol{x}\right|^2\right) \boldsymbol{a}_k \boldsymbol{a}_k^* z.$$

Here, C_r, as defined earlier, is the set of codewords of the code, and $\mathcal{P}_{C_r} : \mathbb{C}^n \to C_r$ denotes the projection operator on this set. That is, for $\boldsymbol{s} \in \mathbb{C}^n$,

$$\mathcal{P}_{C_r}(\boldsymbol{s}) = \arg\min_{\boldsymbol{c} \in C_r} \|\boldsymbol{c} - \boldsymbol{s}\|^2.$$

We show that, under some mild conditions on the initialization, given $m > C\dim_\alpha(\mathcal{F})^2\log_2^2 n$ phase-less measurements, GD-COPER finds an accurate estimate of $\boldsymbol{x}$. Note that the number of measurements GD-COPER requires is considerably larger than what is needed by the combinatorial COPER optimization; in addition to the extra log factor, the number of measurements GD-COPER requires is proportional to $\dim_\alpha(\mathcal{F})^2$, unlike for COPER which only requires $\dim_\alpha(\mathcal{F})$ observations. While it might be the case that the difference is due to our proof techniques and the gap is not something fundamental, based on our study of the problem, it seems more plausible to us that the difference is the cost paid for having a polynomial time algorithm and cannot be closed (except for probably removing the $\log_2$ factors).

Finally, we perform extensive numerical experiments to understand the algorithmic properties of GD-COPER, and evaluate the amount of gain a compression algorithm can offer for a simple 'gradient descent'-type algorithm.

5.1.4 Related work

The problem of phase retrieval has been extensively studied in the literature [24, 25, 139, 38, 26, 140, 141, 142, 143, 21, 22, 144, 145, 146, 147, 148, 149]. (Refer to [140] for a comprehensive review of the literature.) Since, unlike compressed sensing, in phase retrieval, the measurements are a non-linear function of the input, even if the number of measurements is more than the ambient dimension of the signal, the recovery problem is still challenging. Hence, the primary focus of the field has been on developing and analyzing efficient recovery algorithms for *general* input signals. However, similar to compressed sensing, in most applications, the input signals are in fact structured. Therefore, taking such structures into account can lead to more efficient recovery algorithms with a lower number of required measurements or smaller reconstruction error. Hence, in more recent years, there has been work on phase retrieval of structured sources. In this domain, most papers are concerned with standard structures, such as sparsity. Assuming the signal is k-sparse, i.e., all of its coordinates but k of them are 0, a variety of recovery algorithms have been proposed in the literature. In the following, we briefly review some of such methods.

It is assumed in [150] that the signal is sparse, or can be approximated well with few non-zero coefficients. Furthermore, the authors suppose that l_1-norm of the signal is known, and employ an iterative phase retrieval algorithm. However, no theoretical guarantee is offered for the performance of the proposed recovery algorithm. The lifting is used in [18, 151] to convexify the problem and take advantage of semidefinite programming (SDP) for signal recovery. Since $\boldsymbol{x} \in \mathbb{C}^n$ is lifted to the space of $\mathbb{C}^{n\times n}$ matrices, the proposed algorithm is computationally demanding. Furthermore, the performance of the algorithm is guaranteed only under the assumption that the linear operator that appears in the SDP satisfies either the restricted isometry property or the coherence condition. Similarly, [152] poses the problem of sparse phase retrieval as a non-convex optimization problem and uses the alternating direction method of multipliers (ADMM) to solve the problem. Generalized approximate message passing (GAMP) has been used in [153] for the recovery of sparse signals. Despite the success of the ADMM and GAMP in simulation results, the theoretical properties of the algorithms are unknown. Inspired by the Wirtinger flow algorithm, [34] proposes a projected gradient descent for the recovery of k-sparse signals that resembles GD-COPER, proposed in this chapter. However, GD-COPER uses a generic compression algorithm, while the projected gradient descent of [34] uses the projection on the set of all k-sparse vectors. Also, by combining the alternating minimization idea with Compressive Sampling Matching Pursuit (CoSaMP) [154] has obtained another theoretically-supported algorithm for sparse phase retrieval with sample complexity of $O(k^2 \log_2 n)$. In a more general setting, [155, 36] consider the regularized PhaseMax formulation, proposed in [21, 22], and show that if a good anchor is available, then the algorithm is capable of recovering the signal from a number of measurements proportional to the minimum required number of measurements.

More recently, a few papers have used more sophisticated structures that are present in images to improve the performance of the recovery algorithms [156, 157, 158, 159]. For instance, by integrating a generic image denoiser in the iterations of the approximate message passing, similar to the approach of [160], [156] improved the performance of the approximate message passing for the recovery of images. Since the message passing framework works mainly for measurement matrices drawn independent and identically distributed (i.i.d.), [156] used the RED formulation, proposed originally in [161], for the phase retrieval. The simulation results presented in [156] suggest that the algorithms that are based on the RED formulation (and a neural net denoiser) work well on Gaussian as well as coded diffraction and Fourier measurement matrices. Similarly, [158] adds a total variation penalty to the non-convex formulation of phase retrieval problem and uses the ADMM approach for finding a local minimizer. Finally, [162] uses a deep generative network to model images and then uses the learned model as a prior to help the phase retrieval recovery algorithms.

Finally, using generic compression algorithms for compressed sensing and image restoration problems has been investigated before in [138, 163, 164, 165]. However, given the nonlinear nature of the measurement process in phase retrieval, similar to other compressed sensing methods, neither the theoretical nor the algorithmic tools and techniques developed in the area of compression-based compressed sensing are directly applicable to phase retrieval.

In this chapter, we develop a theoretical framework for phase retrieval, i.e., recovering a signal from its under-determined noise-free phase-less measurements, that is applicable to general structures employed by compression codes. This allows developing theoretically-analyzable algorithms that employ structures much beyond those that have been studied so far in the phase retrieval literature. We first propose an idealistic compression-based phase retrieval recovery method that guides us on the potential of such recovery methods. We then propose a computationally-efficient and theoretically-analyzable algorithm that given enough measurements is guaranteed to convergence to the desired solution. We also obtain an upper bound on the gap between the performance of the efficient algorithm and that of the idealistic computationally-infeasible method.

5.1.5 Organization of this chapter

The organization of the chapter is as follows. Sections 5.2 and 5.3 state and prove our main theoretical contributions regarding the performance of COPER and GD-COPER, respectively. Section 5.4 summarizes our simulation results. Finally, section 5.6 gathers lemmas and theorems we have used to obtain the main results and proves them.

5.2 Theoretical Guarantees of COPER

Consider a class of signals $Q \subset \mathbb{C}^n$ and a compression code with encoding and decoding mappings $(\mathcal{E}_r, \mathcal{D}_r)$ and codebook C_r. Using the given compression code, COPER recovers $\boldsymbol{x} \in Q$ from measurements $\boldsymbol{y} = |A\boldsymbol{x}|$ by solving the following combinatorial optimization:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{c} \in C_r} d_A(\boldsymbol{x}, \boldsymbol{c}).$$

The main goal of this section is to analyze the performance of this optimization. Toward this goal, we make the following assumptions:

1. For every $\boldsymbol{x} \in Q$, we have $\|\boldsymbol{x}\|_2 \leq 1$.[1]
2. The elements of A are i.i.d. drawn from $\mathcal{N}(0, 1) + i\mathcal{N}(0, 1)$, where i denotes the square root of -1.

The following theorem obtains an upper bound on the accuracy of the COPER's estimate.

Theorem 8. *Let $(\mathcal{E}_r, C_r)$ be a rate-r compression code with distortion δ. Let $\boldsymbol{x} \in Q$ denotes the desired signal, and define sensing matrix A, as above. Let $\hat{\boldsymbol{x}}$ denotes the solution of COPER optimization. That is, $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{c} \in C_r} d_A(\boldsymbol{x}, \boldsymbol{c})$. Then, we have*

$$\inf_{\theta} \left\| \mathrm{e}^{i\theta} \boldsymbol{x} - \hat{\boldsymbol{x}} \right\|^2 \leq 16\sqrt{3} \frac{1 + \tau_2}{\sqrt{\tau_1}} m\delta, \tag{5.5}$$

with probability at least

$$1 - 2^r \mathrm{e}^{\frac{m}{2}\left(K + \log \tau_1 - \log m\right)} - \mathrm{e}^{-2m\left(\tau_2 - \log(1+\tau_2)\right)}, \tag{5.6}$$

[1] Given the fact that we need the rate-distortion function to be finite for every $\delta > 0$, we expect Q to be a subset of $\{\boldsymbol{x} \in \mathrm{Re}^n | \|\boldsymbol{x}\|_2 \leq R\}$ for a given R. Without any loss of generality and for notational simplicity we have set $R = 1$.

where $K = \log 2\pi e$, and τ_1, τ_2 are arbitrary positive real numbers.

The general form of this theorem enables us to set τ_1, τ_2, and δ, and obtain different types of performance guarantees. Hence, before proving this theorem, we mention one specific choice that connects this result to the α-dimension of the compression code in the next corollary.

Corollary 8. *For large enough r, we have*

$$\mathbb{P}\left(\inf_{\theta}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \hat{\boldsymbol{x}}\right\|^2 \leq C\delta^{\epsilon}\right) \geq 1 - 2^{-c_\eta r} - \mathrm{e}^{-0.6m}, \tag{5.7}$$

where $C = 32\sqrt{3}$, and $m = \eta\frac{r}{\log_2 \frac{1}{\delta}}$. Given $\eta > \frac{1}{1-\epsilon}$, c_η is a positive number less than $\eta(1-\epsilon) - 1$.

Proof. Given $\epsilon > 0, \eta > 0$, in Theorem 8, let $\tau_1 = m^2\delta^{2-2\epsilon}$, and $\tau_2 = 1$. It follows that,

$$\inf_{\theta}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \hat{\boldsymbol{x}}\right\|^2 \leq 32\sqrt{3}\delta^{\epsilon}, \tag{5.8}$$

with probability

$$1 - \mathrm{e}^{r\left(\log 2 + \frac{\eta \log 2}{2\log\frac{1}{\delta}}\left(K + \log m^2\delta^{2-2\epsilon} - \log m\right)\right)} - \mathrm{e}^{-2m(1-\log 2)}. \tag{5.9}$$

Note that $1 - \log 2 > 0.3$, and

$$1 + \frac{\eta}{2\log\frac{1}{\delta}}\left(K + \log m^2\delta^{2-2\epsilon} - \log m\right) = \\ 1 + \frac{\eta(K + \log m)}{2\log\frac{1}{\delta}} - \eta(1-\epsilon).$$

Since K, η are constants, and $m \longrightarrow \eta \dim_\alpha(\mathcal{F})$ as $\delta \longrightarrow 0$. Therefore,

$$\frac{\eta(K + \log m)}{2\log\frac{1}{\delta}} \xrightarrow{\delta\to 0} 0.$$

Set any positive number c_η such that $0 < c_\eta < \eta(1-\epsilon) - 1$, so for large enough r we have

$$1 + \frac{\eta(K + \log m)}{2\log\frac{1}{\delta}} - \eta(1-\epsilon) < -c_\eta,$$

Thus

$$\mathrm{e}^{r\log 2\left(1+\frac{\eta}{2\log\frac{1}{\delta}}\left(K+\log m^2\delta^{2-2\epsilon}-\log m\right)\right)} < 2^{-c_\eta r}.$$

□

We would like to emphasize on a few points about this corollary:

Remark 22. *Corollary 8 shows that COPER recovers the signal* $\boldsymbol{x}$ *from* $\eta\dim_\alpha(Q)$ *measurements for any* $\eta > 1$ *with desired small distortion. This happens with very high probability as* $r \to \infty$*. One simple implication of this result is that, in the case of* k*-sparse complex signals, COPER needs* $2\eta k$ *measurements for almost accurate recovery. Even if we had access to the sign of* $A\mathbf{x}$*, we could not recover* $\mathbf{x}$ *accurately with less than* $2k$ *measurements. Hence, in some sense this result is sharp.*

Remark 23. *This theorem guarantees the minimizer of the COPER optimization. However, note that the COPER optimization is highly non-convex (optimization of a non-convex function over a discrete set). Hence, it is still not clear how we can get a good approximation of* $\hat{\mathbf{x}}$ *in polynomial time. This issue will be discussed in the next section.*

Next we briefly review the main steps of the proof of Theorem 8.

Roadmap of the proof of Theorem 8. Here we mention the roadmap of the proof to help the readers understand the main ideas. The details are presented in section 5.6.3. Let

$$\tilde{\boldsymbol{x}} = \mathcal{D}(\mathcal{E}(\boldsymbol{x})).$$

Clearly, $\tilde{\boldsymbol{x}} \in \mathcal{C}_r$. Note that by definition of $\delta(r)$, $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| \leq \delta(r)$. Moreover, by definition of $\hat{\boldsymbol{x}}$, we have

$$d_A\left(|A\boldsymbol{x}|, |A\hat{\boldsymbol{x}}|\right) \leq d_A\left(|A\boldsymbol{x}|, |A\tilde{\boldsymbol{x}}|\right). \tag{5.10}$$

For a complex vector $\boldsymbol{c}$, let $\lambda_1(\boldsymbol{c}), \lambda_2(\boldsymbol{c})$ denote the two non-zero eigenvalues of $\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*$. Furthermore, let $\lambda_{\max}(\boldsymbol{c})$ denote the one with the largest absolute value. In Theorem 12 (proved in Appendix B) we prove that for any positive τ_1 and τ_2 we have

$$\begin{aligned} \mathbb{P}\left(d_A(|A\boldsymbol{x}|,|A\boldsymbol{c}|) > \lambda_{\max}^2(\boldsymbol{c})\tau_1,\ \forall \boldsymbol{c} \in C_r\right) \\ \geq 1 - 2^r \mathrm{e}^{\frac{m}{2}\left(K + \log\tau_1 - \log m\right)}, \end{aligned} \tag{5.11}$$

where $K = \log 2\pi\mathrm{e}$ and

$$\begin{aligned} \mathbb{P}\left(d_A(|A\boldsymbol{x}|,|A\tilde{\boldsymbol{x}}|) < \lambda_{\max}^2(\tilde{\boldsymbol{x}})\left(4m(1+\tau_2)\right)^2\right) \\ \geq 1 - \mathrm{e}^{-2m\left(\tau_2 - \log(1+\tau_2)\right)}. \end{aligned} \tag{5.12}$$

Combining (5.10), (5.11), and (5.12), we obtain

$$\begin{aligned} \lambda_{\max}^2(\hat{\boldsymbol{x}})\tau_1 &< d_A\left(|A\boldsymbol{x}|,|A\hat{\boldsymbol{x}}|\right) \\ &\leq d\left(|A\boldsymbol{x}|,|A\tilde{\boldsymbol{x}}|\right) \\ &< \lambda_{\max}^2(\tilde{\boldsymbol{x}})\left(4m(1+\tau_2)\right)^2. \end{aligned} \tag{5.13}$$

Therefore,

$$\lambda_{\max}^2(\hat{\boldsymbol{x}}) < \frac{16m^2(1+\tau_2)^2}{\tau_1}\lambda_{\max}^2(\tilde{\boldsymbol{x}}), \tag{5.14}$$

with a probability larger than $1 - 2^r \mathrm{e}^{\frac{m}{2}\left(K + \log\tau_1 - \log m\right)} - \mathrm{e}^{-2m\left(\tau_2 - \log(1+\tau_2)\right)}$. Hence, the main remaining step is to connect $\lambda_{\max}^2(\hat{\boldsymbol{x}})$ with $\inf_\theta \left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \hat{\boldsymbol{x}}\right\|^2$. According to Lemma 52 (proved in the Appendix) we have

$$\begin{aligned}\lambda_{\max}^2(\hat{\boldsymbol{x}}) &\geq \frac{1}{2}\left(\lambda_1^2(\hat{\boldsymbol{x}}) + \lambda_2^2(\hat{\boldsymbol{x}})\right) \\ &= \frac{1}{2}\left(\|\boldsymbol{x}\|^2 - \|\hat{\boldsymbol{x}}\|^2\right)^2 + \left(\|\boldsymbol{x}\|^2\|\hat{\boldsymbol{x}}\|^2 - \left|\boldsymbol{x}^*\hat{\boldsymbol{x}}\right|^2\right).\end{aligned} \tag{5.15}$$

Recall $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\| \leq \delta$ and since $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in Q$ we have $\|\boldsymbol{x}\|, \|\tilde{\boldsymbol{x}}\| \leq 1$, thus

$$\left(\|\boldsymbol{x}\| + \|\tilde{\boldsymbol{x}}\|\right)^2 \left(\|\boldsymbol{x}\| - \|\tilde{\boldsymbol{x}}\|\right)^2 \leq 4\delta^2. \tag{5.16}$$

Moreover,

$$\begin{aligned}\delta^2 &\geq \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2 \\ &= \|\boldsymbol{x}\|^2 + \|\tilde{\boldsymbol{x}}\|^2 - \boldsymbol{x}^*\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}^*\boldsymbol{x} \\ &\geq \|\boldsymbol{x}\|^2 + \|\tilde{\boldsymbol{x}}\|^2 - 2\left|\boldsymbol{x}^*\tilde{\boldsymbol{x}}\right| \\ &\geq 2\left(\|\boldsymbol{x}\|\|\tilde{\boldsymbol{x}}\| - \left|\boldsymbol{x}^*\tilde{\boldsymbol{x}}\right|\right),\end{aligned}$$

so we have $\left(\|\boldsymbol{x}\|\|\tilde{\boldsymbol{x}}\| - |\boldsymbol{x}^*\tilde{\boldsymbol{x}}|\right) \leq \frac{\delta^2}{2}$, which implies

$$\left(\|\boldsymbol{x}\|^2\|\tilde{\boldsymbol{x}}\|^2 - \left|\boldsymbol{x}^*\tilde{\boldsymbol{x}}\right|^2\right) \leq \delta^2. \tag{5.17}$$

Similarly, Lemma 52 implies

$$\begin{aligned}\lambda_{\max}^2(\tilde{\boldsymbol{x}}) &\leq \left(\lambda_1^2(\tilde{\boldsymbol{x}}) + \lambda_2(\tilde{\boldsymbol{x}})^2\right) \\ &= \left(\|\boldsymbol{x}\|^2 - \|\tilde{\boldsymbol{x}}\|^2\right)^2 + 2\left(\|\boldsymbol{x}\|^2\|\tilde{\boldsymbol{x}}\|^2 - \left|\boldsymbol{x}^*\tilde{\boldsymbol{x}}\right|^2\right) \\ &\leq 6\delta^2.\end{aligned} \tag{5.18}$$

Therefore, combining (5.14),(5.15),(5.18), we have

$$\begin{aligned}\frac{1}{2}\left(\|\boldsymbol{x}\|^2-\|\hat{\boldsymbol{x}}\|^2\right)^2+\left(\|\boldsymbol{x}\|^2\|\hat{\boldsymbol{x}}\|^2-\left|\boldsymbol{x}^*\hat{\boldsymbol{x}}\right|^2\right) &\leq \lambda_{\max}^2(\hat{\boldsymbol{x}})\\ &< \frac{16m^2(1+\tau_2)^2}{\tau_1}\lambda_{\max}^2(\tilde{\boldsymbol{x}})\\ &\leq \frac{96m^2(1+\tau_2)^2}{\tau_1}\delta^2 \end{aligned} \tag{5.19}$$

with probability larger than $1-2^r\mathrm{e}^{\frac{m}{2}\left(K+\log\tau_1-\log m\right)}-\mathrm{e}^{-2m\left(\tau_2-\log(1+\tau_2)\right)}$. Finally, Lemma 47 connects the left hand side of (5.19) with $\left(\inf_{\theta}\left\|\mathrm{e}^{i\theta}\boldsymbol{x}-\hat{\boldsymbol{x}}\right\|^2\right)^2$. Hence, using Lemma 47 we have

$$\left(\inf_{\theta}\left\|\mathrm{e}^{i\theta}\boldsymbol{x}-\hat{\boldsymbol{x}}\right\|^2\right)^2 \leq \frac{768m^2(1+\tau_2)^2}{\tau_1}\delta^2,$$

which means

$$\begin{aligned}\mathbb{P}\left(\inf_{\theta}\left\|\mathrm{e}^{i\theta}\boldsymbol{x}-\hat{\boldsymbol{x}}\right\|^2 \leq 16\sqrt{3}\frac{1+\tau_2}{\sqrt{\tau_1}}m\delta\right)&\\ \geq 1-2^r\mathrm{e}^{\frac{m}{2}\left(K+\log\tau_1-\log m\right)}-\mathrm{e}^{-2m\left(\tau_2-\log(1+\tau_2)\right)},&\end{aligned}$$

where $K=\log 2\pi e,\ \tau_1,\tau_2>0$.

□

5.3 Theoretical Guarantees of GD-COPER

As discussed before, COPER is based on an exhaustive search over the space of all codewords, and is hence computationally very demanding, if not infeasible. This section aims to prove that with more measurements GD-COPER, introduced in Section 5.1.3, reaches a good approximation of the solution of COPER in polynomial time. In this section, we assume that

$$\|\boldsymbol{x}\|=1,\ \|z\|=1,\quad \forall\ z\in C_r. \tag{5.20}$$

This assumption enables us to state our theoretical results in a simpler form. We will have a more detailed discussion about this assumption in Section 5.5. Recall that the iterations of the GD-COPER algorithm are given by

$$\begin{aligned} \boldsymbol{s}_{t+1} &\triangleq \boldsymbol{z}_t - \mu \nabla d_A(\boldsymbol{z}_t), \\ \boldsymbol{z}_{t+1} &\triangleq \mathcal{P}_{\mathcal{C}_r}(\boldsymbol{s}_{t+1}), \end{aligned} \tag{5.21}$$

Remark 24. *The projection step in GD-COPER, i.e., $\boldsymbol{z}_{t+1} \triangleq \mathcal{P}_{\mathcal{C}_r}(\boldsymbol{s}_{t+1})$, might seem computationally expensive, as the codebook $\mathcal{C}_r$ is exponentially large. However, for a good compression code, it is natural to expect the projection on the set of codewords to be equivalent to the successive application of the encoder and the decoder mappings of the compression code. In other words, we expect $\mathcal{P}_{\mathcal{C}_r}(\cdot) = \mathcal{D}_r(\mathcal{E}_r(\cdot))$ or, at least, $\mathcal{D}_r(\mathcal{E}_r(\cdot))$ to be very close to $\mathcal{P}_{\mathcal{C}_r}(\cdot)$. We will present an example in Section 5.5 to justify this claim. We will also provide theoretical results regarding the robustness of GD-COPER to this assumption in Theorem 11. Hence, in our simulations, we use this observation and run the GD-COPER algorithm as follows:*

$$\begin{aligned} \boldsymbol{s}_{t+1} &= \boldsymbol{z}_t - \mu \nabla d_A(\boldsymbol{z}_t), \\ \boldsymbol{z}_{t+1} &= \mathcal{D}_r(\mathcal{E}_r(\boldsymbol{s}_{t+1})). \end{aligned}$$

We first mention our generic result. We will then, simplify this result in a few corollaries to interpret it and compare with the existing work.

