Non-Convex Phase Retrieval Algorithms and Performance Analysis

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Gang Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Prof. Georgios B. Giannakis, Advisor

April, 2018

Acknowledgments

There are so many people to whom I wish to acknowledge and thank for making my past four years at the University of Minnesota (UMN) the most rewarding and enlightening journey of my life so far, and for this thesis in particular.

First and foremost, my deepest gratitude goes to my advisor Prof. Georgios B. Giannakis. I am very grateful to him for such an exciting and fruitful experience during my PhD studies. His advice and feedback on the formalization, investigation, and presentation of original research has been extraordinary to me by all means. His invaluable guidance as well as constant encouragement through dedicating extensive amounts of time has not only made me a better researcher, but also a better person. His vision and enthusiasm about innovative research and beyond, his broad and deep knowledge, and his unbounded energy have constantly been a true inspiration for me. He has also provided me with phenomenal environments for conducting research, and (because of him) I have been very fortunate to be always surrounded by other wonderful students and colleagues. This thesis would have not been possible without his support.

I would also like to extend my sincerest appreciation to Profs. Jie Chen and Jian Sun at the Beijing Institute of Technology (BIT), for introducing me to the world of academic research at the very beginning of my graduate studies at BIT, for their true understanding of my concerns throughout the course of my PhD at BIT as well as at UMN, and for many other reasons which cannot be expressed in the space provided here.

Due thanks go to Profs. Mostafa Kaveh, Yousef Saad, and Mehmet Akçakaya for agreeing to serve on my committee as well as all their valuable comments and feedback on my research and thesis. Thanks also go to other professors in the Departments of Electrical Engineering and Computer Science whose graduate level courses helped me build the necessary background to embark on this journey.

During my PhD studies, I had the opportunity to collaborate with several excellent individuals,

and I have greatly benefited from their critical thinking, brilliant ideas, and vision. Particularly, I would like to express my gratitude to Prof. Seung-Jun Kim who was patient enough to train me in the first couple of years at UMN, and Prof. Vassilis Kekatos, with whom we had great research collaboration. I would also like to extend my due credit and warmest thanks to Profs. Mehmet Akçakaya, Antonio J. Conejo (OSU), Yonina C. Eldar (Technion), Yousef Saad, and Nikos D. Sidiropoulos (UVA) for their insightful input to our fruitful collaborations. The material in this thesis has also benefited from discussions with current and former members of the SPiNCOM group at UMN: Dr. Brian Baingana, Prof. Juan-Andrés Bazerque, Dimitris Berberidis, Dr. Jia Chen, Tianyi Chen, Dr. Emiliano Dall'Anese, Vassilis Ioannidis, Georgios V. Karanikolas, Donghoon Lee, Prof. Geert Leus, Bingcong Li, Prof. Qing Ling, Meng Ma, Dr. Morteza Mardani, Prof. Antonio G. Marqués, Prof. Gonzalo Mateos, Dr. Athanasios Nikolakopoulos, Prof. Daniel Romero, Alireza Sadeghi, Fatemeh Sheikholeslami, Yanning Shen, Prof. Konstantinos Slavakis, Panagiotis Traganitis, Dr. Yunlong Wang, Liang Zhang, Prof. Yu Zhang, and Prof. Hao Zhu. I am truly grateful to these people for their continuous help. I would also wish to acknowledge the grants that support financially our research.

I am not forgetting my friends, some of which I have already mentioned above, both the ones here in Minneapolis, and my old friends that are far away in China, in particular: Yongjian Cai, Kexin Guo, Shuai He, Haoji Hu, Kejun Huang, Mingyi Hong, Meng Li, Cheng Qian, Yunmei Shi, Hanghai Tian, Jiaxiang Yan, Bo Yang, and Ahmed S. Zamzam.

Last but not least, I am eternally grateful to my parents, Fengming Wang and Mingju Zhang, who encouraged me and gave me every learning opportunity they could think of. Without you, I would not be standing here today.

Gang Wang, Minneapolis, March 10, 2018.

Dedication

This dissertation is dedicated to my family for their unconditional love and support.

Abstract

High-dimensional signal estimation plays a fundamental role in various science and engineering applications, including optical and medical imaging, wireless communications, and power system monitoring. The ability to devise solution procedures that maintain high computational and statistical efficiency will facilitate increasing the resolution and speed of lensless imaging, identifying artifacts in products intended for military or national security, as well as protecting critical infrastructure including the smart power grid. This thesis contributes in both theory and methods to the fundamental problem of phase retrieval of high-dimensional (sparse) signals from magnitude-only measurements. Our vision is to leverage exciting advances in non-convex optimization and statistical learning to devise algorithmic tools that are simple, scalable, and easy-to-implement, while being computationally and statistically (near-)optimal.