Theorem 9. *For a fixed signal $\boldsymbol{x} \in \mathcal{Q}$, define $\boldsymbol{z}_t \in \mathcal{C}_r$ as in (5.21) with $\mu = \frac{1}{8m}$. Suppose that for all $\theta \in \mathbb{R}$, $\mathrm{e}^{i\theta}\boldsymbol{x} \in \mathcal{Q}$. Define $\theta_t \triangleq \arg\min_{\theta\in\mathbb{R}} \left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta}\boldsymbol{x}\right\|$. For all $\epsilon \geq C_2 m^{-\frac{1}{3}}$, with probability at least $1 - C_3 \mathrm{e}^{-C_1\sqrt{m\epsilon}+(3\log 2)r}$, where $C_1, C_2, C_3 > 0$ are absolute constants, for $t = 1, 2, \ldots$, we have*

$$\left\|\boldsymbol{z}_{t+1} - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| \leq \left(\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + \epsilon\right)\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + 3\delta_r. \tag{5.22}$$

Before proving this theorem, we first simplify the statement of this theorem and compare it with

Corollary 8. The following Corollary shows having enough measurements, we may get arbitrary close to the COPER's solution with this algorithm, with exponentially high probability.

Corollary 9. *Consider the same setup as in Theorem 9. Assume that* $\inf_{\theta\in\mathbb{R}}\left\|e^{i\theta}\boldsymbol{x} - z_0\right\| = 1 - 2\tau < 1$, *for some* $\tau > 0$. *Then, if* $\delta \leq \frac{\tau(1-2\tau)}{3}$, *and*

$$m \geq \max\left\{\left(\frac{C_2}{\tau}\right)^3, \frac{C_4}{\tau}(\dim_\alpha(\mathcal{F})\log_2\frac{1}{\delta})^2\right\},$$

after T iterations of GD-COPER,

$$\inf_{\theta\in\mathbb{R}}\left\|e^{i\theta}\boldsymbol{x} - \boldsymbol{z}_T\right\| \leq (1-2\tau)(1-\tau)^T + \frac{3}{\tau}\delta_r, \tag{5.23}$$

with probability at least

$$1 - C_3 e^{-\frac{C_1\sqrt{\tau}}{2}\sqrt{m}}. \tag{5.24}$$

Here, C_1, C_2, C_3 *are the constants introduced in Theorem 9 with* $\epsilon = \tau$, *and* C_4 *is an absolute constant.*

Proof. We apply Theorem 9 with $\epsilon = \tau$, thus we need $\tau = \epsilon \geq C_2 m^{-\frac{1}{3}}$, hence

$$m \geq \left(\frac{C_2}{\tau}\right)^3. \tag{5.25}$$

With a probability larger than $1 - C_3 e^{-C_1\sqrt{m\tau}+(3\log 2)r}$ at each iteration we have

$$\left\|z_{t+1} - e^{i\theta_{t+1}}\boldsymbol{x}\right\| \leq \left(\left\|z_t - e^{i\theta_t}\boldsymbol{x}\right\| + \tau\right)\left\|z_t - e^{i\theta_t}\boldsymbol{x}\right\| + 3\delta, \tag{5.26}$$

hence,

$$\begin{aligned}\left\|\boldsymbol{z}_{t+1} - \mathrm{e}^{i\theta_{t+1}}\boldsymbol{x}\right\| &\leq \left(\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + \tau\right)\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + 3\delta \\ &\leq (1-\tau)(1-2\tau) + 3\delta \\ &\leq 1 - 2\tau, \end{aligned} \tag{5.27}$$

since $\delta \leq \frac{\tau(1-2\tau)}{3}$. Therefore, by (5.26) and (5.27), we may deduce that

$$\left\|\boldsymbol{z}_{t+1} - \mathrm{e}^{i\theta_{t+1}}\boldsymbol{x}\right\| \leq (1-\tau)\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + 3\delta.$$

Hence we get,

$$\begin{aligned}\left\|\boldsymbol{x} - \mathrm{e}^{i\theta_T}\boldsymbol{z}_T\right\| &\leq (1-\tau)^T\left\|\mathrm{e}^{i\theta_0}\boldsymbol{x} - \boldsymbol{z}_0\right\| \\ &\quad + 3\delta\left(1 + 1 - \tau + (1-\tau)^2 + \ldots + (1-\tau)^{T-1}\right) \\ &\leq (1-2\tau)\,(1-\tau)^T + \frac{3\delta}{\tau}. \end{aligned} \tag{5.28}$$

Moreover, if $\mathcal{G}$ denotes the event under which Theorem 9 holds, i.e. (5.22) is satisfied, then

$$\begin{aligned}\mathbb{P}\left(\mathcal{G}\right) &\geq 1 - C_3\mathrm{e}^{-C_1\sqrt{m\tau}+(3\log 2)r} \\ &\geq 1 - C_3\mathrm{e}^{-\frac{C_1\sqrt{\tau}}{2}\sqrt{m}}, \end{aligned}$$

once we have $(3\log 2)r \leq \frac{C_1\sqrt{\tau m}}{2}$, or equivalently

$$m \geq \frac{C_4'}{\tau}r^2, \quad C_4' = \left(\frac{6\log 2}{C_1}\right)^2.$$

Since $\lim\limits_{r\to\infty}\frac{r}{\log_2\frac{1}{\delta}} = \dim_\alpha(\mathcal{F})$, for large enough r, we have $r \leq 1.5\dim_\alpha(\mathcal{F})\log_2\frac{1}{\delta}$. Hence,

$$m \geq \frac{C_4}{\tau}\left(\dim_\alpha(\mathcal{F})\log_2\frac{1}{\delta}\right)^2, \tag{5.29}$$

where $C_4 = 2.25C'_4$.

Since we assumed $m \geq \max\left(\left(\frac{C_2}{\tau}\right)^3, \frac{C_4}{\tau}\left(\dim_\alpha(\mathcal{F}) \log_2 \frac{1}{\delta}\right)^2\right)$, both (5.25) and (5.29) are satisfied. Then by (5.28) we obtain

$$\inf_{\theta\in\mathbb{R}} \left\| e^{i\theta}\boldsymbol{x} - z_T \right\| \leq (1-2\tau)(1-\tau)^T + \frac{3\delta}{\tau}. \tag{5.30}$$

□

Remark 25. *Consider the same setup as in Corollary 9 and let $\tau = \frac{1}{4}$. Then, for $\delta \leq \frac{1}{24}$, and*

$$m \geq \max\left\{(4C_2)^3, 4C_4(\dim_\alpha(\mathcal{F}) \log_2 \frac{1}{\delta})^2\right\},$$

after T iterations of the GD-COPER algorithm,

$$\inf_{\theta\in\mathbb{R}} \left\| e^{i\theta}\boldsymbol{x} - z_T \right\| \leq \frac{1}{2}\left(\frac{3}{4}\right)^T + 12\delta, \tag{5.31}$$

with a probability greater than $1 - C_3 e^{-\frac{C_1}{4}\sqrt{m}}$, where $C_i,\ i \in \{1, \ldots, 4\}$, are positive constants.

Remark 26. *If n is a large number (which is the case in almost all the applications of the phase retrieval), then we can set $\delta = \frac{1}{n}$ in Remark 25, and conclude that with $m \geq C'_4 \dim_\alpha(\mathcal{F})^2 \log_2^2 n$ measurements, GD-COPER can with high probability obtain an accurate estimate of $\boldsymbol{x}$ (with $O(1/n)$ distortion). Hence, the number of measurements GD-COPER requires is substantially more than the number of observations COPER requires. At this stage, it is not clear whether this discrepancy is an artifact of our proof technique, the limitation of the GD-COPER algorithm, or a fundamental limitation of the polynomial time algorithms. We leave the full study of this phenomenon for future research. We should also mention that in the case of sparse phase retrieval [34] observed that even under a good initialization the thresholded Wirtinger flow algorithm can recover the signal exactly with $k^2 \log_2 n$ measurements, which is again consistent with our result. Furthermore, the paper presented other evidences to suggest that to obtain a good initialization $k^2 \log_2 n$ measurements are required. It is also worth mentioning that [166] has shown that convex relaxation methods will not* *work if the number of measurements is less than $ck^2/\log_2 n$ for constant c.*

Remark 27. *Corollary 9 proves the accuracy of GD-COPER under an initialization that satisfies $\inf_{\theta\in\mathbb{R}} \left\| e^{i\theta}\boldsymbol{x} - \boldsymbol{z}_0 \right\| = 1 - 2\tau < 1$. Finding an initialization that theoretically satisfies this condition is a good research direction for future research. However, as will be clarified in our simulation results and has also be discussed elsewhere [145], the choice of initialization seems to have a minor effect (almost none) on the performance of GD-COPER (and other iterative algorithms). Hence, in our simulation results we have initialized GD-COPER with a white image.*

Roadmap of the proof of Theorem 9. Let $\tilde{\boldsymbol{x}} = \mathcal{P}_{C_r}(e^{i\theta_t}\boldsymbol{x})$. Since $\boldsymbol{z}_{t+1} = \mathcal{P}_{C_r}(\boldsymbol{s}_{t+1})$ and $\tilde{\boldsymbol{x}} \in C_r$, we have

$$\begin{aligned} \|\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}}\|^2 \geq \|\boldsymbol{s}_{t+1} - \boldsymbol{z}_{t+1}\|^2 \\ = \|\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}}\|^2 + \|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|^2 \\ + 2\mathrm{Re}\left((\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})^*(\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}})\right). \end{aligned} \tag{5.32}$$

Therefore,

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|^2 \leq 2\mathrm{Re}\left((\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})^*(\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1})\right). \tag{5.33}$$

Let

$$\boldsymbol{v}_t \triangleq \frac{\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}}{\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|}. \tag{5.34}$$

Using this definition, (5.33) can be written as

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\| \leq 2\mathrm{Re}\left(\boldsymbol{v}_t^*(\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1})\right). \tag{5.35}$$

Recall that $\mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right] = 8m(\boldsymbol{z}\boldsymbol{z}^* - \boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{z}$. Hence,

$$
\begin{aligned}
\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1} &= \tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x} + \mathrm{e}^{i\theta_t}\boldsymbol{x} - \Bigg(\boldsymbol{z}_t - \frac{1}{8m}\mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right] \\
&\quad + \frac{1}{8m}\left(\mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right] - \nabla d_A(\boldsymbol{z}_t)\right)\Bigg) \\
&= \tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x} + \mathrm{e}^{i\theta_t}\boldsymbol{x} - \left(\boldsymbol{z}_t - \boldsymbol{z}_t + (\boldsymbol{x}^*\boldsymbol{z}_t)\boldsymbol{x}\right) \\
&\quad + \frac{1}{8m}\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right) \\
&= \tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x} + (1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^*\boldsymbol{z}_t)\mathrm{e}^{i\theta_t}\boldsymbol{x} \qquad (5.36) \\
&\quad + \frac{1}{8m}\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right).
\end{aligned}
$$

Note that $\left\|\tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| \leq \delta_r$. Also, since by lemma 46 we have $1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^*\boldsymbol{z}_t = \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2$ and $\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| = \|\boldsymbol{v}_t\| = 1$, by the triangle inequality, from (5.35) and (5.36), we have

$$
\begin{aligned}
\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_{t+1}\right\| &\leq \left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \tilde{\boldsymbol{x}}\right\| + \|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\| \\
&\leq \delta_r + 2\mathrm{Re}\left(\boldsymbol{v}_t^*(\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1})\right) \\
&\leq \delta_r + 2\|\boldsymbol{v}_t\|\left\|\tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| \\
&\quad + 2(1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^*\boldsymbol{z}_t)\|\boldsymbol{v}_t\|\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| \\
&\quad + \frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}_t^*\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right) \\
&\leq \delta_r + 2\delta_r + \left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2 \qquad (5.37) \\
&\quad + \frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}_t^*\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right).
\end{aligned}
$$

Define event $\mathcal{G}$ as follows

$$
\begin{aligned}
\mathcal{G} \triangleq \Bigg\{&\frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}^*\left(\nabla d_A(\boldsymbol{z}) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right]\right)\right) \\
&\leq \epsilon \inf_{\theta\in\mathbb{R}}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z}\right\|, \quad \boldsymbol{v} = \frac{\tilde{\boldsymbol{x}} - \boldsymbol{z}'}{\|\tilde{\boldsymbol{x}} - \boldsymbol{z}'\|}, \quad \forall \tilde{\boldsymbol{x}}, \boldsymbol{z}, \boldsymbol{z}' \in \mathcal{C}_r\Bigg\}. \qquad (5.38)
\end{aligned}
$$

One difficulty in bounding $\mathbb{P}\left(\mathcal{G}\right)$ is that $\nabla d_A(z)$ is summation of heavy-tailed random variables. To address this issue, in Lemma 54 (stated and proved in Section 5.6.4), we develop a technique that yields sharp concentration bounds for such summations. Applying Lemma 54 with 4ϵ, for a given $\boldsymbol{v} \in \mathbb{C}^n$ with $\|\boldsymbol{v}\| = 1$ and $\boldsymbol{z} \in \mathcal{C}_r$, we get constants $C_1, C_2, C_3 > 0$ for which, for every $\epsilon \geq C_2 m^{-\frac{1}{3}}$,

$$\begin{aligned} \mathbb{P}\Bigg\{\bigg|\mathrm{Re}\left(\boldsymbol{v}^*\left(\nabla d_A(\boldsymbol{z}) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right]\right)\right)\bigg| \\ > 4m\epsilon \inf_{\theta\in\mathbb{R}}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z}\right\|\Bigg\} \leq C_3 \mathrm{e}^{-C_1\sqrt{m\epsilon}}. \end{aligned} \tag{5.39}$$

Hence, combining (5.39) with the union bound, for every $\epsilon \geq C_2 m^{-\frac{1}{3}}$, we have

$$\mathbb{P}\left(\mathcal{G}\right) \geq 1 - 2^{3r} C_3 \mathrm{e}^{-C_1\sqrt{m\epsilon}}. \tag{5.40}$$

Therefore, conditioned on $\mathcal{G}$ we have

$$\begin{aligned} & \frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}_t^*\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right) \\ \leq \;& \epsilon \inf_{\theta\in\mathbb{R}}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z}_t\right\| \\ = \;& \epsilon\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|, \end{aligned}$$

hence, (5.37) implies that, for all $t \in \{1 \cdots, T\}$,

$$\left\|\boldsymbol{z}_{t+1} - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| \leq \left(\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + \epsilon\right)\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + 3\delta_r, \tag{5.41}$$

which in turn leads to (5.22).

□

5.4 Simulation results

The main goal of this section is to experimentally evaluate the performance of our algorithm. Furthermore, comparisons between our algorithm and Wirtinger flow will be presented to empirically evaluate the amount of gain a compression scheme offers. Since the publicly available compression algorithms work with real-valued images, in our simulation results we focus on real-valued signals and measurements only. Note that even though our theoretical results are presented for complex-valued signals, the extension to real-valued signals is straightforward. For the sake of brevity, we did not include such extensions.

5.4.1 Measurement matrices

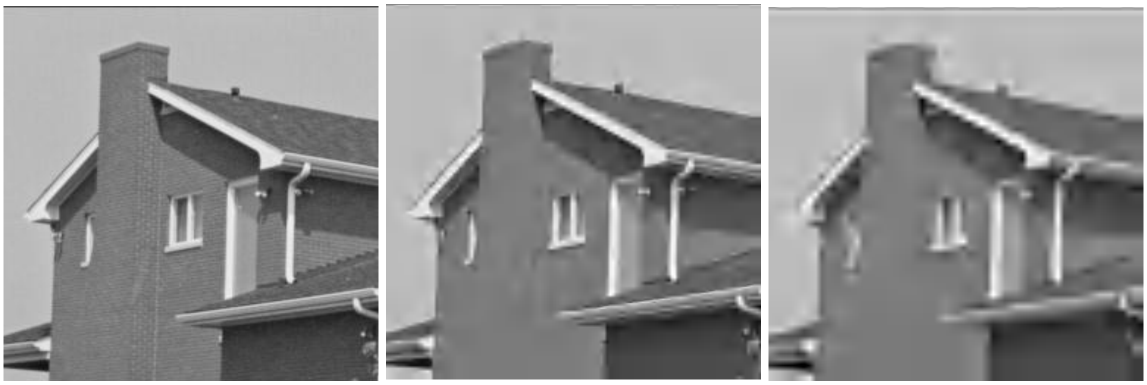
We consider two types of measurement matrices: (i) Gaussian measurement matrices in which $A_{ij} \overset{iid}{\sim} N(0,1)$, and (ii) coded diffraction patterns in which the measurements are constructed in the following way:

$$y_{i,l} = \left| \sum_{k=1}^{n} x_k \cos\left(\frac{i\pi}{n}\left(k + \frac{1}{2}\right)\right) M_{l,k} \right|. \tag{5.42}$$

In these measurements, $M_{l,k}$ modulates the entries of the signal and is drawn from the following distribution:

$$M_{l,k} \overset{iid}{\sim} \begin{cases} 1 & \text{with probability} \quad \frac{1}{4} \\ -1 & \text{with probability} \quad \frac{1}{4} \\ 0 & \text{with probability} \quad \frac{1}{2} \end{cases}, \qquad 1 \leq k \leq n, \quad 1 \leq l \leq L. \tag{5.43}$$

Coded diffraction patterns have recently received attention in the phase retrieval problem since they can outperform the Fourier matrices. Note that due to the construction of the coded diffraction measurement matrices, the imaging system is over-sampled by the factor L. Our simulation results will cover $L \in \{1, 2, \ldots, 15\}$. As we will discuss later, GD-COPER algorithm is capable of



(a) q = 0 (b) q = 50 (c) q = 100

Figure 5.1: Compression with different quality-layers

performing well even when $L = 1$. Note that given that the signs are missing, this can be considered as an under-sampled situation.

5.4.2 Setting the parameters

GD-COPER

In our simulation results, we will be using natural images, and JPEG2000 compression algorithm. In particular, we have used a python implementation of the JPEG2000 which is a part of the PIL package available at : `https://pillow.readthedocs.io/en/3.1.x/handbook/image-file-formats.html#jpeg-2000`. The compression algorithm has multiple inputs. The first one is the image itself. The other parameter that is important in our implementation is the parameter "quality-layer", denoted by q in this chapter, that controls compression ratio (or equivalently the rate). Figure 5.1 shows the result of the compression-decompression for three different values of the parameter q. It is clear from this figure that as q decreases, the distortion in the reconstructed image reduces. The value $q = 0$ corresponds to the lossless compression.

The GD-COPER algorithm has three different parameters that require tuning: (i) initialization, (ii) the quality parameter of the compression algorithm q at every iteration, and (iii) the step size μ. As will be discussed in Section 5.4.4, our algorithm is not sensitive to the the initialization and in our simulation results, we start the algorithm with the white image. Hence, in this section, we only describe how we set the step-size μ_t and the quality parameter q_t at every iteration. The problem of parameter tuning for iterative algorithms is a challenging problem that has not been settled properly yet [167]. Hence, after doing multiple runs of the algorithm we have found a set of parameters that work well in practice. Below we summarize the chosen parameters for the Gaussian and coded diffraction patterns. We should emphasize that better tuning are expected to improve the performance of GD-COPER. Below we discuss our choice of parameters for the Gaussian and coded diffraction patterns separately.

- Gaussian matrices: The “quality-layer" and step-size parameters at step t are set in the following way:

$$\begin{cases} q_t = 40, \quad \mu_t = .2 \times \frac{\|z_t\|}{\|\nabla d_A(z_t)\|} & 1 \leq t \leq 10 \\ q_t = 0, \quad \mu_t = .02 \times \frac{\|z_t\|}{\|\nabla d_A(z_t)\|} & t \geq 11 \end{cases} \tag{5.44}$$

  Note that $q_t = 0$ means that the algorithm employs an almost-lossless compression. The main reason an almost-lossless compression is used in the final iteration is that we are considering noiseless observations. We run the GD-COPER for 50 iterations, since the error does not decrease much after that.

- Coded diffraction patterns: The value of parameters we chose for the coded-diffraction patterns is somewhat different from the ones we chose for Gaussian matrices. For such matrices, we adopt the following parameters:

$$\mu_t = \max\left(\mathrm{e}^{0.7-0.41t}, 0.02\right) \times \frac{\|z_t\|}{\|\nabla d_A(z_t)\|},$$
$$\begin{cases} q_t = 50 & 1 \leq t \leq 5 \\ q_t = 20 & 6 \leq t \leq 30 \\ q_t = 0 & t > 30. \end{cases} \tag{5.45}$$

We run the GD-COPER for 50 iterations, since the error remains almost the same for further iterations. Again these parameters are obtained by comparing a number of choices and choosing the one that seems to perform well on a wide range of images and problem instances.

Setting the parameters of Wirtinger flow

The following parameters of the Wirtinger flow require tuning: (i) initialization, (ii) step size. Most of the papers, including [25] suggest using the spectral method for the initialization of the algorithm. Our simulation results, some of which are reported in Section 5.4.5, show that the algorithm works better when it is initialized with the white image. Hence, in all our simulations, except the ones in Sections 5.4.5 and 5.4.4, we initialize the algorithm with a white image.

For setting the step size, we follow the suggestions of [25], and adopt the following policy:

$$\mu_t = \min\left(1 - \mathrm{e}^{-\frac{t}{t_o}}, \mu_{\max}\right), \tag{5.46}$$

where $t_o = 330$, $\mu_{\max} = 0.4$. Moreover, 300 iterations are used in all runs of Wirtinger Flow (this is the number which is suggested in the simulations of [25]) except for the cases that due to the divergence of algorithm the machine terminates the run earlier. Divergence happens when the norm of z_t goes to infinity.

5.4.3 Results

We present our results for Gaussian and coded diffraction patterns separately.

Gaussian measurement matrices

In our simulations, we consider seven different images shown in the first column of Table 5.1. All these images are chosen from “The Miscellaneous volume data-set”, which is publicly available at http://sipi.usc.edu/database/database.php?volume=misc. Since the images are colored we have extracted the luminance of the image and all the simulations are performed on gray-scale images.

To reduce the computational complexity of our algorithm (in the case of Gaussian measurements only) we downsample images to reduce their size to 128×128. This size reduction helps us avoid the issues we face in storing i.i.d. Gaussian matrices. However, it also reduces the structures that exist in an image. Hence, JPEG2000 loses some of its efficiency. Hence, we expect the GD-COPER to perform better as the image size increases. This will become clearer when we work with larger images in the coded diffraction pattern simulations.

After the downsampling, the signals' dimensions are $n = 16384$. In Table 5.1, we have considered $m = 32786, 16384, 12000$, and 8192. Note that, in most of these systems, not only the measurements are phaseless, but also they are undersampled.

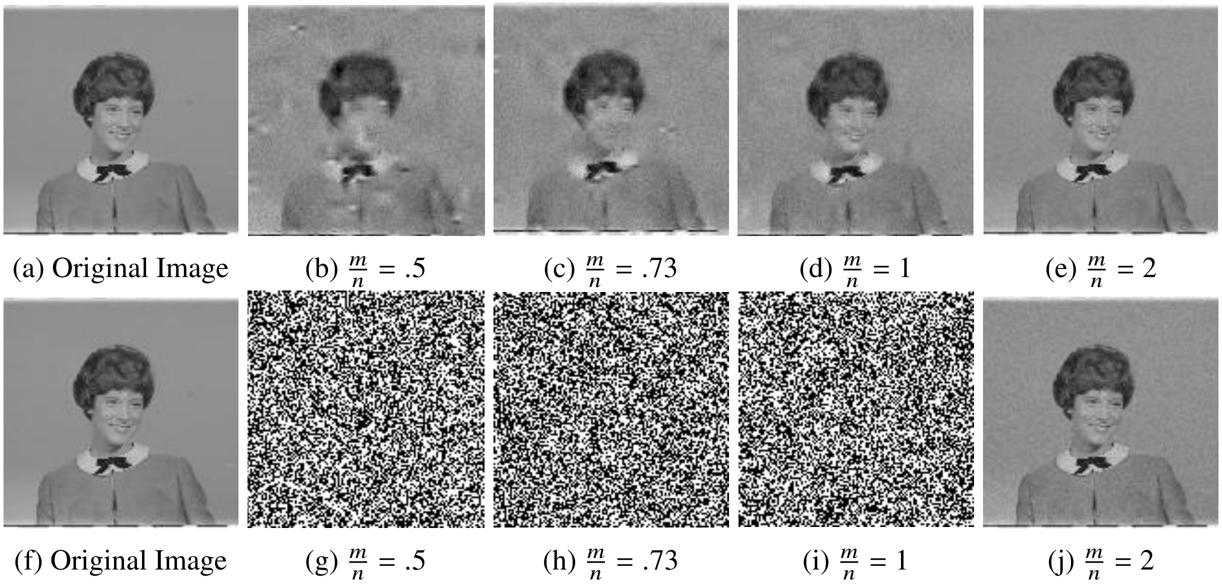
In each setup, we compare the performance of our algorithm with that of the Wirtinger flow. In addition to comparing the quality of the reconstruction via evaluating the peak-signal-to-noise-ratio (PSNR),[2] we report the run time of the algorithms as well. The run times of the algorithm are measured on a laptop computer with 2.8 GHz Intel Core i7 processor and 16 GB RAM. We can draw the following conclusions from the results reported in Table 5.1:

(i) As expected, the Wirtinger flow does not do well when $\frac{m}{n} \leq 1$. This is in contrast to the performance of GD-COPER that can obtain reasonable estimates even for $m/n \leq 1$. Note that in some cases, the Wirtinger flow can do as well as GD-COPER when $\frac{m}{n} = 2$. This happens because we have downsampled the images to 128×128 size, and hence we have removed some structures that JPEG2000 could otherwise employ. In other words, JPEG2000 cannot efficiently reduce the size of such images, and hence GD-COPER is not capable of employing the structures of such images either. In the next section, GD-COPER works with large images (we can do this with coded diffraction patterns), and will outperform the Wirtinger flow with a larger margin. See Figure 5.2 for a visual comparison of the GD-COPER and Wirtinger flow algorithms.

[2] PSNR is defined as

$$\text{PSNR} = 20 \log_2 10 \left(\frac{255}{\sqrt{\text{MSE}}} \right),$$

where MSE is the mean squared error obtained from the last iteration of the algorithm.



(a) Original Image (b) $\frac{m}{n} = .5$ (c) $\frac{m}{n} = .73$ (d) $\frac{m}{n} = 1$ (e) $\frac{m}{n} = 2$

(f) Original Image (g) $\frac{m}{n} = .5$ (h) $\frac{m}{n} = .73$ (i) $\frac{m}{n} = 1$ (j) $\frac{m}{n} = 2$

Figure 5.2: First row: outcomes of GD-COPER for different values of m/n. Second row: outcomes of Wirtinger Flow for different values of m/n. The original image is shown in the left column. The measurement matrix is Gaussian.

(ii) GD-COPER is faster than the Wirtinger flow. Note that each iteration of GD-COPER is computationally more demanding than that of the Wirtinger flow. However, GD-COPER requires less steps to obtain a good estimate of the signal. Figure 5.3 compares the normalized MSE (we have normalized the mean square error, by the energy of the underlying signal) of GD-COPER as a function of the iteration number with that of the Wirtinger flow. We can see that GD-COPER converges in 10 iteration, while Wirtinger flow requires around 200 iterations to get to a comparable error if it does not diverge.

Coded diffraction model

In this section, we evaluate the performance of our algorithm on the more practical coded-diffraction measurements. Again, we work with the seven images we introduced in the last section. However, given the fact that in the case of the coded diffraction patterns the measurement matrix is not explicitly stored we will use images in their original sizes, 256 × 256. We compare the performance of GD-COPER with that of the Wirtinger flow for different m/n ratios. Tables 5.2 and 5.3 summarize our simulation results. Again we should emaphasize that both the GD-COPER and

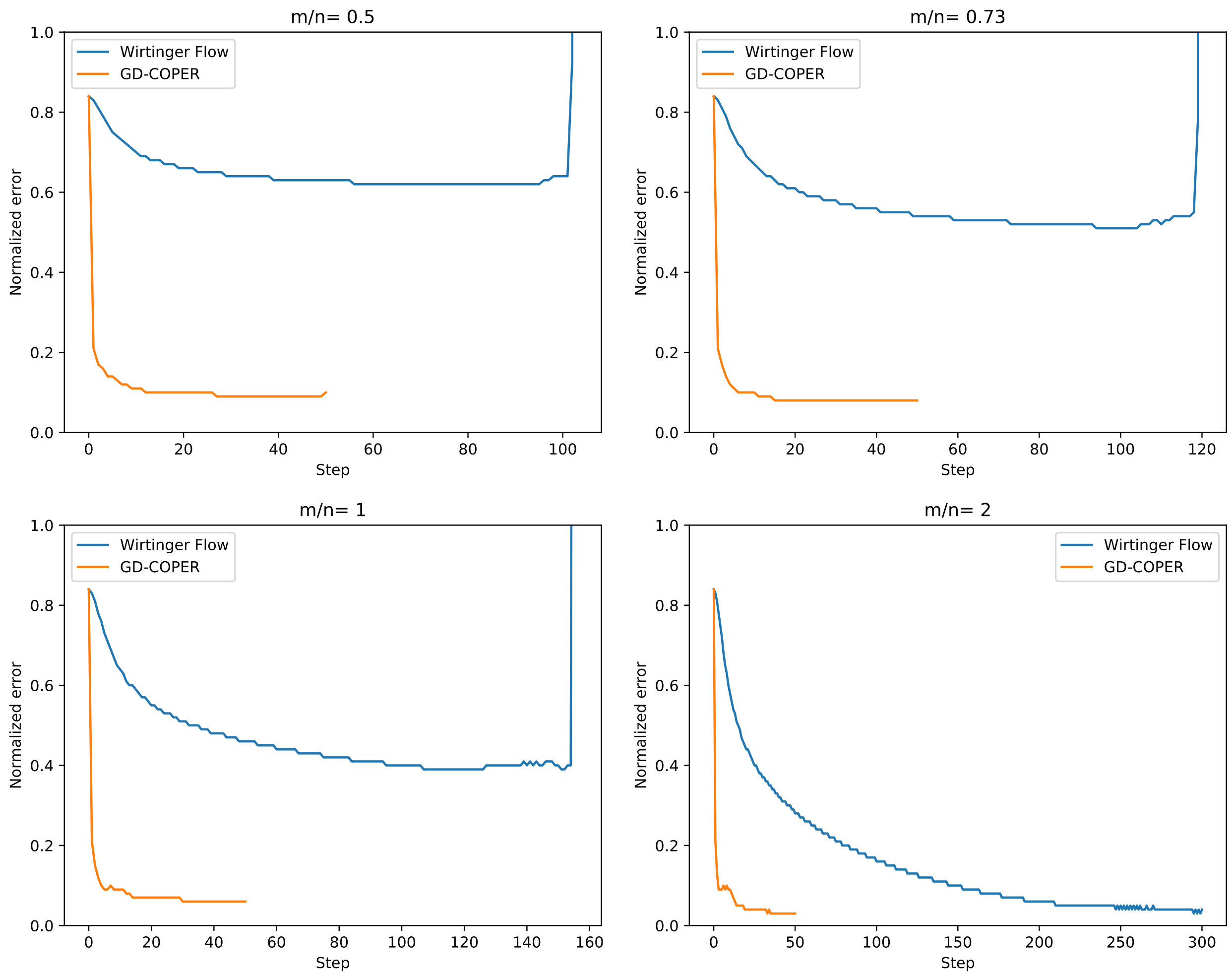


Figure 5.3: Normalized mean square error as a function of the iteration number for four different values of m/n. The original image is the same as the one chosen for Figure 5.2.

Wirtinger flow are initialized with an all-white image. We can draw the following conclusion from Tables 5.2 and 5.3:

1. Again for lower values of the sampling ratio m/n such as $m/n \leq 5$ Wirtinger flow is not capable of finding a good estimate. However, GD-COPER obtains an accurate estimate for $m/n \leq 5$, and even for $m/n = 1$. If we compare these simulations with the ones we had for Gaussian matrices, it seems to be the case that the discrepancy between the performance of the Wirtinger flow and GD-COPER has increased in the coded-diffraction simulation. Part of this is a result of the fact that our simulations have been performed on larger images for which JPEG2000 is more efficient.

2. As we increase the number of masks, usually after $L = 10$ the performance of GD-COPER saturates, while Wirtinger flow continues to improve. There are two effects that cause the saturation of the GD-COPER: (i) Given that the compression is applied at every iteration, even though it is in its loss-less mode, it still imposes some quantization to the estimates. (ii) Suboptimal parameter tuning. We believe that the performance saturation of the algorithm does not cause a major issue in practice since it happens at very high values of PSNR, e.g. 40 dB. However, as a result of the saturation, we see that in most cases, when $m/n > 15$, then the Wirtinger flow outperforms GD-COPER (for the sake of brevity we have not included the results of $m/n > 15$ in our tables). Hence, if extremely accurate estimates of the signal are required (e.g. PSNR= 50 dB) and we have enough masks, then the Wirtinger flow should be preferred over GD-COPER.

3. The computational complexity of GD-COPER is comparable with that of the Wirtinger flow. Note that each iteration of GD-COPER is computationally more demanding than that of the Wirtinger flow. However, GD-COPER requires less steps to obtain a good estimate of the signal.

5.4.4 Robustness of GD-COPER with respect to initialization

As we discussed in Section 5.4.2, the performance of GD-COPER is not sensitive to the initialization. In this section, we present some of our evidence that supports this claim. Given that our simulation results are similar for both coded diffraction patterns and Gaussian measurements, we only report our simulations for the coded diffraction patterns. In order to observe the impact of initialization we considered the following initialization: Let $\boldsymbol{x}$ denote the underlying signal we want to recover, and let $\boldsymbol{x}_o$ denote the vector that corresponds to an all-white image. A simple initialization that we can use in practice is $\boldsymbol{x}_o$, while the best oracle-initialization is $\boldsymbol{x}$. Hence, we can consider the family of initializations

$$\boldsymbol{x}_{\text{init}} = \lambda \boldsymbol{x}_o + (1 - \lambda)\boldsymbol{x},$$

for $\lambda \in \{0, 0.1, 0.2, \ldots, 0.9, 1\}$. We expect the smaller values of λ to return better initializations. Tables 5.6 and 5.7 evaluate the performance of GD-COPER for different initializations and different images. The other parameters of GD-COPER are set according to the strategy described in Section 5.4.2. As is clear from our simulation results, the initialization schemes have much larger impacts on the Wirtinger flow compared to GD-COPER. In fact, the GD-COPER is not very sensitive to the choice of initialization and in most cases, the difference between the best initialization and worst initialization is less than 2 dB. In contrast to GD-COPER, the performance of the Wirtinger flow is very sensitive to the choice of the initialization. For this reason, the spectral method is often used for the initialization of the Wirtinger flow algorithm. In the next section, we will show that the initialization of the Wirtinger flow algorithm with an all-white image is often better than the spectral initialization. However, we should emphasize that this phenomenon is only true for the real-valued signals, and has not been tested on complex-valued signals.

5.4.5 Spectral Initialization

Another claim we made in Section 5.4.2 regarding the initialization was the fact that Spectral initialization does not seem to help the Wirtinger flow beyond what is offered by an all-white image initialization. We show part of our evidence regarding this claim. Tables 5.4 - 5.5 summarize some of our findings. In these table the ‘n-init-err’ shows the normalized mean square error of the initialization. Note that in most cases the spectral methods does not offer a closer point than the all-white image except when we have $\frac{m}{n} \geq 7$. Moreover, when we have many observations and the initial point offered by the Spectral method is closer than the white image, Wirtinger Flow usually performs better starting from the white image. This shows the initial distance is not the only important factor to the convergence of Wirtinger flow (this is an artifact of the fixed parameter tuning that has been proposed for the Wirtinger flow). For instance, if the norm of the gradient at the starting steps, when the step-size defined in (5.46) is large, remains high, then the algorithm may diverges.

5.5 Discussion of our assumptions

In the proof of the convergence of GD-COPER in Theorem 9, we made two assumptions:

1. $\|\boldsymbol{x}\|_2^2 = 1$ and $\|\boldsymbol{z}\|_2^2 = 1$ for all $\boldsymbol{z} \in C_r$.
2. $\mathcal{P}_{C_r}(\cdot) = \mathcal{D}_r(\mathcal{E}_r(\cdot))$, i.e. the application of the encoder and decoder of a compression algorithm is equivalent to projecting a signal on the closest code-word.

First, note that we can obtain a good estimate of the norm of the signal, and normalize the measurements and pretend that the signal satisfies $\|\boldsymbol{x}\|_2^2 = 1$. Below we present one approach to execute this normalization. Suppose that $\boldsymbol{y} = |A\boldsymbol{x}|$. We have

$$\mathbb{E}\left[\left|y_k\right|^2\right] = \mathbb{E}\left[y_k^* y_k\right] = \mathbb{E}\left[\boldsymbol{x}^* \boldsymbol{a}_k^* \boldsymbol{a}_k \boldsymbol{x}\right] = 2n\|\boldsymbol{x}\|^2 .$$

Hence, $\frac{1}{2nm}\left\|\boldsymbol{y}\right\|^2 \xrightarrow{\mathbb{P}} \|\boldsymbol{x}\|^2$, where the notation $\xrightarrow{\mathbb{P}}$ denotes convergence in probability. Hence, if we divide our measurements by $\sqrt{\frac{1}{2nm}\left\|\boldsymbol{y}\right\|^2}$, then we can assume that $\|\boldsymbol{x}\|_2 = 1$. Once we know that the magnitude of the signal is equal to one, we can modify any compression algorithm to have the property $\|\boldsymbol{z}\|_2^2 = 1$ for all $\boldsymbol{z} \in C_r$, by dividing the output of the decoder by its magnitude. One question that we still have to address though is the following: Often times the estimate of the magnitude of the signal is random and may deviate from what we expect. Hence, we may end up having a signal $\boldsymbol{x}$ whose magnitude satisfies $|\|\boldsymbol{x}\|_2^2 - 1| \leq \gamma$ where γ is a small number. What would be the impact of such an error in the performance of GD-COPER? In particular, one would hope that this error does not accumulate in the iterations of the algorithm. Our next theorem proves this claim.

Theorem 10. *Consider a fixed signal* $\boldsymbol{x} \in Q$ *that satisfies* $|\|\boldsymbol{x}\|_2^2 - 1| \leq \gamma$. *Define* $\boldsymbol{z}_t \in C_r$ *as in* (5.21) *with* $\mu = \frac{1}{8m}$. *Suppose that for all* $\theta \in \mathbb{R}$, $\mathrm{e}^{i\theta}\boldsymbol{x} \in Q$. *Define* $\theta_t \triangleq \arg\min_{\theta \in \mathbb{R}} \left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta}\boldsymbol{x}\right\|$. *For all* $\epsilon \geq C_2 m^{-\frac{1}{3}}$, *with probability at least* $1 - C_3 \mathrm{e}^{-C_1\sqrt{m\epsilon} + (3\log 2)r}$, *where* $C_1, C_2, C_3 > 0$ *are absolute* *constants, for $t = 1, 2, \ldots$, we have*

$$\left\| \boldsymbol{z}_{t+1} - \mathrm{e}^{i\theta_t}\boldsymbol{x} \right\| \leq \left(\left\| \boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x} \right\| + \epsilon \right) \left\| \boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x} \right\| + 3\delta_r + \gamma. \tag{5.47}$$

Proof. Note that we still assume that for all $\boldsymbol{z} \in \mathcal{C}_r$, we have $\|\boldsymbol{z}\|_2^2 = 1$, since this condition is straightforward to satisfy exactly (given that the output of decoder is available and hence we can directly normalize it). Since the proof of this theorem is similar to the proof of Theorem Theorem 9 we skip most of the steps, and only mention the ones that are different. By following the steps in the proof of Theorem 9 that led to (5.36) we obtain

$$\begin{aligned} \tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1} = \tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x} + (1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^* \boldsymbol{z}_t)\mathrm{e}^{i\theta_t}\boldsymbol{x} \\ + \frac{1}{8m}\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t) \right] \right). \end{aligned}$$

In the proof of Theorem 9, we claimed that $(1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^* \boldsymbol{z}_t) = \frac{1}{2}\left\| \mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t \right\|^2$. Clearly, this is not true any more. Instead we have

$$\frac{1}{2}\left\| \mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t \right\|^2 = \frac{1}{2}\|\boldsymbol{x}\|^2 - \frac{1}{2} + (1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^* \boldsymbol{z}_t).$$

Hence, we will conclude that

$$\begin{aligned}
\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_{t+1}\right\| &\leq \left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \tilde{\boldsymbol{x}}\right\| + \|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\| \\
&\leq \delta_r + 2\mathrm{Re}\left(\boldsymbol{v}_t^*(\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1})\right) \\
&\leq \delta_r + 2\|\boldsymbol{v}_t\|\left\|\tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + 2(1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^*\boldsymbol{z}_t)\|\boldsymbol{v}_t\|\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| \\
&\quad + \frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}_t^*\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right) \\
&\leq \delta_r + 2\delta_r + \left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2 + |\|\boldsymbol{x}\|^2 - 1| \\
&\quad + \frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}_t^*\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right) \\
&\leq 3\delta_r + \gamma + \left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2 \\
&\quad + \frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}_t^*\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right).
\end{aligned}$$

The rest of the proof is exactly the same as the proof of Theorem 9, and is hence skipped. □

We would like to emphasize that given the linear convergence of the GD-COPER, the accumulation of the error due to γ will be negligible and the error in the estimation of the magnitude of $\|\boldsymbol{x}\|^2$ does not have any major impact on the performance of GD-COPER.

We now turn our attention to the second assumption, i.e. the assumption that $\mathcal{P}_{\mathcal{C}_r}(\cdot) = \mathcal{D}_r(\mathcal{E}_r(\cdot))$. We would like to first emphasize that ideally, this is what a compression algorithm should do. If an image compression algorithm maps an image to a codeword that is far from the original image, that is an indication of the fact that the compression algorithm is not good. However, it is also reasonable to consider situations in which multiple codewords are close to an image and the compression algorithm does not pick the one which is the closest to the image because of some non-ideal strategies that is chosen to reduce the computational complexity. Hence, again we can ask whether the GD-COPER algorithm is robust to such non-ideal compression algorithms? In the rest of this section, we pursue the following two goals:

1. Provide a few examples to convince the readers that most of the standard compression

algorithms try to mimic a projection onto the codewords.

2. Suppose that even though the compression algorithm is non-ideal and does not find the closest codeword, it is still capable of finding a codeword that is in the vicinity of the closest codeword. We aim to show that the performance of GD-COPER algorithm is robust to such non-idealities.

Let us start with an example that is the cornerstone of several important compression algorithms. Suppose $Q \subset [0,1]^n$ is the set of approximately sparse signals. For instance, for some $p < 1$

$$Q = \left\{ \boldsymbol{x} \in [0,1)^n,\ \|\boldsymbol{x}\|_p \leq \zeta \right\}.$$

The main idea of many compression algorithms is to approximate the signals in Q with k-sparse signals and encode the k-sparse signal. For simplicity suppose that we are given the k. Let $r_t = k\lceil \log_2 n \rceil + k(t+1)$ denote the rate of our compression algorithm. Let $\mathcal{E}_1 : Q \rightarrow \{0,1\}^{k\lceil \log_2 n \rceil}$ encode the location of k largest elements of $\boldsymbol{x}$. Furthermore, to code the magnitudes of the non-zero coefficients we consider $\mathcal{E}_2 : Q \rightarrow \{0,1\}^{k(t+1)}$ that consider the k largest components of $\boldsymbol{x}$ and codes each of them with $t+1$ bits (does a binary expansion and keeps the $t+1$ most significant bits).

More precisely, if $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n) \in Q$, where $1 \leq i_1 < i_2 < ... < i_k \leq n$ are the location of its k-largest elements, then

$$\mathcal{E}_1(\boldsymbol{x}) = \left(B(i_1), ..., B(i_k)\right), \tag{5.48}$$

where $B(i)$ denotes the binary expansion of positive integer i. Note that since indices are less than or equal to n, $\log_2 n$ bits are enough to code each of them. Moreover, if $x_i = \sum_{j=1}^{\infty} \epsilon_{i,j} 2^{-j}$ with $\epsilon_{i,j} \in \{0,1\}$, denote the binary expansion of x_i, then

$$\mathcal{E}_2(\boldsymbol{x}) = \left((\epsilon_{i_1,1}, \ldots, \epsilon_{i_1,t+1}), ..., (\epsilon_{i_k,1}, \ldots, \epsilon_{i_k,t+1})\right). \tag{5.49}$$

Note that this type of coding is very close to what happens in e.g. JPEG and embedded zero tree wavelet (EZW) compression algorithms. Both compression algorithms first transform the image to a domain that is more compressible, e.g. Fourier and wavelet, and then code the location and magnitudes of the largest coefficients similar to what we did above.[3] The decoder of the compression algorithms has access to the locations of the largest coefficients from the $k \log_2 n$ bits that it received from the encoder. Hence, it can easily use $(\epsilon_{i_1,1}, \epsilon_{i_1,2}, \dots, \epsilon_{i_1,t+1})$, ..., $(\epsilon_{i_k,1}, \epsilon_{i_k,2}, \dots, \epsilon_{i_k,t+1})$ to find the magnitudes of the signals at those locations. Define $\Gamma_k = \{\boldsymbol{x} \in [0,1]^n \ : \ \|\boldsymbol{x}\|_0 \leq k\}$. One can easily confirm that

$$\begin{aligned} C_{r_t} &= \mathcal{D}_{r_t}\left(\mathcal{E}_{r_t}(Q)\right) \\ &= \left\{ \boldsymbol{y} \in \Gamma_k, \ y_i = \sum_{j=1}^{t} \epsilon_{i,j} 2^{-j}, \quad \epsilon_{i,j} \in \{0,1\} \right\}. \end{aligned}$$

It is straightforward to show that in these types of compression algorithms $\mathcal{P}_{C_r}(\cdot) = \mathcal{D}_r \circ \mathcal{E}_r(\cdot)$. For the sake of completeness we include a brief proof below. Suppose that the choice of codeword for the projection is unique. For notational simplicity we drop the subscript t. To prove this formally, let $\boldsymbol{x} \in [0,1)^n$ be an arbitrary vector and let $\boldsymbol{y} = \mathcal{D}_r(\mathcal{E}_r(\boldsymbol{x}))$ and $\boldsymbol{z} = \mathcal{P}_{C_r}(\boldsymbol{x})$. We have to show $\boldsymbol{y} = \boldsymbol{z}$. Since $\boldsymbol{z} \in C_r$, it has at most k non-zero coordinates. Firstly, we claim location of these non-zero coordinates have to match with the largest coordinates of $\boldsymbol{x}$. If this does not hold, one can swap two coordinates of $\boldsymbol{z}$ and get smaller distance to $\boldsymbol{x}$ by noting that if $x_i < x_j$ and $z_i > 0, z_j = 0$ then

$$(x_i - z_i)^2 + x_j^2 < \left(x_j - z_i\right)^2 + x_i^2,$$

which contradicts with $\boldsymbol{z}$ being the projection of $\boldsymbol{x}$. Furthermore, if $y_i = \sum_{j=1}^{t} \epsilon_{i,j} 2^{-j}$ and $z_i = \sum_{j=1}^{t} \tilde{\epsilon}_{i,j} 2^{-j}$ then$\left|x_i - y_i\right| \leq 2^{-t-1}$ and$|x_i - z_i| \leq 2^{-t-1}$ implies$\left|y_i - z_i\right| \leq 2^{-t}$ which yields$\left|x_i - y_i\right| = |x_i - z_i|$. Note that there can be a case where $y_i \neq z_i$ while they have the same distance from x_i. As an example, consider $t = 0, \ x_i = 0.5, \ y_i = 0, \ \boldsymbol{z}_i = 1$. This yields for every i that$\left|x_i - y_i\right| \leq |x_i - z_i|$.

[3] There are some minor tweaks in the actual JPEG and EZW. Since coding the locations of the largest coefficients requires a large number of bits, they often use techniques such as counting zero runs or coding along the trees to reduce the number of bits.

Hence

$$\left\|\boldsymbol{x}-\boldsymbol{y}\right\| \le \|\boldsymbol{x}-\boldsymbol{z}\|,$$

which means $\boldsymbol{y} = \mathcal{D}(\mathcal{E}(\boldsymbol{x}))$ is also a projection on C_r.

Now, let us turn to another point we would like to make, that is, even if the compression algorithm is not an accurate projection, GD-COPER can still perform an accurate recovery. Towards this goal we assume that the operation of $\mathcal{D}(\mathcal{E}(\boldsymbol{x}))$ is not a projection, but has some error. In other words, we assume that

$$\|\mathcal{D}(\mathcal{E}(\boldsymbol{x})) - \mathcal{P}_{C_r}(\boldsymbol{x})\| \le \gamma.$$

Our next theorem proves that if γ is not too large, then GD-COPER given by the following iteration can still perform well:

$$\begin{aligned} \boldsymbol{s}_{t+1} &= \boldsymbol{z}_t - \mu \nabla d_A(\boldsymbol{z}_t), \\ \boldsymbol{z}_{t+1} &= \mathcal{D}_r(\mathcal{E}_r(\boldsymbol{s}_{t+1})). \end{aligned}$$

Theorem 11. *For a fixed signal $\boldsymbol{x} \in \mathcal{Q}$, define $\boldsymbol{z}_t \in C_r$ as in (5.21) with $\mu = \frac{1}{8m}$. Suppose that for all $\theta \in \mathbb{R}$, $\mathrm{e}^{i\theta}\boldsymbol{x} \in \mathcal{Q}$. Define $\theta_t \triangleq \arg\min_{\theta\in\mathbb{R}} \left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta}\boldsymbol{x}\right\|$. For all $\epsilon \ge C_2 m^{-\frac{1}{3}}$, with probability at least $1 - C_3 \mathrm{e}^{-C_1\sqrt{m}\epsilon + (3\log 2)r}$, where $C_1, C_2, C_3 > 0$ are absolute constants, for $t = 1, 2, \ldots$, we have*

$$\begin{aligned} \left\|\boldsymbol{z}_{t+1} - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| \le &\left(\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + 2\epsilon\right)\left\|\boldsymbol{z}_t - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| + 3\delta_r \\ &+ 2\gamma + \sqrt{2\gamma(\delta_r + 1 + \epsilon)}. \end{aligned} \tag{5.50}$$

Before we prove this theorem, let us interpret it. Everything in the theorem is similar to what we had in Theorem 9. The only difference, is the term $2\gamma + \sqrt{2\gamma(\delta_r + 1 + \epsilon)}$ added to the error. Again given the geometric convergence of the algorithm the total error after T iterations does not accumulate much and remains at the same order. It is clear that if γ is small, then the GD-COPER algorithm performs well.

Proof of Theorem 11. Since the proof is very similar to the proof of Theorem 9 we do not repeat the entire proof and only emphasize on the aspects of this proof that change. Let $\tilde{\boldsymbol{x}} = \mathcal{P}_{C_r}(\mathrm{e}^{i\theta_t}\boldsymbol{x})$, and define $\boldsymbol{w}_{t+1} = \mathcal{P}_{C_r}(\boldsymbol{s}_{t+1})$. In this case, we know that $\boldsymbol{z}_{t+1} = \mathcal{D}(\mathcal{E}(\boldsymbol{s}_{t+1}))$ and we have

$$\|\boldsymbol{w}_{t+1} - \boldsymbol{z}_{t+1}\| \leq \gamma. \tag{5.51}$$

Since $\tilde{\boldsymbol{x}} \in C_r$, we have

$$\begin{aligned}
\|\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}}\|^2 &\geq \|\boldsymbol{s}_{t+1} - \boldsymbol{w}_{t+1}\|^2 \\
&= \|\boldsymbol{s}_{t+1} - \boldsymbol{z}_{t+1}\|^2 + \|\boldsymbol{z}_{t+1} - \boldsymbol{w}_{t+1}\|^2 \\
&\quad + 2\mathrm{Re}\left((\boldsymbol{z}_{t+1} - \boldsymbol{w}_{t+1})^*(\boldsymbol{s}_{t+1} - \boldsymbol{z}_{t+1})\right) \\
&\geq \|\boldsymbol{s}_{t+1} - \boldsymbol{z}_{t+1}\|^2 \\
&\quad + 2\mathrm{Re}\left((\boldsymbol{z}_{t+1} - \boldsymbol{w}_{t+1})^*(\boldsymbol{s}_{t+1} - \boldsymbol{z}_{t+1})\right),
\end{aligned}$$

where to obtain the last inequality we used the Cauchy-Schwartz inequality, (5.51), and the fact that both $\boldsymbol{s}_{t+1}$ and $\boldsymbol{z}_{t+1}$ have unit norms. Therefore, we have

$$\begin{aligned}
\|\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}}\|^2 &\geq \|\boldsymbol{s}_{t+1} - \boldsymbol{z}_{t+1}\|^2 \\
&\quad + 2\mathrm{Re}\left((\boldsymbol{z}_{t+1} - \boldsymbol{w}_{t+1})^*(\boldsymbol{s}_{t+1} - \boldsymbol{z}_{t+1})\right) \\
&= \|\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}}\|^2 + \|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|^2 \\
&\quad + 2\mathrm{Re}\left((\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})^*(\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}})\right) \\
&\quad + 2\mathrm{Re}\left((\boldsymbol{z}_{t+1} - \boldsymbol{w}_{t+1})^*(\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}})\right) \\
&\quad + 2\mathrm{Re}\left((\boldsymbol{z}_{t+1} - \boldsymbol{w}_{t+1})^*(\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})\right).
\end{aligned}$$

Hence,

$$\begin{aligned}\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|^2 \leq\ & 2\mathrm{Re}\left((\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})^*(\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1})\right) \\ & + 2\mathrm{Re}\left((\boldsymbol{w}_{t+1} - \boldsymbol{z}_{t+1})^*(\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}})\right) \\ & + 2\mathrm{Re}\left((\boldsymbol{w}_{t+1} - \boldsymbol{z}_{t+1})^*(\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})\right) \end{aligned} \tag{5.52}$$

Recall that $\mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right] = 8m(\boldsymbol{z}\boldsymbol{z}^* - \boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{z}$. Thus,

$$\begin{aligned}\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1} &= \tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x} + \mathrm{e}^{i\theta_t}\boldsymbol{x} - \Bigg(\boldsymbol{z}_t - \frac{1}{8m}\mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right] \\ &\quad + \frac{1}{8m}\Big(\mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right] - \nabla d_A(\boldsymbol{z}_t)\Big)\Bigg) \\ &= \tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x} + \mathrm{e}^{i\theta_t}\boldsymbol{x} - \left(\boldsymbol{z}_t - \boldsymbol{z}_t + (\boldsymbol{x}^*\boldsymbol{z}_t)\boldsymbol{x}\right) \\ &\quad + \frac{1}{8m}\Big(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\Big) \\ &= \tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x} + (1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^*\boldsymbol{z}_t)\mathrm{e}^{i\theta_t}\boldsymbol{x} \\ &\quad + \frac{1}{8m}\Big(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\Big). \end{aligned} \tag{5.53}$$

Note that $\left\|\tilde{\boldsymbol{x}} - \mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| \leq \delta_r$. Also, since $1 - (\mathrm{e}^{i\theta_t}\boldsymbol{x})^*\boldsymbol{z}_t = \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2$ and $\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x}\right\| = \|\boldsymbol{v}_t\| = 1$. Let

$$\boldsymbol{v}_t \triangleq \frac{\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}}{\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|}. \tag{5.54}$$

and

$$\tilde{\boldsymbol{v}}_t \triangleq \frac{\boldsymbol{w}_{t+1} - \boldsymbol{z}_{t+1}}{\|\boldsymbol{w}_{t+1} - \boldsymbol{z}_{t+1}\|}. \tag{5.55}$$

Then we have

$$
\begin{aligned}
\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|^2 \leq{}& 2\mathrm{Re}\left((\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})^*(\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1})\right) \\
&+ 2\mathrm{Re}\left((\boldsymbol{w}_{t+1} - \boldsymbol{z}_{t+1})^*(\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}})\right) \\
&+ 2\mathrm{Re}\left((\boldsymbol{w}_{t+1} - \boldsymbol{z}_{t+1})^*(\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})\right) \\
\leq{}& 2\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\||\mathrm{Re}\left(\boldsymbol{v}_t^*(\tilde{\boldsymbol{x}} - \boldsymbol{s}_{t+1})\right)| \\
&+ 2\gamma|\mathrm{Re}\left((\tilde{\boldsymbol{v}}_{t+1})^*(\boldsymbol{s}_{t+1} - \tilde{\boldsymbol{x}})\right)| \\
&+ 2\gamma|\mathrm{Re}\left((\tilde{\boldsymbol{v}}_{t+1})^*(\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1})\right)| \\
\leq{}& 2\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\| \left(\delta_r + \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2\right. \\
&+ \left.\left|\mathrm{Re}\left(\boldsymbol{v}_t^*(\frac{1}{8m}\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right)\right|\right) \\
&+ 2\gamma\left(\delta_r + \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2\right. \\
&+ \left.\left|\mathrm{Re}\left(\boldsymbol{v}_t^*(\frac{1}{8m}\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right)\right|\right) \\
&+ 2\gamma\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|.
\end{aligned}
\tag{5.56}
$$

Define events $\mathcal{G}$ and $\tilde{\mathcal{G}}$ as follows

$$
\begin{aligned}
\mathcal{G} \triangleq \Bigg\{&\frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}^*\left(\nabla d_A(\boldsymbol{z}) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right]\right)\right) \\
&\leq \epsilon \inf_{\theta\in\mathbb{R}}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z}\right\|, \quad \boldsymbol{v} = \frac{\tilde{\boldsymbol{x}} - \boldsymbol{z}'}{\|\tilde{\boldsymbol{x}} - \boldsymbol{z}'\|}, \quad \forall \boldsymbol{z}, \tilde{\boldsymbol{x}} \in C_r\Bigg\},
\end{aligned}
\tag{5.57}
$$

$$
\begin{aligned}
\tilde{\mathcal{G}} \triangleq \Bigg\{&\frac{1}{4m}\mathrm{Re}\left(\tilde{\boldsymbol{v}}^*\left(\nabla d_A(\boldsymbol{z}) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right]\right)\right) \\
&\leq \epsilon \inf_{\theta\in\mathbb{R}}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z}\right\|, \quad \tilde{\boldsymbol{v}} = \frac{\boldsymbol{z} - \boldsymbol{z}'}{\|\boldsymbol{z} - \boldsymbol{z}'\|}, \quad \forall \boldsymbol{z}, \boldsymbol{z}' \in C_r\Bigg\}.
\end{aligned}
\tag{5.58}
$$

Similar to the proof of Theorem 9, we can bound the probabilities of these events. In particular, we have constants $C_1, C_2, C_3 > 0$ such that for every $\epsilon \geq C_2 m^{-\frac{1}{3}}$,

$$\mathbb{P}\left(\left|\mathrm{Re}\left(\boldsymbol{v}^*\left(\nabla d_A(\boldsymbol{z}) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right]\right)\right)\right| \right.$$
$$\left. > 4m\epsilon \inf_{\theta \in \mathbb{R}} \left\| \mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z} \right\|\right) \leq C_3 \mathrm{e}^{-C_1\sqrt{m\epsilon}}. \tag{5.59}$$

Hence, combining (5.59) with the union bound, for every $\epsilon \geq C_2 m^{-\frac{1}{3}}$, we have

$$\mathbb{P}\left(\mathcal{G}\right) \geq 1 - 2^{3r} C_3 \mathrm{e}^{-C_1\sqrt{m\epsilon}}. \tag{5.60}$$

Similarly, we have

$$\mathbb{P}\left(\tilde{\mathcal{G}}\right) \geq 1 - 2^{3r} C_3 \mathrm{e}^{-C_1\sqrt{m\epsilon}}. \tag{5.61}$$

Therefore, conditioned on $\mathcal{G} \cap \tilde{\mathcal{G}}$ we have

$$\begin{aligned} &\frac{1}{4m}\mathrm{Re}\left(\boldsymbol{v}_t^*\left(\nabla d_A(\boldsymbol{z}_t) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z}_t)\right]\right)\right) \\ &\quad \leq \epsilon \inf_{\theta \in \mathbb{R}} \left\| \mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z}_t \right\| = \epsilon \left\| \mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t \right\|. \end{aligned}$$

Hence, (5.56) implies that, for all $t \in \{1 \cdots, T\}$,

$$\begin{aligned} &\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|^2 \\ &\quad \leq 2\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\| \left(\delta_r + \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2 + \epsilon\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|\right) \\ &\qquad + 2\gamma\left(\delta_r + \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2 + \epsilon\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|\right) \\ &\qquad + 2\gamma\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\| \\ &\quad \leq 2\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\| \left(\delta_r + \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2 \right. \\ &\qquad \left. + \epsilon\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\| + \gamma\right) + 2\gamma(\delta_r + 1 + \epsilon). \end{aligned}$$

Given that we have a quadratic function of $\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|$ with one negative and one positive root, it is straightforward to see that $\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\|$ should be smaller than the positive root. By bounding the positive root we obtain

$$\begin{aligned}
\|\tilde{\boldsymbol{x}} - \boldsymbol{z}_{t+1}\| \le 2\left(\delta_r + \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2 + \epsilon\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\| + \gamma\right) \\
+ \sqrt{2\gamma(\delta_r + 1 + \epsilon)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_{t+1}\right\| &\le \delta_r + 2\left(\delta_r + \frac{1}{2}\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|^2\right. \\
&\qquad \left. + \epsilon\left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\| + \gamma\right) + \sqrt{2\gamma(\delta_r + 1 + \epsilon)} \\
&\le 3\delta_r + 2\gamma + \sqrt{2\gamma(\delta_r + 1 + \epsilon)} \\
&\qquad + \left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|\left(2\epsilon + \left\|\mathrm{e}^{i\theta_t}\boldsymbol{x} - \boldsymbol{z}_t\right\|\right). \qquad (5.62)
\end{aligned}$$

□

5.6 Proofs

5.6.1 Preliminaries

Lemma 46. $\inf_{\theta\in[0,2\pi)}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{y}\right\|$ *achieves its minimum at a value of* θ *that makes* $\mathrm{e}^{-i\theta}\boldsymbol{x}^*\boldsymbol{y}$ *a positive real number, and for that* θ *we have*

$$\begin{aligned}
\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{y}\right\|^2 &= \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\left|\boldsymbol{x}^*\boldsymbol{y}\right| \\
&= \left(\|\boldsymbol{x}\| - \|\boldsymbol{y}\|\right)^2 + 2\left(\|\boldsymbol{x}\|\|\boldsymbol{y}\| - \left|\boldsymbol{x}^*\boldsymbol{y}\right|\right).
\end{aligned}$$

Proof. Let $\boldsymbol{z} = \mathrm{e}^{i\theta}\boldsymbol{x}$

$$\begin{aligned}\left\|\boldsymbol{z}-\boldsymbol{y}\right\|^2 &= \left(\boldsymbol{z}-\boldsymbol{y}\right)^*\left(\boldsymbol{z}-\boldsymbol{y}\right)\\ &= \|\boldsymbol{z}\|^2+\left\|\boldsymbol{y}\right\|^2-2\mathrm{Re}(\boldsymbol{z}^*\boldsymbol{y})\\ &\geq \|\boldsymbol{z}\|^2+\left\|\boldsymbol{y}\right\|^2-2\left|\boldsymbol{z}^*\boldsymbol{y}\right|\\ &= \|\boldsymbol{x}\|^2+\left\|\boldsymbol{y}\right\|^2-2\left|\boldsymbol{x}^*\boldsymbol{y}\right|.\end{aligned}$$

Note that equality holds only when $\mathrm{Re}(\boldsymbol{z}^*\boldsymbol{y}) = \left|\boldsymbol{z}^*\boldsymbol{y}\right|$, which proves our claim. □

Lemma 47. *For any two vectors* $\boldsymbol{x}$ *and* $\hat{\boldsymbol{x}}$ *in* $\mathbb{C}^n$*, we have*

$$\begin{aligned}&\frac{1}{8}\left(\inf_{\theta}\left\|\mathrm{e}^{i\theta}\boldsymbol{x}-\hat{\boldsymbol{x}}\right\|^2\right)^2\\ &\qquad\leq \frac{1}{2}\left(\|\boldsymbol{x}\|^2-\|\hat{\boldsymbol{x}}\|^2\right)^2+\left(\|\boldsymbol{x}\|^2\|\hat{\boldsymbol{x}}\|^2-\left|\boldsymbol{x}^*\hat{\boldsymbol{x}}\right|^2\right).\end{aligned}$$

Proof. Note that according to Lemma 46 we have

$$\begin{aligned}&\left(\left(\|\boldsymbol{x}\|-\|\hat{\boldsymbol{x}}\|\right)^2+2\left(\|\boldsymbol{x}\|\|\hat{\boldsymbol{x}}\|-\left|\boldsymbol{x}^*\hat{\boldsymbol{x}}\right|\right)\right)^2\\ &\qquad\leq (1+1)\left(\left(\|\boldsymbol{x}\|-\|\hat{\boldsymbol{x}}\|\right)^4+4\left(\|\boldsymbol{x}\|\|\hat{\boldsymbol{x}}\|-\left|\boldsymbol{x}^*\hat{\boldsymbol{x}}\right|\right)^2\right)\\ &\qquad\leq 2\left(\|\boldsymbol{x}\|^2-\|\hat{\boldsymbol{x}}\|^2\right)^2+8\left(\|\boldsymbol{x}\|^2\|\hat{\boldsymbol{x}}\|^2-\left|\boldsymbol{x}^*\hat{\boldsymbol{x}}\right|^2\right).\end{aligned}$$

Hence,

$$\begin{aligned}&\frac{1}{8}\left(\inf_{\theta}\left\|\mathrm{e}^{i\theta}\boldsymbol{x}-\hat{\boldsymbol{x}}\right\|^2\right)^2\\ &\qquad\leq \frac{1}{4}\left(\|\boldsymbol{x}\|^2-\|\hat{\boldsymbol{x}}\|^2\right)^2+\left(\|\boldsymbol{x}\|^2\|\hat{\boldsymbol{x}}\|^2-\left|\boldsymbol{x}^*\hat{\boldsymbol{x}}\right|^2\right)\\ &\qquad\leq \frac{1}{2}\left(\|\boldsymbol{x}\|^2-\|\hat{\boldsymbol{x}}\|^2\right)^2+\left(\|\boldsymbol{x}\|^2\|\hat{\boldsymbol{x}}\|^2-\left|\boldsymbol{x}^*\hat{\boldsymbol{x}}\right|^2\right).\end{aligned}$$

□

Lemma 48. *Let $\Phi(x)$ denote the CDF of a standard normal variable. Then, for any $u > 0$,*

$$g(u) = \mathrm{e}^{\frac{1}{u^2}} \Phi\left(-\frac{\sqrt{2}}{u}\right) \leq 1.$$

Proof. With a change of variable $v = \frac{\sqrt{2}}{u}$, proving $g(u) \leq 1$ is equivalent to proving $h(v) = \mathrm{e}^{-\frac{v^2}{2}} - \Phi(-v) \geq 0$ for all $v \geq 0$. We have

$$h'(v) = \left(-v + \frac{1}{\sqrt{2\pi}}\right) \mathrm{e}^{-\frac{v^2}{2}} \implies \begin{cases} h'(v) \geq 0 & v \leq \dfrac{1}{\sqrt{2\pi}} \\ h'(v) < 0 & v > \dfrac{1}{\sqrt{2\pi}} \end{cases}.$$

In addition, $h(0) = \frac{1}{2} > 0$ and $h(\infty) = 0$. □

Lemma 49 (Chi squared concentration)**.** *For any $\tau \geq 0$, we have*

$$\mathbb{P}\left(\chi^2(m) > m(1+\tau)\right) \leq \mathrm{e}^{-\frac{m}{2}(\tau - \log(1+\tau))}.$$

The proof of this lemma can be found in [138].

5.6.2 Heavy-tailed concentration

In this section, we discuss a few lemmas regarding the concentration of heavy-tailed random variables. A more complete discussion of such concentration results can be found in [44].

Lemma 50 (Bounded random variable MGF upper bound)**.** *Let X be a random variable and c, c' positive constants such that*

$$\mathbb{P}\left(\left|X - \mathbb{E}[X]\right| > \tau\right) \leq c' \exp\left(-c\sqrt{\tau}\right) \quad \forall \tau \geq 0.$$

Then there exist constants c_2, c_3, depending only on the distribution of X, such that for all $L \geq c_3$

and $\lambda = \frac{c}{2\sqrt{L}}$

$$\log_2 \mathbb{E}\left[\exp\left(\lambda\left(X\mathbf{1}_{X\leq L} - \mathbb{E}[X]\right)\right)\right] \leq \frac{c_2}{2}\lambda^2.$$

Proof. For simplicity of the notation let $X_L \triangleq X\mathbf{1}_{X\leq L}$ denotes the truncated version of the X. By Taylor expansion of exponential function at $\mathbb{E}[X]$, one can get

$$\begin{aligned}\exp(\lambda X_L) = \exp\left(\lambda\mathbb{E}[X]\right) + \lambda\left(X_L - \mathbb{E}[X]\right)\exp\left(\lambda\mathbb{E}[X]\right)\\ +\frac{\lambda^2}{2}\left(X_L - \mathbb{E}[X]\right)^2\exp(\lambda Y),\end{aligned}$$

where Y is a random variable whose value is between $\mathbb{E}[X]$ and X_L. Therefore

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\left(X_L - \mathbb{E}[X]\right)\right)\right] &= 1 + \lambda\left(\mathbb{E}[X_L] - \mathbb{E}[X]\right)\\ &+\frac{\lambda^2}{2}\mathbb{E}\left[\left(X_L - \mathbb{E}[X]\right)^2\exp\left(\lambda\left(Y - \mathbb{E}[X]\right)\right)\right].\end{aligned} \tag{5.63}$$

Note that $\mathbb{E}[X_L - X] \leq 0$ and $\log_2(1+x) \leq x \quad \forall x \geq 0$, hence by (5.63) we obtain

$$\begin{aligned}&\log_2 \mathbb{E}\left[\exp\left(\lambda\left(X_L - \mathbb{E}[X]\right)\right)\right]\\ &\qquad\leq \frac{\lambda^2}{2}\mathbb{E}\left[\left(X_L - \mathbb{E}[X]\right)^2\exp\left(\lambda\left(Y - \mathbb{E}[X]\right)\right)\right],\end{aligned}$$

which means if for some c_3 we show

$$\sup_{L\geq c_3,\ \lambda=\frac{c}{2\sqrt{L}}} \mathbb{E}\left[\left(X_L - \mathbb{E}[X]\right)^2\exp\left(\lambda\left(Y - \mathbb{E}[X]\right)\right)\right] \leq c_2, \tag{5.64}$$

the Lemma holds with c_2, c_3.

Note that since Y is bounded between $\mathbb{E}[X]$ and X_L we get

$$\mathbb{E}\left[(X_L - \mathbb{E}[X])^2 \exp\left(\lambda (Y - \mathbb{E}[X])\right)\right] \leq \mathbb{E}\left[X_L - \mathbb{E}[X]\right]^2 + \mathbb{E}\left[(X_L - \mathbb{E}[X])^2 \exp\left(\lambda (X_L - \mathbb{E}[X])\right)\right]. \tag{5.65}$$

Since $X_L \xrightarrow{L\to\infty} X$ and it is dominated by X, by the dominated convergence Theorem we get $\mathbb{E}\left[X_L - \mathbb{E}[X]\right]^2 \to Var(X)$, hence for $L > c_3'$ we have

$$\mathbb{E}\left[X_L - \mathbb{E}[X]\right]^2 \leq 2Var(X). \tag{5.66}$$

Now, we bound the second term in (5.65).

$$\begin{aligned}&\mathbb{E}\left[(X_L - \mathbb{E}[X])^2 \exp\left(\lambda (X_L - \mathbb{E}[X])\right)\right]\\ &= \mathbb{E}\left[(X_L - \mathbb{E}[X])^2 \exp\left(\lambda (X_L - \mathbb{E}[X])\right) \mathbf{1}_{X_L < \mathbb{E}[X]}\right]\\ &\quad + \mathbb{E}\left[(X_L - \mathbb{E}[X])^2 \exp\left(\lambda (X_L - \mathbb{E}[X])\right) \mathbf{1}_{X_L \geq \mathbb{E}[X]}\right].\end{aligned}$$

Note that

$$\mathbb{E}\left[(X_L - \mathbb{E}[X])^2 \exp\left(\lambda (X_L - \mathbb{E}[X])\right) \mathbf{1}_{X_L < \mathbb{E}[X]}\right] \leq \mathbb{E}\left[X_L - \mathbb{E}[X]\right]^2 \leq 2Var(X), \tag{5.67}$$

for $L > c_3'$.

Moreover, if $U \triangleq X_L - \mathbb{E}[X]$

$$
\begin{aligned}
&\mathbb{E}\left[U^2 \exp\left(\lambda U\right) \mathbf{1}_{U\geq 0}\right] \\
&\quad = \int_0^\infty \mathbb{P}\left(U^2 \exp\left(\lambda U\right) > u,\ U \geq 0\right) du \\
&\quad = \int_0^\infty \mathbb{P}\left(X_L - \mathbb{E}\left[X\right] > t\right) du, \qquad t^2 \exp\left(\lambda t\right) = u, \\
&\quad = \int_0^{L-\mathbb{E}[X]} \mathbb{P}\left(X - \mathbb{E}\left[X\right] > t\right) \exp\left(\lambda t\right) \left(2t + \lambda t^2\right) dt \\
&\quad \leq \int_0^{L-\mathbb{E}[X]} c' \exp\left(-c\sqrt{t} + \frac{c}{2\sqrt{L}} t\right) \left(2t + \lambda t^2\right) dt. \qquad (5.68)
\end{aligned}
$$

Note that for large enough L and $0 \leq t \leq L - \mathbb{E}\left[X\right]$, we have $-c\sqrt{t} + \frac{c}{2\sqrt{L}} t \leq -\frac{c\sqrt{t}}{3}$. More specifically,

$$
-\sqrt{t} + \frac{t}{2\sqrt{L}} \leq -\frac{\sqrt{t}}{3},\ \forall L \geq \frac{-9\mathbb{E}\left[X\right]}{7},\ 0 \leq t \leq L - \mathbb{E}\left[X\right],
$$

hence by (5.68) we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left(X_L - \mathbb{E}\left[X\right]\right)^2 \exp\left(\lambda\left(X_L - \mathbb{E}\left[X\right]\right)\right) \mathbf{1}_{X_L \geq \mathbb{E}[X]}\right] \\
&\quad \leq c' \int_0^{L-\mathbb{E}[X]} \exp\left(-\frac{c\sqrt{t}}{3}\right) (2t + \lambda t^2) dt \\
&\quad \leq c' \int_0^\infty \exp\left(-\frac{c\sqrt{t}}{3}\right) (2t + \lambda t^2) dt \\
&\quad = c' \int_0^\infty \exp(-z) \left(\frac{18}{c^2} z^2 + \frac{81\lambda}{c^4} z^4\right) \frac{18z}{c^2} dz \\
&\quad = \frac{18^2 c'}{c^4} \Gamma(4) + \frac{18c' \times 81\lambda}{c^6} \Gamma(6) \\
&\quad = \frac{18^2 c'}{c^4} \Gamma(4) + \frac{9c' \times 81}{c^5\sqrt{L}} \Gamma(6) \\
&\quad \leq \frac{324c'}{c^4} \Gamma(4) + \frac{729c'}{c^5\sqrt{c_3}} \Gamma(6), \qquad \forall L \geq c_3. \qquad (5.69)
\end{aligned}
$$

Hence. if we set $c_2 = 4Var(X) + \frac{324c'}{c^4}\Gamma(4) + \frac{729c'}{c^5\sqrt{c_3}}\Gamma(6)$ and $c_3 \geq c_3'$, (5.66), (5.67), (5.69) yield

(5.64) which concludes the proof. □

Lemma 51 (Heavy tail concentration)**.** *Let $\{Y_k\}_{k\in\mathbb{N}}$ be i.i.d. random variables. Assume there are constants $c > 0,\ c' \geq 1$ such that $\mathbb{P}\left(\left|Y_k - \mathbb{E}\left[Y_k\right]\right| > \tau\right) \leq c' \mathrm{e}^{-c\sqrt{\tau}}$, for all $\tau > 0$. Then, there exist a positive constant $C_3 > 0$, such that, for every $\epsilon > C_3 m^{-\frac{1}{3}}$,*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{k=1}^{m} Y_k - \mathbb{E}\left[Y_k\right]\right| > \epsilon\right) \leq 4\mathrm{e}^{-\frac{c}{2}\sqrt{m\epsilon}}. \tag{5.70}$$

Proof. Following the notion in the proof of the Lemma 50, let $Y_k^L = Y_k \mathbf{1}_{Y_k \leq L}$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{m}\sum_{k=1}^{m} Y_k - \mathbb{E}\left[Y_k\right] > \epsilon\right) \\
&\quad\leq \mathbb{P}\left(\frac{1}{m}\sum_{k=1}^{m} Y_k^L - \mathbb{E}\left[Y_k\right] > \epsilon\right) + \mathbb{P}\left(\exists k,\ Y_k > L\right) \\
&\quad\leq \exp\left(-\lambda\epsilon\right)\mathbb{E}\left[\exp\left(\frac{\lambda}{m}(Y_1^L - \mathbb{E}\left[Y_1\right])\right)\right]^m \\
&\qquad + m\mathbb{P}\left(Y_1 - \mathbb{E}\left[Y_1\right] > L - \mathbb{E}\left[Y_1\right]\right) \\
&\quad\leq \exp\left(-\lambda\epsilon + \frac{c_2}{2}\frac{\lambda^2}{m^2}m\right) + mc'\exp\left(-c\sqrt{L - \mathbb{E}\left[Y_1\right]}\right),
\end{aligned} \tag{5.71}$$

for $\frac{\lambda}{m} = \frac{c}{2\sqrt{L}}$. Note that the last inequality is obtained by the Lemma 50. Set $L = m\epsilon$ and hence $\lambda = \frac{c\sqrt{m}}{2\sqrt{\epsilon}}$, then by (5.71) we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{m}\sum_{k=1}^{m} Y_k - \mathbb{E}\left[Y_k\right] > \epsilon\right) &\leq \exp\left(-c\sqrt{m\epsilon} + \frac{c_2 c^2}{8}\frac{1}{\epsilon}\right) \\
&\quad + mc'\exp\left(-c\left(\sqrt{m\epsilon} - \sqrt{\left|\mathbb{E}\left[Y_1\right]\right|}\right)\right).
\end{aligned} \tag{5.72}$$

Note that if $\epsilon \geq \left(\frac{c_2 c}{4}\right)^{\frac{2}{3}} m^{-\frac{1}{3}}$, we have $\frac{c_2 c^2}{8}\frac{1}{\epsilon} \leq \frac{c}{2}\sqrt{m\epsilon}$, hence,

$$\exp\left(-c\sqrt{m\epsilon} + \frac{c_2 c^2}{8}\frac{1}{\epsilon}\right) \leq \exp\left(-\frac{c}{2}\sqrt{m\epsilon}\right). \tag{5.73}$$

Furthermore,

$$\begin{aligned} &mc' \exp\left(-c\left(\sqrt{m\epsilon} - \sqrt{\left|\mathbb{E}\left[Y_1\right]\right|}\right)\right) \\ &\quad = \exp\left(\log_2 c' + \log_2 m + c\sqrt{\left|\mathbb{E}\left[Y_1\right]\right|} - c\sqrt{m\epsilon}\right) \\ &\quad \leq \exp\left(-\frac{c}{2}\sqrt{m\epsilon}\right), \end{aligned} \tag{5.74}$$

whenever

$$\log_2 c' + \log_2 m + c\sqrt{\left|\mathbb{E}\left[Y_1\right]\right|} \leq \frac{c}{2}\sqrt{m\epsilon}. \tag{5.75}$$

since $m\epsilon \geq C_3 m^{\frac{2}{3}}$ for $\epsilon \geq C_3 m^{-\frac{1}{3}}$ we have

$$\frac{c}{2}\sqrt{m\epsilon} \geq \frac{c\sqrt{C_3}}{2} m^{\frac{1}{3}}. \tag{5.76}$$

Given $m^{\frac{1}{3}}$ grows faster than $\log_2 m$, by choosing large enough C_3 we can make (5.73) and (5.74) hold for all integer m, thus we obtain

$$\mathbb{P}\left(\frac{1}{m}\sum_{k=1}^{m} Y_k - \mathbb{E}\left[Y_k\right] > \epsilon\right) \leq 2\exp\left(-\frac{c}{2}\sqrt{m\epsilon}\right). \tag{5.77}$$

By repeating the exact same line of the proof for $-Y_k$ instead of Y_k we can obtain

$$\mathbb{P}\left(\frac{1}{m}\sum_{k=1}^{m} Y_k - \mathbb{E}\left[Y_k\right] < -\epsilon\right) \leq 2\exp\left(-\frac{c}{2}\sqrt{m\epsilon}\right). \tag{5.78}$$

Combining (5.77) and (5.78) yields

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{k=1}^{m} Y_k - \mathbb{E}\left[Y_k\right]\right| > \epsilon\right) \leq 4\exp\left(-\frac{c}{2}\sqrt{m\epsilon}\right). \tag{5.79}$$

□

5.6.3 Properties of $d_A(\cdot,\cdot)$

Lemma 52. *If $\lambda_1(\boldsymbol{c})$ and $\lambda_2(\boldsymbol{c})$ denote the two non-zero eigenvalues of $\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*$, then we have*

1. $\lambda_1(\boldsymbol{c}) + \lambda_2(\boldsymbol{c}) = \|\boldsymbol{x}\|^2 - \|\boldsymbol{c}\|^2$.
2. $\lambda_1(\boldsymbol{c})^2 + \lambda_2(\boldsymbol{c})^2 = \left(\|\boldsymbol{x}\|^2 - \|\boldsymbol{c}\|^2\right)^2 + 2\left(\|\boldsymbol{x}\|^2\|\boldsymbol{c}\|^2 - |\boldsymbol{x}^*\boldsymbol{c}|^2\right)$.
3. $\lambda_1(\boldsymbol{c})\lambda_2(\boldsymbol{c}) = \left(|\boldsymbol{x}^*\boldsymbol{c}|^2 - \|\boldsymbol{x}\|^2\|\boldsymbol{c}\|^2\right) \leq 0$

Proof. First note that

$$\lambda_1(\boldsymbol{c}) + \lambda_2(\boldsymbol{c}) = \mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*) = \|\boldsymbol{x}\|^2 - \|\boldsymbol{c}\|^2. \tag{5.80}$$

Similarly,

$$\begin{aligned}\lambda_1(\boldsymbol{c})^2 + \lambda_2(\boldsymbol{c})^2 &= \mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*)^2 \\ &= \mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^*\boldsymbol{x}\boldsymbol{x}^*) + \mathrm{Tr}(\boldsymbol{c}\boldsymbol{c}^*\boldsymbol{c}\boldsymbol{c}^*) - \mathrm{Tr}(\boldsymbol{c}\boldsymbol{c}^*\boldsymbol{x}\boldsymbol{x}^*) \\ &\quad - \mathrm{Tr}(\boldsymbol{x}\boldsymbol{x}^*\boldsymbol{c}\boldsymbol{c}^*) \\ &= \|\boldsymbol{x}\|^4 + \|\boldsymbol{c}\|^4 - 2\left|\boldsymbol{x}^*\boldsymbol{c}\right|^2 \\ &= \left(\|\boldsymbol{x}\|^2 - \|\boldsymbol{c}\|^2\right)^2 + 2\left(\|\boldsymbol{x}\|^2\|\boldsymbol{c}\|^2 - \left|x^*c\right|^2\right). \end{aligned} \tag{5.81}$$

Finally,

$$\begin{aligned}
2\lambda_1(\boldsymbol{c})\lambda_2(\boldsymbol{c}) &= (\lambda_1(\boldsymbol{c}) + \lambda_2(\boldsymbol{c}))^2 - (\lambda_1(\boldsymbol{c})^2 + \lambda_2(\boldsymbol{c})^2) \\
&= 2\left(\left|\boldsymbol{x}^*\boldsymbol{c}\right|^2 - \|\boldsymbol{x}\|^2\|\boldsymbol{c}\|^2\right) \\
&\leq 0.
\end{aligned}$$

□

Lemma 53. *Let $Z = (\lambda_1(\boldsymbol{c})U + \lambda_2(\boldsymbol{c})V)^2$, where U and V are independent $\chi^2(2)$. Then, for any $\alpha > 0$, we have*

$$f(\alpha) \triangleq \mathbb{E}\left[\mathrm{e}^{-\alpha Z}\right] \leq \left(\frac{\pi}{\lambda_{\max}(\boldsymbol{c})^2\alpha}\right)^{\frac{1}{2}}.$$

Proof.

$$f(\alpha) = \int_{x,y\geq 0} \mathrm{e}^{-\alpha\left(\lambda_1(\boldsymbol{c})x+\lambda_2(\boldsymbol{c})y\right)^2}\frac{\mathrm{e}^{-\frac{x}{2}}}{2}\frac{\mathrm{e}^{-\frac{y}{2}}}{2}dxdy. \tag{5.82}$$

Consider changing the variable (x, y) in the above integral to (u, v) defined as

$$(u, v) = \left(\lambda_1(\boldsymbol{c})x + \lambda_2(\boldsymbol{c})y, \frac{x+y}{2}\right).$$

The determinent of the Jacobian of this mapping is given by

$$\left|\frac{\partial u, v}{\partial x, y}\right| = \begin{vmatrix} \lambda_1(\boldsymbol{c}) & \lambda_2(\boldsymbol{c}) \\ \frac{1}{2} & \frac{1}{2} \end{vmatrix} = \frac{\lambda_1(\boldsymbol{c}) - \lambda_2(\boldsymbol{c})}{2}.$$

Furthermore,

$$v - \frac{u}{2\lambda_2(\boldsymbol{c})} = \left(\frac{1}{2} + \frac{\lambda_1(\boldsymbol{c})}{-2\lambda_2(\boldsymbol{c})}\right)x.$$

Since $\frac{\lambda_1(\boldsymbol{c})}{-2\lambda_2(\boldsymbol{c})} > 0$, we have

$$x \geq 0 \iff v \geq \frac{u}{2\lambda_2(\boldsymbol{c})}.$$

Similarly,

$$v \geq \frac{u}{2\lambda_1(\boldsymbol{c})}.$$

Therefore,

$$\begin{aligned}
f(\alpha) &= \frac{2}{4(\lambda_1(\boldsymbol{c}) - \lambda_2(\boldsymbol{c}))} \int \int_{v \geq \frac{u}{2\lambda_1(\boldsymbol{c})}, v \geq \frac{u}{2\lambda_2(\boldsymbol{c})}} \mathrm{e}^{-\alpha u^2 - v} dv, u \\
&= \frac{1}{2(\lambda_1(\boldsymbol{c}) - \lambda_2(\boldsymbol{c}))} \int_{u \geq 0} \int_{v = \frac{u}{2\lambda_1(\boldsymbol{c})}}^{\infty} \mathrm{e}^{-\alpha u^2} \mathrm{e}^{-v} dv du + \frac{1}{2(\lambda_1(\boldsymbol{c}) - \lambda_2(\boldsymbol{c}))} \int_{u < 0} \int_{v = \frac{u}{2\lambda_2(\boldsymbol{c})}}^{\infty} \mathrm{e}^{-\alpha u^2} \mathrm{e}^{-v} dv du \\
&= \frac{1}{2(\lambda_1(\boldsymbol{c}) - \lambda_2(\boldsymbol{c}))} \int_{u=0}^{\infty} \mathrm{e}^{-\alpha u^2 - \frac{u}{2\lambda_1(\boldsymbol{c})}} du + \frac{1}{2(\lambda_1(\boldsymbol{c}) - \lambda_2(\boldsymbol{c}))} \int_{u=-\infty}^{0} \mathrm{e}^{-\alpha u^2 - \frac{u}{2\lambda_2(\boldsymbol{c})}} du \\
&= \frac{\mathrm{e}^{\frac{1}{16\lambda_1(\boldsymbol{c})^2 \alpha}}}{2(\lambda_1(\boldsymbol{c}) - \lambda_2(\boldsymbol{c}))} \int_{u=0}^{\infty} \mathrm{e}^{-\alpha(u + \frac{1}{4\lambda_1 \alpha})^2} du + \frac{\mathrm{e}^{\frac{1}{16\lambda_2(\boldsymbol{c})^2 \alpha}}}{2(\lambda_1(\boldsymbol{c}) - \lambda_2(\boldsymbol{c}))} \int_{u=-\infty}^{0} \mathrm{e}^{-\alpha(u + \frac{1}{4\lambda_2(\boldsymbol{c})\alpha})^2} du \\
&= \frac{\sqrt{\pi} \mathrm{e}^{\frac{1}{16\lambda_1(\boldsymbol{c})^2 \alpha}}}{2(|\lambda_1(\boldsymbol{c})| + |\lambda_2(\boldsymbol{c})|)\sqrt{\alpha}} \Phi\left(-\frac{\sqrt{2}}{4|\lambda_1(\boldsymbol{c})|\sqrt{\alpha}}\right) + \frac{\sqrt{\pi} \mathrm{e}^{\frac{1}{16\lambda_2(\boldsymbol{c})^2 \alpha}}}{2(|\lambda_1(\boldsymbol{c})| + |\lambda_2(\boldsymbol{c})|)\sqrt{\alpha}} \Phi\left(-\frac{\sqrt{2}}{4|\lambda_2(\boldsymbol{c})|\sqrt{\alpha}}\right),
\end{aligned} \tag{5.83}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \mathrm{e}^{-\frac{1}{2}u^2}$. According to Lemma 48 we have

$$\mathrm{e}^{\frac{1}{16\lambda_1(\boldsymbol{c})^2 \alpha}} \Phi\left(-\frac{\sqrt{2}}{4|\lambda_1(\boldsymbol{c})|\sqrt{\alpha}}\right) = g\left(4|\lambda_1(\boldsymbol{c})|\sqrt{\alpha}\right) \leq 1. \tag{5.84}$$

Hence, by combining (5.83), (5.84), and the fact that $\frac{1}{|\lambda_1(\boldsymbol{c})| + |\lambda_2(\boldsymbol{c})|} \leq \frac{1}{|\lambda_{\max}(\boldsymbol{c})|}$ we can complete the proof. □

Theorem below is showing how the distance function d_A concentrates when we have sufficient measurements.

Theorem 12 (Concentration of $d_A(\cdot, \cdot)$). *Let C_r denote the set of codewords at rate r, and $\boldsymbol{x}$ denotes the signal of interest. For a given $\boldsymbol{c} \in \mathbb{C}^n$, let $\lambda_{\min}^2(\boldsymbol{c}) \leq \lambda_{\max}^2(\boldsymbol{c})$ be squared of the two non-zero* *eigenvalues of* $\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*$. *For any positive real numbers* τ_1, τ_2,

$$\begin{aligned}\mathbb{P}\left(d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) > \lambda_{\max}^2(\boldsymbol{c})\tau_1, \ \forall \boldsymbol{c} \in C_r\right)& \\ \geq 1 - 2^r \mathrm{e}^{\frac{m}{2}\left(K + \log\tau_1 - \log m\right)}&, \end{aligned} \tag{5.85}$$

where $K = \log 2\pi\mathrm{e}$ *and*

$$\begin{aligned}\mathbb{P}\left(d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) < \lambda_{\max}^2(\boldsymbol{c})\left(4m(1+\tau_2)\right)^2\right)& \\ \geq 1 - \mathrm{e}^{-2m\left(\tau_2 - \log(1+\tau_2)\right)}&. \end{aligned} \tag{5.86}$$

Proof. Recall from (5.2) that

$$d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) = \sum_{k=1}^{m} \left(\boldsymbol{a}_k{}^*(\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*)\boldsymbol{a}_k\right)^2. \tag{5.87}$$

First, for fixed $\boldsymbol{x}$ and $\boldsymbol{c}$, we derive the distribution and the moment-generating function (mgf) of $d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|)$. Note that $\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*$ is a Hermitian matrix of rank at most two, and therefore it can be written as

$$\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^* = Q^T \begin{pmatrix} \lambda_1(\boldsymbol{c}) & & & \\ & \lambda_2(\boldsymbol{c}) & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \overline{Q}, \tag{5.88}$$

where $Q^T\overline{Q} = I_n$. Combining (5.87) and (5.88), we have

$$\begin{aligned}
&\sum_{k=1}^{m} \left(\boldsymbol{a}_k{}^*(\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{c}\boldsymbol{c}^*)\boldsymbol{a}_k\right)^2 \\
&= \sum_{k=1}^{m} \left(\boldsymbol{a}_k{}^* Q^T \begin{pmatrix} \lambda_1(\boldsymbol{c}) & & & \\ & \lambda_2(\boldsymbol{c}) & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \overline{Q}\boldsymbol{a}_k \right)^2 \\
&= \sum_{k=1}^{m} \left(\boldsymbol{B}_k^* \begin{pmatrix} \lambda_1(\boldsymbol{c}) & & & \\ & \lambda_2(\boldsymbol{c}) & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \boldsymbol{B}_k \right)^2 \\
&= \sum_{k=1}^{m} \left(\lambda_1(\boldsymbol{c}) \left|B_{k,1}\right|^2 + \lambda_2(\boldsymbol{c}) \left|B_{k,2}\right|^2 \right)^2,
\end{aligned}$$

where $\boldsymbol{B}_k = \overline{Q}\boldsymbol{a}_k$.

Since $\overline{Q}$ is an orthonormal matrix, $B = \overline{QA}$ has the same distribution as A, and therefore the χ^2 variables in the above sum are all independent. Let $Z_k = \left(\lambda_1(\boldsymbol{c})\left|B_{k,1}\right|^2 + \lambda_2(\boldsymbol{c})\left|B_{k,2}\right|^2\right)^2$. Then we have

$$d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) = \sum_{i=1}^{m} Z_i, \tag{5.89}$$

where $Z_1, \ldots, Z_m$ are i.i.d. as $(\lambda_1(\boldsymbol{c})U + \lambda_2(\boldsymbol{c})V)^2$, where U and V are independent $\chi^2(2)$ random variables. Define $\lambda_{\min}(\boldsymbol{c}), \lambda_{\max}(\boldsymbol{c})$ to denote $\lambda_1(\boldsymbol{c}), \lambda_2(\boldsymbol{c})$ with smaller and larger absolute value respectively, i.e.,

$$\left|\lambda_{\min}(\boldsymbol{c})\right| = \min\left\{\left|\lambda_1(\boldsymbol{c})\right|, \left|\lambda_2(\boldsymbol{c})\right|\right\},$$

$$\left|\lambda_{\max}(\boldsymbol{c})\right| = \max\left\{\left|\lambda_1(\boldsymbol{c})\right|, \left|\lambda_2(\boldsymbol{c})\right|\right\}.$$

To derive (5.85), note that according to Lemma 53 for any $\alpha > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) \leq t\right) &= \mathbb{P}\left(\mathrm{e}^{-\alpha \sum_{i=1}^{m} Z_i} \geq \mathrm{e}^{-\alpha t}\right) \\
&\leq \mathrm{e}^{\alpha t}\mathbb{E}\left[\mathrm{e}^{-\alpha Z_1}\right]^m \\
&\leq \mathrm{e}^{\alpha t} f(\alpha)^m \\
&\leq \mathrm{e}^{\alpha t}\left(\frac{\pi}{\lambda_{\max}(c)^2 \alpha}\right)^{\frac{m}{2}},
\end{aligned}$$

where $\alpha > 0$ is a free parameter. Let $\alpha = \frac{m}{2\lambda_{\max}^2(\boldsymbol{c})\tau_1}$ and $t = \lambda_{\max}^2(\boldsymbol{c})\tau_1$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) \leq \lambda_{\max}^2(\boldsymbol{c})\tau_1\right) &\leq \mathrm{e}^{\frac{m}{2}}\left(\frac{2\pi\tau_1}{m}\right)^{\frac{m}{2}} \\
&\leq \mathrm{e}^{\frac{m}{2}\left(K + \log\tau_1 - \log m\right)},
\end{aligned}$$

where $K = \log 2\pi e$. Hence, we have

$$\mathbb{P}\left(d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) > \lambda_{\max}^2(\boldsymbol{c})\tau_1\right) \geq 1 - \mathrm{e}^{\frac{m}{2}\left(K + \log\tau_1 - \log m\right)},$$

and with an union bound on C_r we get

$$\begin{aligned}
&\mathbb{P}\left(d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) > \lambda_{\max}^2(\boldsymbol{c})\tau_1 \quad \forall \boldsymbol{c} \in C_r\right) \\
&\qquad \geq 1 - 2^r \mathrm{e}^{\frac{m}{2}\left(K + \log\tau_1 - \log m\right)}.
\end{aligned}$$

To prove (5.86), note that for Z_i defined in (5.89), one has $Z_i \le \left(\left|\lambda_{\max}(\boldsymbol{c})\right| \chi^2(4)\right)^2$, thus

$$
\begin{aligned}
\sum_{i=1}^{m} Z_i &\le \lambda_{\max}^2 \sum_{i=1}^{m} \chi^4(4) \\
&\le \lambda_{\max}^2 \left(\sum_{i=1}^{m} \chi^2(4) \right)^2 \\
&\overset{d}{=} \lambda_{\max}^2 \left(\chi^2(4m) \right)^2,
\end{aligned}
$$

where the notation $\overset{d}{=}$ implies that they have the same distributions. Therefore, by Lemma 49 we have

$$
\begin{aligned}
&\mathbb{P}\left(d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) \ge \lambda_{\max}^2(\boldsymbol{c}) \left(4m(1+\tau_2)\right)^2 \right) \\
&\quad = \mathbb{P}\left(\sum_{i=1}^{m} Z_i \ge \lambda_{\max}^2 \left(4m(1+\tau_2)\right)^2 \right) \\
&\quad \le \mathbb{P}\left(\chi^2(4m) \ge 4m(1+\tau_2) \right) \\
&\quad \le \mathrm{e}^{-2m\left(\tau_2 - \log(1+\tau_2)\right)}.
\end{aligned}
$$

Hence, for any $\tau_2 > 0$, we have

$$
\begin{aligned}
&\mathbb{P}\left(d_A(|A\boldsymbol{x}|, |A\boldsymbol{c}|) < \lambda_{\max}^2(\boldsymbol{c}) \left(4m(1+\tau_2)\right)^2 \right) \\
&\qquad\qquad \ge 1 - \mathrm{e}^{-2m\left(\tau_2 - \log(1+\tau_2)\right)}.
\end{aligned}
$$

$\square$

Remark 28 (Expectation of $d_A(.,.)$)**.** *Note that* (5.89) *implies*

$$
\mathbb{E}\left[d\left(|A\boldsymbol{x}|, |A\boldsymbol{c}|\right) \right] = 8m \left(\lambda_1(\boldsymbol{c})^2 + \lambda_2(\boldsymbol{c})^2 + \lambda_1(\boldsymbol{c})\lambda_2(\boldsymbol{c}) \right) \tag{5.90}
$$

Proof. By (5.89) we obtain

$$\begin{aligned}
\mathbb{E}\left[d\left(|A\boldsymbol{x}|,|A\boldsymbol{c}|\right)\right] &= m\mathbb{E}\left[Z_1\right] \\
&= m\left(\lambda_1(\boldsymbol{c})^2\mathbb{E}\left[U^2\right] + \lambda_2(\boldsymbol{c})^2\mathbb{E}\left[V^2\right]\right. \\
&\quad \left. +2\lambda_1(\boldsymbol{c})\lambda_2(\boldsymbol{c})\mathbb{E}\left[UV\right]\right) \\
&= m\left(8\lambda_1(\boldsymbol{c})^2 + 8\lambda_2(\boldsymbol{c})^2 + 2\times 4\lambda_1(\boldsymbol{c})\lambda_2(\boldsymbol{c})\right) \\
&= 8m\left(\lambda_1(\boldsymbol{c})^2 + \lambda_2(\boldsymbol{c})^2 + \lambda_1(\boldsymbol{c})\lambda_2(\boldsymbol{c})\right)
\end{aligned}$$

□

5.6.4 Concentration of the gradient

Lemma 54. *Let $\boldsymbol{v} \in \mathbb{C}^n$ with $\|\boldsymbol{v}\| = 1$ and $\boldsymbol{z} \in \mathcal{C}_r$ be fixed. Then there exist constants $C_1, C_2, C_3 > 0$ such that,*

$$\begin{aligned}
&\mathbb{P}\left(\left|\mathrm{Re}\left(\boldsymbol{v}^*\left(\nabla d_A(\boldsymbol{z}) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right]\right)\right)\right| > m\epsilon \inf_{\theta\in\mathbb{R}}\left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z}\right\|\right) \\
&\qquad \le C_2\mathrm{e}^{-C_1\sqrt{m\epsilon}}, \qquad \forall\, \epsilon \ge C_3 m^{-\frac{1}{3}}.
\end{aligned}$$

Proof. In this proof, we will use the notations we introduced in (5.88) in the proof of Theorem 12. Since we have assumed that for any codeword $\boldsymbol{c}$, $\|\boldsymbol{x}\|_2 = \|\boldsymbol{c}\|_2 = 1$, according to Lemma 52,

$\lambda_1(\boldsymbol{c}) + \lambda_2(\boldsymbol{c}) = 0$. Hence,

$$
\begin{aligned}
&\nabla d_A(\boldsymbol{z}) \\
&\quad = 2\sum_{k=1}^{m} \left(\left|\boldsymbol{a}_k^*\boldsymbol{z}\right|^2 - \left|\boldsymbol{a}_k^*\boldsymbol{x}\right|^2\right) \boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}, \\
&\quad = 2\sum_{k=1}^{m} \boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z} \left(\overline{Q}\boldsymbol{a_k}\right)^* \begin{pmatrix} -\lambda_1(\boldsymbol{z}) & & & \\ & \lambda_1(\boldsymbol{z}) & & \\ & & \ddots & \\ & & & 0 \end{pmatrix} \overline{Q}\boldsymbol{a}_k \\
&\quad = 2\lambda_1(\boldsymbol{z})\sum_{k=1}^{m} \boldsymbol{a}_k\boldsymbol{a}_k^*\boldsymbol{z}U_k, \qquad (5.91)
\end{aligned}
$$

where $\lambda_i(\boldsymbol{z}), Q$ are as defined in (5.88), and

$$U_k \triangleq \left(\left|\left(\overline{Q}\boldsymbol{a}_k\right)_2\right|^2 - \left|\left(\overline{Q}\boldsymbol{a}_k\right)_1\right|^2\right). \qquad (5.92)$$

It is straightforward to check that

$$\mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right] = 8m(\boldsymbol{z}\boldsymbol{z}^* - \boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{z}. \qquad (5.93)$$

We also have

$$\lambda_1(\boldsymbol{z}) = -\lambda_2(\boldsymbol{z}) = \lambda_{\max}(\boldsymbol{z}).$$

By (5.91) we have,

$$
\begin{aligned}
&\operatorname{Re}\left(\boldsymbol{v}^*\left(\nabla d_A(\boldsymbol{z})-\mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right]\right)\right)\\
&\quad=2\lambda_{\max}(\boldsymbol{z})\sum_{k=1}^{m}\operatorname{Re}\Big((\boldsymbol{v}^*\boldsymbol{a}_k)(\boldsymbol{a}_k^*\boldsymbol{z})U_k\\
&\qquad-\mathbb{E}\left[(\boldsymbol{v}^*\boldsymbol{a}_k)(\boldsymbol{a}_k^*\boldsymbol{z})U_k\right]\Big)\\
&\quad=2\lambda_{\max}(\boldsymbol{z})\left(\sum_{k=1}^{m}\operatorname{Re}\left((\boldsymbol{v}^*\boldsymbol{a}_k)(\boldsymbol{a}_k^*\boldsymbol{z})U_k\right)\right.\\
&\qquad\left.-\sum_{k=1}^{m}\mathbb{E}\left[\operatorname{Re}\left((\boldsymbol{v}^*\boldsymbol{a}_k)(\boldsymbol{a}_k^*\boldsymbol{z})U_k\right)\right]\right)\\
&\quad=2\lambda_{\max}(\boldsymbol{z})\sum_{k=1}^{m}Y_k-\mathbb{E}\left[Y_k\right],
\end{aligned}
\tag{5.94}
$$

where $Y_k=\operatorname{Re}\left((\boldsymbol{v}^*\boldsymbol{a}_k)(\boldsymbol{a}_k^*\boldsymbol{z})U_k\right)$. We claim Y_k satisfies all assumptions of Lemma 51. To prove this note that since $\|\boldsymbol{v}\|=\|\boldsymbol{z}\|=1$, all $\boldsymbol{v}^*\boldsymbol{a}_k,\ \boldsymbol{a}_k^*\boldsymbol{z},\ (\overline{Q}\boldsymbol{a}_k)_1,\ (\overline{Q}\boldsymbol{a}_k)_2$ have the same distribution as $\mathcal{N}(0,1)+i\mathcal{N}(0,1)$. Therefore, Y_k can be written as

$$
Y_k=\sum_{j=1}^{16}W_{1,j,k}W_{2,j,k}W_{3,j,k}W_{4,j,k},
\tag{5.95}
$$

where $W_{l,j,k}\sim\mathcal{N}(0,1)\quad 1\le l\le 4,\ 1\le j\le 16,\ 1\le k\le m.$ We should emphasize that $W_{1,j,k},W_{2,j,k},W_{3,j,k},W_{4,j,k}$ may be dependent on each other but are independent of $W_{1,j,k'},W_{2,j,k'}$, $W_{3,j,k'},W_{4,j,k'}$, if $k\neq k'$. Hence, we have

$$
\begin{aligned}
\mathbb{P}\left(\left|Y_k\right| > \tau\right) &\leq \mathbb{P}\left(\exists j \leq 16; \quad \left|W_{1,j,k}W_{2,j,k}W_{3,j,k}W_{4,j,k}\right| > \frac{\tau}{16}\right) \\
&\leq \mathbb{P}\left(\exists j \leq 16,\ l \leq 4; \quad \left|W_{l,j,k}\right| > \sqrt[4]{\frac{\tau}{16}}\right) \\
&\leq 16 \times 4 \times \mathrm{e}^{-\frac{1}{c^2}\sqrt{\frac{\tau}{16}}} \\
&\leq 64\mathrm{e}^{-c'\sqrt{\tau}}. \qquad (5.96)
\end{aligned}
$$

To have (5.96), one may choose $c' = \frac{1}{4c^2}$, where c is a constant for which $\mathbb{P}\left(\left|\mathcal{N}(0,1)\right| > \tau\right) \leq \mathrm{e}^{-\frac{\tau^2}{c^2}}$.

Hence, by Lemma 51, there exist constants C_3 such that

$$
\mathbb{P}\left(\left|\sum_{k=1}^{m} Y_k - \mathbb{E}\left[Y_k\right]\right| > m\frac{\epsilon}{2}\right) \leq 4\mathrm{e}^{-\frac{c'}{2}\sqrt{m\epsilon}}, \quad \forall\ \epsilon \geq C_3 m^{-\frac{1}{3}}. \qquad (5.97)
$$

Thus,

$$
\begin{aligned}
\mathbb{P}\left(\left|\mathrm{Re}\left(\boldsymbol{v}^*\left(\nabla d_A(\boldsymbol{z}) - \mathbb{E}\left[\nabla d_A(\boldsymbol{z})\right]\right)\right)\right| > m\epsilon\lambda_{\max}(\boldsymbol{z})\right) \\
\leq C_2\mathrm{e}^{-C_1\sqrt{m\epsilon}}, \quad \forall\ \epsilon \geq C_3 m^{-\frac{1}{3}}.
\end{aligned}
$$

Furthermore, note that by (5.81) and Lemma 47 we have

$$
\begin{aligned}
\lambda_{\max}(\boldsymbol{z})^2 &\leq \lambda_1(\boldsymbol{z})^2 + \lambda_2(\boldsymbol{z})^2 \\
&= 2(1 - \left|\boldsymbol{x}^*\boldsymbol{z}\right|) \\
&= \inf_{\theta\in\mathbb{R}} \left\|\mathrm{e}^{i\theta}\boldsymbol{x} - \boldsymbol{z}\right\|^2.
\end{aligned}
$$

□

5.7 Conclusions

In this chapter, we have studied the problem of employing compression codes to solve the phase retrieval problem. Given a class of structured signals and a corresponding compression code, we have proposed COPER, which provably recovers structured signals in that class from their phaseless measurements using the compression code. Our results have shown that, in noiseless phase retrieval, asymptotically, the required sampling rate for almost zero-distortion recovery, modulo the phase, is the same as noiseless compressed sensing.

COPER is based on a combinatorial optimization problem. Hence, we have also introduced an iterative algorithm named gradient descent COPER (GD-COPER). We have shown that GD-COPER can return an accurate estimate of the signal in polynomial time (under mild assumptions on the compression code and the initialization of the algorithm). However, GD-COPER requires more measurements than COPER. The simulation results not only confirms the excellent performance of GD-COPER, but also shows the GD-COPER can perform pretty well even with a far initial point from the target. This confirms that the very mild condition we had in Corollary 9 for the theoretical guarantee, also works in practice.

Table 5.1: Results for the Gaussian measurement matrices. Both the GD-COPER algorithm and the Wirtinger flow are initialized with a white image. The setting of all the other parameters is described in Section 5.4.2. The notation DVG in the table refers to the fact that the algorithm either stops since the norm of z diverges to infinity, or returns a result with negative PSNR.

Target	$\frac{m}{n}$	GD-COPER		Wirtinger Flow	
		PSNR	Run time	PSNR	Run time
	0.5	23.22	11.2	DVG	8.68
	0.73	24.44	15.2	DVG	15.2
	1.0	25.63	18.9	DVG	30.6
	2.0	31.79	29.3	DVG	106.
	0.5	22.58	13.1	4.83	39.3
	0.73	24.79	15.6	6.5	60.3
	1.0	26.43	17.9	8.68	79.6
	2.0	31.91	31.3	17.71	135.
	0.5	21.42	11.9	DVG	13.4
	0.73	23.73	15.2	DVG	33.1
	1.0	25.84	18.8	10.94	82.8
	2.0	32.36	30.1	29.66	136.
	0.5	25.5	12.2	DVG	14.2
	0.73	27.43	13.9	DVG	22.1
	1.0	29.15	18.3	DVG	41.7
	2.0	34.76	29.6	33.36	140.
	0.5	22.03	12.4	3.92	43.1
	0.73	24.08	15.1	5.68	59.0
	1.0	26.67	17.4	7.94	74.2
	2.0	33.07	28.4	14.35	143.
	0.5	21.83	11.2	DVG	7.64
	0.73	23.35	15.7	DVG	20.7
	1.0	24.52	19.9	DVG	34.1
	2.0	32.67	28.8	35.65	135.
	0.5	17.49	10.9	DVG	10.9
	0.73	18.68	14.4	DVG	20.9
	1.0	21.44	19.0	DVG	37.8
	2.0	29.04	29.8	32.74	140.

Table 5.2: Comparison between the Wirtinger flow and the GD-COPER with the coded diffraction patterns for different values of m/n. The true images of the simulations are shown in the first column.

Target	$\frac{m}{n}$	GD-COPER		Wirtinger Flow	
		PSNR	Run time	PSNR	Run time
	1	27.8	13.0	DVG	1.2
	2	34.7	16.9	DVG	2.0
	3	36.2	18.1	DVG	2.7
	4	39.7	19.8	DVG	4.2
	5	42.1	14.2	DVG	4.1
	6	38.5	14.6	DVG	4.3
	7	42.7	15.4	DVG	5.7
	8	44.5	15.6	DVG	6.2
	9	38.9	16.1	23.6	18.9
	10	49.1	15.1	17.8	12.7
	15	38.6	17.3	13.0	23.0
	1	19.4	14.3	4.1	2.8
	2	28.6	19.5	7.2	5.1
	3	33.4	17.6	10.1	7.4
	4	34.5	14.4	13.1	5.9
	5	39.0	14.9	16.2	7.4
	6	40.2	15.0	18.9	8.0
	7	44.0	14.8	22.4	9.1
	8	45.9	15.3	25.2	10.0
	9	45.6	15.1	28.0	11.4
	10	47.4	15.7	31.8	12.9
	15	50.9	19.6	44.1	29.8

Table 5.3: Comparison between the Wirtinger flow and the GD-COPER with the coded diffraction patterns for different values of m/n. The true images in the simulations are shown in the leftmost column.

Target	$\frac{m}{n}$	GD-COPER		Wirtinger Flow	
		PSNR	Run time	PSNR	Run time
	1	29.3	14.2	DVG	1.4
	2	34.0	18.5	DVG	2.7
	3	36.8	17.6	DVG	4.7
	4	38.0	15.1	17.6	6.1
	5	40.7	15.9	20.4	8.0
	6	44.2	14.6	22.8	8.2
	7	42.1	14.9	28.1	9.2
	8	40.7	16.2	30.5	9.8
	9	42.2	16.0	33.8	11.8
	10	49.9	16.3	37.6	13.1
	15	41.8	16.9	52.1	23.0
	1	27.2	13.8	DVG	1.4
	2	32.3	16.6	DVG	2.1
	3	35.8	16.9	DVG	3.9
	4	36.4	17.1	15.6	7.3
	5	38.7	15.1	17.9	6.7
	6	39.4	14.9	20.1	7.8
	7	42.9	15.1	27.1	8.9
	8	47.5	15.6	30.5	10.1
	9	40.9	18.2	33.2	18.9
	10	47.6	15.9	36.9	13.0
	15	48.6	17.7	53.3	23.0
	1	23.1	13.8	DVG	1.4
	2	28.0	17.9	DVG	2.6
	3	32.0	17.9	DVG	3.7
	4	34.3	18.4	16.2	7.1
	5	38.1	16.9	19.0	9.2
	6	38.5	15.3	21.2	8.0
	7	42.0	15.2	22.6	9.2
	8	44.6	15.6	29.2	9.9
	9	43.2	19.4	32.3	14.7
	10	43.0	16.3	36.1	13.1
	15	52.0	17.5	49.7	23.4

Table 5.4: Wirtinger Flow performance with spectral and all-white initialization

Target	$\frac{m}{n}$	All-white			Spectral		
		n-init-err	PSNR	Run time	n-init-err	PSNR	Run time
	1	0.57	DVG	1.7	1.39	DVG	2.4
	2	0.57	DVG	1.4	1.39	DVG	4.4
	3	0.57	17.1	4.3	1.39	DVG	6.2
	4	0.57	20.3	5.5	1.37	DVG	7.4
	5	0.57	23.2	6.5	1.37	DVG	9.3
	6	0.57	26.9	7.7	1.38	DVG	11.2
	7	0.57	29.4	9.8	1.13	DVG	12.3
	8	0.57	32.8	13.5	0.89	DVG	16.2
	9	0.57	36.2	17.5	0.63	9.4	74.8
	10	0.57	39.0	44.5	0.64	DVG	32.5
	15	0.57	51.3	50.6	0.49	20.5	99.4

Target	$\frac{m}{n}$	All-white			Spectral		
		n-init-err	PSNR	Run time	n-init-err	PSNR	Run time
	1	0.86	DVG	2.8	1.39	DVG	4.8
	2	0.86	12.1	9.1	1.39	DVG	8.4
	3	0.86	15.1	11.4	1.39	DVG	10.9
	4	0.86	18.2	14.9	1.39	DVG	13.6
	5	0.86	21.1	20.0	1.41	DVG	15.1
	6	0.86	24.2	25.1	1.37	DVG	17.1
	7	0.86	27.4	28.2	1.06	DVG	19.5
	8	0.86	30.4	28.4	0.9	DVG	22.8
	9	0.86	33.4	31.6	1.33	DVG	26.0
	10	0.86	35.5	32.2	0.6	24.4	56.1
	15	0.86	56.7	43.7	0.48	11.0	81.7

Target	$\frac{m}{n}$	All-white			Spectral		
		n-init-err	PSNR	Run time	n-init-err	PSNR	Run time
	1	0.98	DVG	2.6	1.39	DVG	5.0
	2	0.98	DVG	2.8	1.39	DVG	7.2
	3	0.98	14.0	9.6	1.39	DVG	8.9
	4	0.98	17.0	11.9	1.4	DVG	10.6
	5	0.98	20.0	15.9	1.38	DVG	12.6
	6	0.98	23.2	17.7	1.21	DVG	15.0
	7	0.98	26.1	21.8	1.31	DVG	17.0
	8	0.98	29.0	24.1	1.39	DVG	17.9
	9	0.98	32.2	26.2	0.65	20.4	30.8
	10	0.98	34.7	13.6	0.6	21.3	30.9
	15	0.98	57.1	21.9	0.48	21.2	55.3

Table 5.5: Wirtinger Flow performance with spectral and all-white initialization

Target	$\frac{m}{n}$	All-white			Spectral		
		n-init-err	PSNR	Run time	n-init-err	PSNR	Run time
	1	2.84	5.1	2.6	1.39	DVG	3.1
	2	2.84	8.1	3.6	1.4	DVG	4.9
	3	2.84	11.0	4.7	1.39	DVG	7.1
	4	2.84	14.1	5.5	1.4	DVG	8.8
	5	2.84	17.3	6.9	1.38	DVG	10.6
	6	2.84	20.3	8.3	1.36	DVG	12.3
	7	2.84	22.9	9.7	1.36	DVG	14.2
	8	2.84	26.2	11.1	1.39	DVG	16.2
	9	2.84	28.4	12.8	1.1	DVG	17.8
	10	2.84	32.3	13.4	0.6	20.8	30.8
	15	2.84	45.0	22.2	0.48	39.7	53.4

Target	$\frac{m}{n}$	All-white			Spectral		
		n-init-err	PSNR	Run time	n-init-err	PSNR	Run time
	1	0.83	DVG	1.4	1.39	DVG	3.1
	2	0.83	12.6	3.8	1.39	DVG	5.1
	3	0.83	15.6	4.7	1.39	DVG	6.8
	4	0.83	18.6	5.6	1.39	DVG	8.7
	5	0.83	21.5	6.7	1.31	DVG	10.1
	6	0.83	24.8	8.1	1.36	DVG	12.2
	7	0.83	28.0	9.5	1.27	DVG	14.1
	8	0.83	30.9	10.7	0.7	DVG	21.6
	9	0.83	33.6	12.3	1.05	DVG	17.9
	10	0.83	36.3	13.0	0.85	DVG	19.6
	15	0.83	59.3	22.5	0.48	28.5	54.0

Target	$\frac{m}{n}$	All-white			Spectral		
		n-init-err	PSNR	Run time	n-init-err	PSNR	Run time
	1	1.25	7.4	2.5	1.4	DVG	3.0
	2	1.25	10.4	3.3	1.39	DVG	4.9
	3	1.25	13.5	4.9	1.39	DVG	6.5
	4	1.25	16.5	5.7	1.38	DVG	8.1
	5	1.25	19.3	7.1	1.4	DVG	10.2
	6	1.25	22.6	8.0	1.32	DVG	12.2
	7	1.25	25.9	9.9	1.02	DVG	13.7
	8	1.25	28.6	10.6	0.71	11.1	23.9
	9	1.25	31.9	22.3	0.77	DVG	22.1
	10	1.25	34.3	27.3	0.64	DVG	25.2
	15	1.25	55.2	42.8	0.49	24.2	79.6

Table 5.6: The impact of initialization on the performance of GD-COPER and Wirtinger flow. “n-init-error" is the normalized mean square error of the initialization. The initializations chosen in this simulation are in the form of $\boldsymbol{x}_{\text{init}} = \lambda \boldsymbol{x}_o + (1-\lambda)\boldsymbol{x}$, where $\boldsymbol{x}_o$ is an all-white image and $\boldsymbol{x}$ is the true signal.

Target	n-init-error	λ	$\frac{m}{n}=1$		$\frac{m}{n}=2$		$\frac{m}{n}=3$	
			GD-C	WF	GD-C	WF	GD-C	WF
	0.0	0.0	29.94	inf	32.55	inf	34.23	inf
	0.49	0.1	29.46	DVG	32.03	DVG	33.79	26.78
	0.98	0.2	29.25	DVG	32.03	DVG	34.03	24.11
	1.48	0.3	28.36	14.55	32.19	17.57	33.96	20.59
	1.97	0.4	27.12	12.06	31.22	15.07	33.18	18.09
	2.46	0.5	25.0	10.13	30.63	13.13	33.59	16.15
	2.95	0.6	23.03	8.54	30.67	11.55	33.32	14.57
	3.44	0.7	21.41	7.2	29.66	10.21	33.21	13.22
	3.94	0.8	20.59	6.04	28.51	9.04	31.37	12.05
	4.43	0.9	20.36	5.01	27.69	8.01	30.89	11.01
	4.92	1.0	18.52	4.09	27.95	7.09	31.82	10.07

Target	n-init-error	λ	$\frac{m}{n}=1$		$\frac{m}{n}=2$		$\frac{m}{n}=3$	
			GD-C	WF	GD-C	WF	GD-C	WF
	0.0	0.0	30.17	inf	34.68	inf	37.67	inf
	0.08	0.1	29.53	DVG	34.59	DVG	37.32	DVG
	0.17	0.2	29.6	DVG	33.67	DVG	37.35	DVG
	0.25	0.3	29.65	DVG	33.6	DVG	37.74	DVG
	0.34	0.4	29.32	DVG	33.7	DVG	37.51	DVG
	0.42	0.5	28.18	DVG	34.19	DVG	36.64	18.73
	0.5	0.6	27.68	DVG	34.92	DVG	35.95	19.86
	0.59	0.7	28.37	DVG	35.08	DVG	35.92	18.66
	0.67	0.8	28.13	DVG	35.12	14.24	36.23	17.56
	0.76	0.9	29.21	DVG	34.79	13.43	36.04	16.54
	0.84	1.0	29.15	DVG	34.16	12.59	35.64	15.63

Table 5.7: The impact of initialization on the performance of GD-COPER and Wirtinger flow. “n-init-error" is the normalized mean square error of the initialization. The initializations chosen in this simulation are in the form of $\boldsymbol{x}_{\text{init}} = \lambda \boldsymbol{x}_o + (1-\lambda)\boldsymbol{x}$, where $\boldsymbol{x}_o$ is an all-white image and $\boldsymbol{x}$ is the true signal.

Target	n-init-error	λ	$\frac{m}{n} = 1$		$\frac{m}{n} = 2$		$\frac{m}{n} = 3$	
			GD-C	WF	GD-C	WF	GD-C	WF
	0.0	0.0	27.84	inf	31.55	inf	35.11	inf
	0.09	0.1	28.04	DVG	31.5	DVG	35.19	DVG
	0.17	0.2	27.44	DVG	31.24	DVG	35.12	DVG
	0.26	0.3	26.99	DVG	31.47	DVG	35.26	DVG
	0.35	0.4	26.68	DVG	31.23	DVG	35.02	DVG
	0.43	0.5	26.89	DVG	31.62	DVG	34.66	19.12
	0.52	0.6	26.5	DVG	32.18	DVG	33.89	18.97
	0.61	0.7	26.69	DVG	32.4	DVG	33.54	17.94
	0.7	0.8	26.56	DVG	31.97	13.86	33.71	17.13
	0.78	0.9	26.26	DVG	31.74	12.92	34.16	16.12
	0.87	1.0	26.71	DVG	32.0	12.11	34.6	15.21

Target	n-init-error	λ	$\frac{m}{n} = 1$		$\frac{m}{n} = 2$		$\frac{m}{n} = 3$	
			GD-C	WF	GD-C	WF	GD-C	WF
	0.0	0.0	23.65	inf	26.23	inf	27.53	inf
	0.1	0.1	23.55	DVG	26.26	DVG	27.65	DVG
	0.2	0.2	23.69	DVG	26.14	DVG	27.68	DVG
	0.3	0.3	23.49	DVG	26.28	DVG	27.46	DVG
	0.39	0.4	23.45	DVG	26.14	DVG	27.49	DVG
	0.49	0.5	23.49	DVG	26.13	DVG	27.6	DVG
	0.59	0.6	23.45	DVG	26.19	DVG	27.56	DVG
	0.69	0.7	23.48	DVG	26.18	DVG	27.43	16.88
	0.79	0.8	22.82	DVG	26.44	12.72	27.53	15.9
	0.89	0.9	22.97	DVG	26.43	11.9	27.5	14.92
	0.99	1.0	22.62	DVG	26.27	11.03	27.56	14.0

Chapter 6: Future research directions

In this chapter, we cover several open questions and challenges regarding topics we discussed in this thesis.

6.1 Probabilistic toolbox

The growing literature of high-dimensional statistics and machine learning has created a major demand for understanding the behavior of heavy-tailed probability measures. For instance, it is straightforward to see how heavy-tailed distributions appear in the analyses of deep learning models. In a deep neural network of the form

$$\boldsymbol{y} = \sigma\left(\boldsymbol{W}_d\left(...\sigma\left(\boldsymbol{W}_2\sigma\left(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1\right) + \boldsymbol{b}_2\right) + ...\right) + \boldsymbol{b}_d\right),$$

where d is the depth of the network and $\boldsymbol{W}_i, \boldsymbol{b}_i, \sigma$ denote weights, biases, and the non-linearity respectively, we usually have a heavy-tailed distribution for the output $\boldsymbol{y}$ because of the iterative multiplications from a layer to the next one. More examples of the appearance of heavy-tailed distributions in current studies are discussed in [48, 168].

Although there are several results in the probability theory literature to address questions about heavy-tailed distributions [49, 64], they are either too general with many free parameters to tune, or are incapable of offering sharp results we require for statistical analyses. Sometimes, the optimization problems that have to be solved for setting free parameters appropriately in such results is as challenging as the concentration problem we aim to address. Hence, many of the existing results are not suitable to address questions raised in the statistical analyses in the high-dimensional settings. For more detailed discussion about this issue refer to Chapter 2.

While the average of heavy-tailed iid random variables have been discussed in Chapter 2, there are many more forms of statistics of dependent heavy-tailed distributions that need to be analyzed. For instance, in the study of random matrices, studying quadratic sums of the form $\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}$ is a random vector and $\boldsymbol{A}$ is a matrix independent from $\boldsymbol{x}$ is quite important [169]. Note that the well-known Hanson-Wright inequality is applicable only when $\boldsymbol{x}$ is a subGaussian vector [170, 65]. Hence, obtaining accurate and easy-to-use concentration inequalities for the functions of dependent heavy-tailed distributions is one of the open questions that need to be investigated more carefully in the future.

6.2 Theoretical gap for recovery of a structured signal

As discussed in Chapter 5, there is a gap between the sufficient number of measurements to recover a structured signal and the sample complexity that is required for a convergence guarantee of almost all proposed algorithms. If the signal is known to lie in a compact set with effective dimension k, then $O(k \log n)$ random measurements are sufficient to determine the signal uniquely [42, 41]. However, the required sample complexity for most of the known algorithms is at least $O(k^2 \log n)$. A few results have suggested that it is possible to reduce the sample complexity of practical methods to $O(k \log n)$. For instance, [171] shows that for a very specific form of sensing vectors, which is far from what one has in practice, it is possible to prove convergence with $O(k \log n)$ observations. A more realistic result for real signals, i.e. $\boldsymbol{x} \in \mathbb{R}^n$, with optimal sample complexity appeared in [43] where it is assumed a generative model for representing the signal of interest and a gradient descent method is used on the input domain of the generative model. Nevertheless, this result suffers from some limitations as well; First, it is proved only for real valued signals. Second, it assumes some technical conditions on the sensing vectors which makes the result inapplicable to many practical settings. Finally, training generative models requires thousands of samples which may not be available for cutting-edge applications such as nano-particle imaging. In general, the optimal sample complexity and the optimal recovery method for generic structures and practical setups for the measurement system is still unknown. Exploring this issue is an excellent direction for future studies.

6.3 Rigorous analysis in practical setups

Despite the major progress that has been made in the theory of the phase retrieval problem, the existing theory is still limited and incapable of addressing many challenges that arise in applications. We discuss an instance of such limitations in more details below.

In non-convex optimization problems, which are used ubiquitously for solving inverse problems such as phase retrieval, it is essential to start with a point close to the signal of interest. Finding such an initial point, however, is a cumbersome task. As discussed in Chapters 3 and 4, spectral methods are very common for such initializations. In these methods, one uses the leading eigenvector of the following matrix

$$\boldsymbol{M} = \frac{1}{m}\sum_{i=1}^{m} \mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^* \tag{6.1}$$

as an initial point. In an ideal scenario, we want to understand the performance of the spectral methods for matrix ensembles that are used in practice and for realistic prior assumptions about the signal of interest. Unfortunately, the scope of the most of the theoretical works in phase retrieval have remained limited to simple random sensing matrices, such as Gaussian ensembles [29], uniform sample from orthogonal group [32], right invariant distributions [172]; and independent [172] and Gaussian distribution for entries of the signal [33]. On the other hand, the sensing matrices that are used in real systems, are often structured with limited amount of randomness. There are recent empirical evidence that some of the results on simpler matrix ensembles explain the behavior of more structured sensing matrices [173]. However, most such hypotheses have remained unproven.

We already have taken some steps toward filling this gap. For instance, [172] conjectures, based on tools borrowed from statistical physics, the optimal performance of spectral methods is in fact attained by a matrix of the form (6.1). Hence, working with this specific form has no cost for generality of our results. Moreover, Chapter 4 which is based on [33] extends the results obtained in Chapter 3, for partial orthogonal matrices, to partial Hadamard matrices which are better approximation for the sensing matrices that are being used in practice. We hope that this progress starts a more serious investigations of structured sensing matrices and signals.

6.4 Extension to other inverse problems

Phase retrieval is a very specific instance of inverse problems that emerge in imaging systems. Extension of the tools we developed in the context of phase retrieval, such as GD-COPER and spectral methods for initialization, to more complex versions of the phase retrieval or even other imaging systems, such as phase retrieval with multiplicative noise [174], coherent diffraction single-shot imaging with random rotation [175], hyperspectral imaging [176], and Nuclear Magnetic Resonance (NMR) [177], or other inverse problems, such as blind deconvolution [178], is an interesting direction of research. Below, I describe a few examples in more details.

Phase retrieval from a randomly rotated object appears in the imaging of nano-particles, where we can obtain a single shot from a particle and have no control over the orientation of the particle in the imaging system [175]. In this case, the mathematical formulation of the measurements would be

$$y_i = \left| \boldsymbol{a}_i^* \boldsymbol{R}_i \boldsymbol{x} \right| + \epsilon_i, \tag{6.2}$$

where $\boldsymbol{R}_i$s are uniform independent samples from the Unitary or the Orthogonal group.

Another example is blind deconvolution problem. In this problem, one would like to recover the signal $g(\boldsymbol{x})$, from blurry and noisy measurements

$$\boldsymbol{y} = \int h(\boldsymbol{t}) g(\boldsymbol{x} - \boldsymbol{t}) d\boldsymbol{t} + \boldsymbol{\epsilon}, \tag{6.3}$$

where $\boldsymbol{\epsilon}$ denotes the measurement noise and h represents an unknown convolution kernel.

Utilizing compression codes, discussed in Chapter 5 for COPER, to take advantage of the prior knowledge about the signal in order to minimize the sample complexity sounds promising for the inverse problems mentioned above. Furthermore, insights developed in the context of simple phase retrieval can pave the path for investigating more challenging problems with a similar nature. For instance, the geometry of local minimas and formulations of the cost functions for the phase retrieval which have shown the state-of-the-art performances can be inspiring to obtain the same level of knowledge for such more challenging inverse problems.

References

[1] C Fienup and J Dainty, “Phase retrieval and image reconstruction for astronomy,” *Image recovery: theory and application*, vol. 231, p. 275, 1987.

[2] R. P. Millane, “Phase retrieval in crystallography and optics,” *JOSA A*, vol. 7, no. 3, pp. 394–411, 1990.

[3] I. Vartanyants and I. Robinson, “Partial coherence effects on the imaging of small crystals using coherent x-ray diffraction,” *Journal of Physics: Condensed Matter*, vol. 13, no. 47, p. 10 593, 2001.

[4] M. V. Klibanov, P. E. Sacks, and A. V. Tikhonravov, “The phase retrieval problem,” *Inverse problems*, vol. 11, no. 1, p. 1, 1995.

[5] A. Walther, “The question of phase retrieval in optics,” *Optica Acta: International Journal of Optics*, vol. 10, no. 1, pp. 41–49, 1963.

[6] J. R. Fienup, “Phase retrieval algorithms: A personal tour,” *Applied optics*, vol. 52, no. 1, pp. 45–56, 2013.

[7] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, “Phase retrieval with application to optical imaging: A contemporary overview,” *IEEE signal processing magazine*, vol. 32, no. 3, pp. 87–109, 2015.

[8] A. Fannjiang and T. Strohmer, “The numerics of phase retrieval,” *arXiv preprint arXiv:2004.05788*, 2020.

[9] M. Eckert, *Max von laue and the discovery of x-ray diffraction in 1912*, 2012.

[10] R. W. Harrison, “Phase problem in crystallography,” *JOSA a*, vol. 10, no. 5, pp. 1046–1055, 1993.

[11] J. Miao, P. Charalambous, J. Kirz, and D. Sayre, “Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens,” *Nature*, vol. 400, no. 6742, pp. 342–344, 1999.

[12] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013.

[13] P Achuthan and K Venkatesan, “General principles of quantum mechanics,” *Handbuch der Physik*, vol. 5, no. Part 1, 1958.

[14] H. Reichenbach, *Philosophic foundations of quantum mechanics*. Courier Corporation, 1998.

[15] P. Jaming, "Uniqueness results in an extension of pauli's phase retrieval problem," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 413–441, 2014.

[16] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[17] E. J. Candès and X. Li, "Solving quadratic equations via phaselift when there are about as many equations as unknowns," *Foundations of Computational Mathematics*, vol. 14, no. 5, pp. 1017–1026, 2014.

[18] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming*," *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 89 –94, 2012, 16th IFAC Symposium on System Identification.

[19] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Comptes rendus mathematique*, vol. 346, no. 9-10, pp. 589–592, 2008.

[20] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Mathematical Programming*, vol. 149, no. 1-2, pp. 47–81, 2015.

[21] S. Bahmani and J. Romberg, "Phase retrieval meets statistical learning theory: A flexible convex relaxation," in *Artificial Intelligence and Statistics*, 2017, pp. 252–260.

[22] T. Goldstein and C. Studer, "Phasemax: Convex phase retrieval via basis pursuit," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2675–2689, 2018.

[23] R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

[24] J. R. Fienup, "Phase retrieval algorithms: A comparison," *Applied optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[25] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[26] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 773–794, 2017.

[27] J. Kishore, C. E. Yonina, and B. Hassibi, “Phase retrieval: An overview of recent developments,” *arXiv*, 2015.

[28] P. Netrapalli, P. Jain, and S. Sanghavi, “Phase retrieval using alternating minimization,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[29] Y. M. Lu and G. Li, “Phase transitions of spectral initialization for high-dimensional nonconvex estimation,” *Information and Inference, to appear*, 2019.

[30] W. Luo, W. Alghamdi, and Y. M. Lu, “Optimal spectral initialization for signal recovery with applications to phase retrieval,” *IEEE Transactions on Signal Processing*, vol. 67, no. 9, pp. 2347–2356, 2019.

[31] R. Dudeja, J. Ma, and A. Maleki, “Information theoretic limits for phase retrieval with subsampled haar sensing matrices,” *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 8002–8045, 2020.

[32] R. Dudeja, M. Bakhshizadeh, J. Ma, and A. Maleki, “Analysis of spectral methods for phase retrieval with random orthogonal matrices,” *IEEE Transactions on Information Theory*, 2020.

[33] R. Dudeja and M. Bakhshizadeh, “Universality of linearized message passing for phase retrieval with structured sensing matrices,” *arXiv preprint arXiv:2008.10503*, 2020.

[34] T. T. Cai, X. Li, and Z. Ma, “Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow,” *Ann. Statist.*, vol. 44, no. 5, pp. 2221–2251, Oct. 2016.

[35] G. Jagatap and C. Hegde, “Fast, sample-efficient algorithms for structured phase retrieval,” in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 4917–4927.

[36] P. Hand and V. Voroninski, “Compressed sensing from phaseless gaussian measurements via linear programming in the natural parameter space,” *arXiv preprint arXiv:1611.05985*, 2016.

[37] M. Moravec, J. Romberg, and R. Baraniuk, “Compressive phase retrieval,” *Proc SPIE*, vol. 6701, Oct. 2007.

[38] Y. Shechtman, A. Beck, and Y. C. Eldar, “Gespar: Efficient phase retrieval of sparse signals,” *IEEE transactions on signal processing*, vol. 62, no. 4, pp. 928–938, 2014.

[39] Y. Wang and Z. Xu, “Phase retrieval for sparse signals,” *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 531–544, 2014.

[40] B. Shi, S. Chen, Y. Tian, X. Fan, and Q. Lian, “Faspr: A fast sparse phase retrieval algorithm via the epigraph concept,” *Digital Signal Processing*, vol. 80, pp. 12–26, 2018.

[41] M. Bakhshizadeh, A. Maleki, and S. Jalali, “Using black-box compression algorithms for phase retrieval,” *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7978–8001, 2020.

[42] Y. C. Eldar and S. Mendelson, “Phase retrieval: Stability and recovery guarantees,” *Applied and Computational Harmonic Analysis*, vol. 36, no. 3, pp. 473–494, 2014.

[43] P. Hand, O. Leong, and V. Voroninski, “Phase retrieval under a generative prior,” in *Advances in Neural Information Processing Systems*, 2018, pp. 9136–9146.

[44] M. Bakhshizadeh, A. Maleki, and V. H. de la Pena, “Sharp concentration results for heavy-tailed distributions,” *arXiv preprint arXiv:2003.13819*, 2020.

[45] M. Rudelson and R. Vershynin, “Non-asymptotic theory of random matrices: Extreme singular values,” in *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, World Scientific, 2010, pp. 1576–1602.

[46] M. Bakhshizadeh, A. Maleki, and S. Jalali, “Using black-box compression algorithms for phase retrieval,” *IEEE Transactions on Information Theory*, vol. 66, no. 12, pp. 7978–8001, 2020.

[47] M. Vladimirova and J. Arbel, “Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier-tailed distributions,” *arXiv preprint arXiv:1905.04955*, 2019.

[48] M. Gurbuzbalaban, U. Simsekli, and L. Zhu, “The heavy-tail phenomenon in sgd,” *arXiv preprint arXiv:2006.04740*, 2020.

[49] S. V. Nagaev, “Large deviations of sums of independent random variables,” *The Annals of Probability*, pp. 745–789, 1979.

[50] M. G. Hahn, M. J. Klass, *et al.*, “Approximation of partial sums of arbitrary iid random variables and the precision of the usual exponential upper bound,” *The Annals of Probability*, vol. 25, no. 3, pp. 1451–1470, 1997.

[51] M. Klass, K. Nowicki, *et al.*, “Uniformly accurate quantile bounds via the truncated moment generating function: The symmetric case,” *Electronic Journal of Probability*, vol. 12, pp. 1276–1298, 2007.

[52] M. J. Klass and K. Nowicki, "Uniform bounds on the relative error in the approximation of upper quantiles for sums of arbitrary independent random variables," *Journal of Theoretical Probability*, vol. 29, no. 4, pp. 1485–1509, 2016.

[53] P. Hitczenko and S. Montgomery-Smith, "Measuring the magnitude of sums of independent random variables," *Annals of probability*, pp. 447–466, 2001.

[54] P. Hitczenko, S. J. Montgomery-Smith, and K. Oleszkiewicz, "Moment inequalities for sums of certain independent symmetric random variables," *Studia Math*, vol. 123, no. 1, pp. 15–42, 1997.

[55] A. K. Kuchibhotla and A. Chakrabortty, "Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression," *arXiv preprint arXiv:1804.02605*, 2018.

[56] L. V. Rozovskii, "Probabilities of large deviations on the whole axis," *Theory of Probability & Its Applications*, vol. 38, no. 1, pp. 53–79, 1994.

[57] L. Rozovskii, "Probabilities of large deviations of sums of independent random variables with common distribution function in the domain of attraction of the normal law," *Theory of Probability & Its Applications*, vol. 34, no. 4, pp. 625–644, 1990.

[58] D. Denisov, A. B. Dieker, V. Shneer, *et al.*, "Large deviations for random walks under subexponentiality: The big-jump domain," *The Annals of Probability*, vol. 36, no. 5, pp. 1946–1991, 2008.

[59] I. Kontoyiannis, M. Madiman, *et al.*, "Measure concentration for compound poisson distributions," *Electronic Communications in Probability*, vol. 11, pp. 45–57, 2006.

[60] M. Bazhba, J. Blanchet, C.-H. Rhee, and B. Zwart, "Sample-path large deviations for lévy processes and random walks with weibull increments," *arXiv preprint arXiv:1710.04013*, 2017.

[61] A. A. Borovkov, "Large deviation probabilities for random walks with semiexponential distributions.," *Siberian Mathematical Journal*, vol. 41, no. 6, pp. 1290–1324, 2000.

[62] A. A. Borovkov and A. A. Mogulskii, "Integro-local and integral theorems for sums of random variables with semiexponential distributions," *Siberian Mathematical Journal*, vol. 47, no. 6, pp. 990–1026, 2006.

[63] J. Nolan, *Stable distributions: models for heavy-tailed data*. Birkhauser New York, 2003.

[64] T. Mikosch and A. V. Nagaev, "Large deviations of heavy-tailed sums with applications in insurance," *Extremes*, vol. 1, no. 1, pp. 81–110, 1998.

[65] M. Rudelson, R. Vershynin, *et al.*, “Hanson-wright inequality and sub-gaussian concentration,” *Electronic Communications in Probability*, vol. 18, 2013.

[66] A. Dembo and O. Zeitouni, “Large deviations techniques and applications,” 1998.

[67] Y. Chen and E. J. Candes, “Solving random quadratic systems of equations is nearly as easy as solving linear systems,” *Communications on Pure and Applied Mathematics*, vol. 70, pp. 822–883, 2017.

[68] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, “Reshaped wirtinger flow and incremental algorithm for solving quadratic system of equations,” *arXiv preprint arXiv:1605.07719*, 2016.

[69] M. Mondelli and A. Montanari, “Fundamental limits of weak recovery with applications to phase retrieval,” in *Conference On Learning Theory*, PMLR, 2018, pp. 1445–1450.

[70] J. R. Fienup, “Reconstruction of an object from the modulus of its Fourier transform,” *Opt. Lett.*, vol. 3, no. 1, pp. 27–29, 1978.

[71] J. Ma, R. Dudeja, J. Xu, A. Maleki, and X. Wang, “Spectral method for phase retrieval: An expectation propagation perspective,” *arXiv preprint arXiv:1903.02505*, 2019.

[72] E. J. Candès, X. Li, and M. Soltanolkotabi, “Phase retrieval from coded diffraction patterns,” *Applied and Computational Harmonic Analysis*, vol. 39, no. 2, pp. 277–299, 2015.

[73] T. P. Minka, “Expectation propagation for approximate bayesian inference,” in *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[74] M. Opper and O. Winther, “Expectation consistent approximate inference,” *Journal of Machine Learning Research*, vol. 6, no. Dec, pp. 2177–2204, 2005.

[75] J. Ma, X. Yuan, and L. Ping, “Turbo compressed sensing with partial DFT sensing matrix,” *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 158–161, 2015.

[76] J. Ma and L. Ping, “Orthogonal AMP,” *IEEE Access*, vol. 5, pp. 2020–2033, 2017.

[77] S. Rangan, P. Schniter, and A. K. Fletcher, “Vector approximate message passing,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 1588–1592.

[78] K. Takeuchi, “Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements,” in *Information Theory (ISIT), 2017 IEEE International Symposium on*, IEEE, 2017, pp. 501–505.

[79] S. T. Belinschi, H. Bercovici, M. Capitaine, M. Fevrier, *et al.*, "Outliers in the spectrum of large deformed unitarily invariant models," *The Annals of Probability*, vol. 45, no. 6A, pp. 3571–3625, 2017.

[80] J. A. Mingo and R. Speicher, *Free probability and random matrices*. Springer, 2017, vol. 35.

[81] S. T. Belinschi, "The atoms of the free multiplicative convolution of two probability distributions," *Integral Equations and Operator Theory*, vol. 46, no. 4, pp. 377–386, 2003.

[82] S. T. Belinschi and H. Bercovici, "A new approach to subordination results in free probability," *Journal d'Analyse Mathématique*, vol. 101, no. 1, pp. 357–365, 2007.

[83] S. T. Belinschi, R. Speicher, J. Treilhard, and C. Vargas, "Operator-valued free multiplicative convolution: Analytic subordination theory and applications to random matrix theory," *International Mathematics Research Notices*, vol. 2015, no. 14, pp. 5933–5958, 2014.

[84] D. Voiculescu, "Limit laws for random matrices and free products," *Inventiones Mathematicae*, vol. 104, no. 1, pp. 201–220, 1991.

[85] S. T. Belinschi, "A note on regularity for free convolutions," in *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, vol. 42, 2006, pp. 635–648.

[86] M. Spruill *et al.*, "Asymptotic distribution of coordinates on high dimensional spheres," *Electronic communications in probability*, vol. 12, pp. 234–247, 2007.

[87] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

[88] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Foundations of Computational Mathematics*, vol. 18, no. 5, pp. 1131–1198, 2018.

[89] T. Bendory, R. Beinert, and Y. C. Eldar, "Fourier phase retrieval: Uniqueness and algorithms," in *Compressed Sensing and its Applications*, Springer, 2017, pp. 55–91.

[90] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, "Phase retrieval with polarization," *SIAM Journal on Imaging Sciences*, vol. 7, no. 1, pp. 35–66, 2014.

[91] A. S. Bandeira, Y. Chen, and D. G. Mixon, "Phase retrieval from power spectra of masked signals," *Information and Inference: a Journal of the IMA*, vol. 3, no. 2, pp. 83–102, 2014.

[92] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM review*, vol. 57, no. 2, pp. 225–251, 2015.

[93] D. Gross, F. Krahmer, and R. Kueng, "A partial derandomization of phaselift using spherical designs," *Journal of Fourier Analysis and Applications*, vol. 21, no. 2, pp. 229–266, 2015.

[94] ——, “Improved recovery guarantees for phase retrieval from coded diffraction patterns,” *Applied and Computational Harmonic Analysis*, vol. 42, no. 1, pp. 37–64, 2017.

[95] F. Krahmer and H. Rauhut, “Structured random measurements in signal processing,” *GAMM-Mitteilungen*, vol. 37, no. 2, pp. 217–238, 2014.

[96] V. Elser, T.-Y. Lan, and T. Bendory, “Benchmark problems for phase retrieval,” *SIAM Journal on Imaging Sciences*, vol. 11, no. 4, pp. 2429–2455, 2018.

[97] E. Abbasi, F. Salehi, and B. Hassibi, “Universality in learning from linear measurements,” *arXiv preprint arXiv:1906.08396*, 2019.

[98] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, “Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms,” *arXiv preprint arXiv:1805.09555*, 2018.

[99] M. Bayati and A. Montanari, “The dynamics of message passing on dense graphs, with applications to compressed sensing,” *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[100] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, “Optimal errors and phase transitions in high-dimensional generalized linear models,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 12, pp. 5451–5460, 2019.

[101] P. Schniter, S. Rangan, and A. K. Fletcher, “Vector approximate message passing for the generalized linear model,” in *2016 50th Asilomar Conference on Signals, Systems and Computers*, IEEE, 2016, pp. 1525–1529.

[102] S. Rangan, P. Schniter, and A. K. Fletcher, “Vector approximate message passing,” *IEEE Transactions on Information Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[103] K. Takeda, S. Uda, and Y. Kabashima, “Analysis of cdma systems that are characterized by eigenvalue spectrum,” *EPL (Europhysics Letters)*, vol. 76, no. 6, p. 1193, 2006.

[104] K. Takeda, A. Hatabu, and Y. Kabashima, “Statistical mechanical analysis of the linear vector channel in digital communication,” *Journal of Physics A: Mathematical and Theoretical*, vol. 40, no. 47, p. 14 085, 2007.

[105] Y. Kabashima, “Inference from correlated patterns: A unified theory for perceptron learning and linear vector channels,” in *Journal of Physics: Conference Series*, IOP Publishing, vol. 95, 2008, p. 012 001.

[106] J. Barbier, N. Macris, A. Maillard, and F. Krzakala, “The mutual information in random linear estimation beyond iid matrices,” in *2018 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2018, pp. 1390–1394.

[107] A. Maillard, B. Loureiro, F. Krzakala, and L. Zdeborová, "Phase retrieval in high dimensions: Statistical and computational phase transitions," *arXiv preprint arXiv:2006.05228*, 2020.

[108] D. Donoho and J. Tanner, "Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[109] H. Monajemi, S. Jafarpour, M. Gavish, D. L. Donoho, S. C. Collaboration, *et al.*, "Deterministic matrices matching the compressed sensing phase transitions of gaussian random matrices," *Proceedings of the National Academy of Sciences*, vol. 110, no. 4, pp. 1181–1186, 2013.

[110] S. Oymak and B. Hassibi, "A case for orthogonal measurements in linear inverse problems," in *2014 IEEE International Symposium on Information Theory*, IEEE, 2014, pp. 3175–3179.

[111] A. Abbara, A. Baker, F. Krzakala, and L. Zdeborová, "On the universality of noiseless linear estimation with respect to the measurement matrix," *arXiv preprint arXiv:1906.04735*, 2019.

[112] E Bolthausen, "On the high-temperature phase of the sherrington-kirkpatrick model," in *Seminar at EURANDOM, Eindhoven*, 2009.

[113] W.-K. Chen and W.-K. Lam, *Universality of approximate message passing algorithms*, 2020. arXiv: `2003.10431 [math.PR]`.

[114] S. B. Korada and A. Montanari, "Applications of the lindeberg principle in communications and statistical learning," *IEEE transactions on information theory*, vol. 57, no. 4, pp. 2440–2450, 2011.

[115] S. Oymak and J. A. Tropp, "Universality laws for randomized dimension reduction, with applications," *Information and Inference: A Journal of the IMA*, vol. 7, no. 3, pp. 337–446, 2018.

[116] D. L. Donoho and J. Tanner, "Counting the faces of randomly-projected hypercubes and orthants, with applications," *Discrete & computational geometry*, vol. 43, no. 3, pp. 522–541, 2010.

[117] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*. Cambridge university press, 2010, vol. 118.

[118] A. M. Tulino, G. Caire, S. Shamai, and S. Verdu, "Capacity of channels with frequency-selective and time-selective fading," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1187–1215, 2010.

[119] B. Farrell, “Limiting empirical singular value distribution of restrictions of discrete fourier transform matrices,” *Journal of Fourier Analysis and Applications*, vol. 17, no. 4, pp. 733–753, 2011.

[120] G. W. Anderson and B. Farrell, “Asymptotically liberating sequences of random unitary matrices,” *Advances in Mathematics*, vol. 255, pp. 381 –413, 2014.

[121] B. Cakmak, M. Opper, O. Winther, and B. H. Fleury, “Dynamical functional theory for compressed sensing,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 2143–2147.

[122] B. Çakmak and M. Opper, “Memory-free dynamics for the tap equations of ising models with arbitrary rotation invariant ensembles of random coupling matrices,” *arXiv preprint arXiv:1901.08583*, 2019.

[123] B. Cakmak and M. Opper, “Analysis of bayesian inference algorithms by the dynamical functional approach,” *Journal of Physics A: Mathematical and Theoretical*, 2020.

[124] B. Çakmak and M. Opper, “A dynamical mean-field theory for learning in restricted boltzmann machines,” *arXiv preprint arXiv:2005.01560*, 2020.

[125] M. Opper and B. Çakmak, “Understanding the dynamics of message passing algorithms: A free probability heuristics,” *arXiv preprint arXiv:2002.02533*, 2020.

[126] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, “Modelling the influence of data structure on learning in neural networks,” *arXiv preprint arXiv:1909.11500*, 2019.

[127] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, “Generalisation error in learning with random features and the hidden manifold model,” *arXiv preprint arXiv:2002.09339*, 2020.

[128] S. Goldt, G. Reeves, M. Mézard, F. Krzakala, and L. Zdeborová, “The gaussian equivalence of generative models for learning with two-layer neural networks,” *arXiv preprint arXiv:2006.14709*, 2020.

[129] F. G. Mehler, “Ueber die entwicklung einer function von beliebig vielen variablen nach laplaceschen functionen höherer ordnung.,” *Journal für die reine und angewandte Mathematik*, vol. 1866, no. 66, pp. 161–176, 1866.

[130] D. Slepian, “On the symmetrized kronecker power of a matrix and extensions of mehler’s formula for hermite polynomials,” *SIAM Journal on Mathematical Analysis*, vol. 3, no. 4, pp. 606–616, 1972.

[131] A. M. Tulino, G. Caire, S. Shamai, and S. Verdú, "Capacity of channels with frequency-selective and time-selective fading," *IEEE Transactions on Information Theory*, vol. 56, no. 3, pp. 1187–1215, 2010.

[132] R. N. Bhattacharya, "On errors of normal approximation," *The Annals of Probability*, pp. 815–828, 1975.

[133] E. S. Meckes, *The random matrix theory of the classical compact groups*. Cambridge University Press, 2019, vol. 218.

[134] K. Ball *et al.*, "An elementary introduction to modern convex geometry," *Flavors of geometry*, vol. 31, pp. 1–58, 1997.

[135] M. Gromov and V. D. Milman, "A topological application of the isoperimetric inequality," *American Journal of Mathematics*, vol. 105, no. 4, pp. 843–854, 1983.

[136] R. van Handel, "Probability in high dimension," PRINCETON UNIV NJ, Tech. Rep., 2014.

[137] B. A. Schmitt, "Perturbation bounds for matrix square roots and pythagorean sums," *Linear algebra and its applications*, vol. 174, pp. 215–227, 1992.

[138] S. Jalali and A. Maleki, "From compression to compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 40, no. 2, pp. 352–385, 2016.

[139] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM review*, vol. 57, no. 2, pp. 225–251, 2015.

[140] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "13 phase retrieval," *Optical Compressive Imaging*, p. 263, 2016.

[141] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Communications on pure and applied mathematics*, vol. 70, no. 5, pp. 822–883, 2017.

[142] O. Dhifallah and Y. M. Lu, "Fundamental limits of phasemax for phase retrieval: A replica analysis," in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, IEEE, 2017, pp. 1–5.

[143] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, "Phase retrieval via linear programming: Fundamental limits and algorithmic improvements," in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2017, pp. 1071–1077.

[144] R. Ghods, A. Lan, T. Goldstein, and C. Studer, "Linear spectral estimators and an application to phase retrieval," in *International Conference on Machine Learning*, PMLR, 2018, pp. 1734–1743.

[145] J. Ma, J. Xu, and A. Maleki, “Optimization-based amp for phase retrieval: The impact of initialization and ℓ_2 regularization,” *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3600–3629, 2019.

[146] Y. Chen, Y. Chi, J. Fan, and C. Ma, “Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval,” *Mathematical Programming*, vol. 176, no. 1, pp. 5–37, 2019.

[147] G. Wang, L. Zhang, G. B. Giannakis, and J. Chen, “Sparse phase retrieval via iteratively reweighted amplitude flow,” in *2018 26th European Signal Processing Conference (EUSIPCO)*, IEEE, 2018, pp. 712–716.

[148] R. Xu, M. Soltanolkotabi, J. P. Haldar, W. Unglaub, J. Zusman, A. F. Levi, and R. M. Leahy, “Accelerated wirtinger flow: A fast algorithm for ptychography,” *arXiv preprint arXiv:1806.05546*, 2018.

[149] C. Lucibello, L. Saglietti, and Y. Lu, “Generalized approximate survey propagation for high-dimensional estimation,” in *International Conference on Machine Learning*, PMLR, 2019, pp. 4173–4182.

[150] M. L. Moravec, J. K. Romberg, and R. G. Baraniuk, “Compressive phase retrieval,” in *Wavelets XII*, International Society for Optics and Photonics, vol. 6701, 2007, p. 670 120.

[151] H. Ohlsson, A. Yang, R. Dong, and S. Sastry, “Cprl–an extension of compressive sensing to the phase retrieval problem,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 1367–1375, 2012.

[152] Z. Yang, C. Zhang, and L. Xie, “Robust compressive phase retrieval via l1 minimization with application to image reconstruction,” *arXiv preprint arXiv:1302.0081*, 2013.

[153] P. Schniter and S. Rangan, “Compressive phase retrieval via generalized approximate message passing,” *IEEE Transactions on Signal Processing*, vol. 63, no. 4, pp. 1043–1055, Feb. 2015.

[154] G. Jagatap and C. Hegde, “Fast, sample-efficient algorithms for structured phase retrieval,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4924–4934.

[155] F. Salehi, E. Abbasi, and B. Hassibi, “Learning without the phase: Regularized phasemax achieves optimal sample complexity,” Neural Information Processing Systems, 2018.

[156] C. A. Metzler, A. Maleki, and R. G. Baraniuk, “Bm3d-prgamp: Compressive phase retrieval based on bm3d denoising,” in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2504–2508.

[157] C. Metzler, P. Schniter, A. Veeraraghavan, *et al.*, "Prdeep: Robust phase retrieval with a flexible deep network," in *International Conference on Machine Learning*, PMLR, 2018, pp. 3501–3510.

[158] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pająk, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian, *et al.*, "Flexisp: A flexible camera image processing framework," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, pp. 1–13, 2014.

[159] M. Soltanolkotabi, "Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization," *IEEE Transactions on Information Theory*, vol. 65, no. 4, pp. 2374–2400, 2019.

[160] C. A. Metzler, A. Maleki, and R. G. Baraniuk, "From denoising to compressed sensing," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5117–5144, 2016.

[161] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (red)," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.

[162] F. Shamshad and A. Ahmed, "Robust compressive phase retrieval via deep generative priors," *arXiv preprint arXiv:1808.05854*, 2018.

[163] S. Beygi, S. Jalali, A. Maleki, and U. Mitra, "An efficient algorithm for compression-based compressed sensing," *Information and Inference: A Journal of the IMA*, vol. 8, no. 2, pp. 343–375, 2019.

[164] Y. Dar, M. Elad, and A. M. Bruckstein, "Restoration by compression," *IEEE Transactions on Signal Processing*, vol. 66, no. 22, pp. 5833–5847, 2018.

[165] F. E. Rezagah, S. Jalali, E. Erkip, and H. V. Poor, "Compression-based compressed sensing," *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6735–6752, 2017.

[166] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM Journal on Mathematical Analysis*, vol. 45, no. 5, pp. 3019–3033, 2013.

[167] A. Maleki and D. L. Donoho, "Optimally tuned iterative reconstruction algorithms for compressed sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 330–341, 2010.

[168] M. Vladimirova, S. Girard, H. Nguyen, and J. Arbel, "Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions," *Stat*, vol. 9, no. 1, e318, 2020.

[169] T. Tao, *Topics in random matrix theory*. American Mathematical Soc., 2012, vol. 132.

[170] D. L. Hanson and F. T. Wright, "A bound on tail probabilities for quadratic forms in independent random variables," *The Annals of Mathematical Statistics*, vol. 42, no. 3, pp. 1079–1083, 1971.

[171] K. Jaganathan, S. Oymak, and B. Hassibi, "Sparse phase retrieval: Uniqueness guarantees and recovery algorithms," *arXiv preprint arXiv:1311.2745*, 2013.

[172] A. Maillard, F. Krzakala, Y. M. Lu, and L. Zdeborová, "Construction of optimal spectral methods in phase retrieval," *arXiv preprint arXiv:2012.04524*, 2020.

[173] M. Haikin, R. Zamir, and M. Gavish, "Random subsets of structured deterministic frames have manova spectra," *Proceedings of the National Academy of Sciences*, vol. 114, no. 26, E5024–E5033, 2017.

[174] J. P. Chen and R. A. Kirian, "Phase retrieval in the presence of multiplicative noise," in *Unconventional and Indirect Imaging, Image Reconstruction, and Wavefront Sensing 2017*, International Society for Optics and Photonics, vol. 10410, 2017, 104100K.

[175] N. Loh, M. J. Bogan, V. Elser, A. Barty, S. Boutet, S. Bajt, J. Hajdu, T. Ekeberg, F. R. Maia, J. Schulz, *et al.*, "Cryptotomography: Reconstructing 3d fourier intensities from randomly oriented single-shot diffraction patterns," *Physical review letters*, vol. 104, no. 22, p. 225 501, 2010.

[176] C.-I. Chang, *Hyperspectral imaging: techniques for spectral detection and classification*. Springer Science & Business Media, 2003, vol. 1.

[177] D. I. Hoult and B. Bhakar, "Nmr signal reception: Virtual photons and coherent spontaneous emission," *Concepts in Magnetic Resonance: An Educational Journal*, vol. 9, no. 5, pp. 277–297, 1997.

[178] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *IEEE signal processing magazine*, vol. 13, no. 3, pp. 43–64, 1996.