Phase retrieval is approached from a non-convex optimization perspective. To gain statistical and computational efficiency, the magnitude data (instead of the intensities) are fitted based on the least-squares or maximum likelihood criterion, which leads to optimization models that trade off smoothness for 'low-order' non-convexity. To solve the resultant challenging non-convex and non-smooth optimization, the present thesis introduces a two-stage algorithmic framework, that is termed amplitude flow. The amplitude flows start with a careful initialization, which is subsequently refined by a sequence of regularized gradient-type iterations. Both stages are lightweight, and they scale well with problem dimensions. Due to the highly non-convex landscape, judicious gradient regularization techniques such as trimming (i.e., truncation) and iterative reweighting are devised to boost the exact phase recovery performance. It is shown that successive iterates of the amplitude flows provably converge to the global optimum at a geometric rate, corroborating their efficiency in terms of computational, storage, and data resources. The amplitude flows are also demonstrated to be stable vis-à-vis additive noise.

Sparsity plays a critical role in many fields - what has led to the upsurge of research referred to as compressive sampling. In diverse applications, the signal is naturally sparse or admits a sparse representation after some known and deterministic linear transformation. This thesis also accounts for phase retrieval of sparse signals, by putting forth sparsity-cognizant amplitude flow variants. Although analysis, comparisons, and corroborating tests focus on non-convex phase retrieval in this thesis, a succinct overview of other areas is provided to highlight the universality of the novel algorithmic framework to a number of intriguing future research directions.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Phase Retrieval Problem

Detecting visible light and radiation of high frequencies relies on energy intensity measurements of the sought radiating field. The field on the other hand is can be characterized by a complex function, comprising its modulus and phase parameters. The intensity is proportional to the square of the modulus, but the phase information is lost at the energy detector [16].

Consider for example an image formation experiment. Suppose that the field in the object (primary) space is denoted by $\mathcal{E}(p)$. If an image indexed by i is formed, the field in the image space is given by $E_i(p')$. In addition, at a large enough distance from the imaging plane, the complex functions $\mathcal{E}(p)$, and $E_i(p')$ are known to be related through Fourier transform relations, determination of one of which provides the other. In physical scattering experiments however, one can only measure $|E_i(p')|$. In general, the modulus data $|E_i(p')|$ provides only geometrical information concerning the object of interest. To recover the image, namely to determine the structure of the object, one has to recover the missing phases of $E_i(p')$ first, which is known as the phase retrieval problem [55].

To set up the phase retrieval problem mathematically, this thesis focuses on the discretized one-dimensional (1D) setting. Suppose we have an object of interest described by $\boldsymbol{x} \in \mathbb{C}^n$, and that we would like to measure $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ for some known sampling vectors $\boldsymbol{a}_i \in \mathbb{C}^n$, but only have access to the modulus of the linear transformations, namely

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 , \quad i = 1, 2, \ldots, m. \tag{1.1}$$

The goal here is to recover the missing phase of the linear transformations $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$, which is known as the (generalized) phase retrieval problem. In the original phase retrieval problem, the sampling vectors $\boldsymbol{a}_i$ correspond to rows of the discrete Fourier transform (DFT) matrix of suitable dimensions, in which case one is tasked with recovering a signal vector from the modulus of its Fourier transform. It is clear that once the phase becomes available, one can easily find the object $\boldsymbol{x}$ by solving linear equations. Succinctly stated, we are interested in solving systems of quadratic equations of the form

$$\text{find} \quad \boldsymbol{z} \in \mathbb{C}^n \tag{1.2a}$$

$$\text{subject to} \quad y_i = |\langle \boldsymbol{a}_i, \boldsymbol{z} \rangle|^2, \quad i = 1, 2, \ldots, m \tag{1.2b}$$

where $\boldsymbol{z} \in \mathbb{C}^n$ is the decision vector, $\{\boldsymbol{a}_i \in \mathbb{C}^n\}$ are the known sampling vectors, and $\{y_i \in \mathbb{R}\}$ are the observed data.

1.2 Motivation and Context

The phase problem appears in many fields of science and engineering [47]. X-ray crystallography is one such field which, as is well known, led to the discovery of the double helix structure of DNA [55]. Besides X-ray crystallography, other relevant application domains include optics [86], array and high-power coherent diffractive imaging [25], astronomical imaging [46], ptychography [144], and microscopy [85]. Similar problems are also encountered in areas such as acoustics [7], channel estimation in wireless communications [2, 96], computational biology [114], speech processing [1], quantum mechanics [98], and quantum information [38].

It has been shown that reconstructing a discrete, finite-duration signal from its Fourier transform magnitude is generally NP-complete [102]. Even checking quadratic feasibility (i.e., whether a solution to a given quadratic system of the form (1.2) exists or not) is itself an NP-hard problem [57, Thm. 2.6]. Therefore, despite its simple form and practical relevance across various fields, tackling the quadratic system in (1.2) is challenging and NP-hard in general.

The problem in (1.2) constitutes an instance of non-convex quadratic programming. Specifically for real $\boldsymbol{a}_i$ and $\boldsymbol{x}$, problem (1.2) can be understood as a combinatorial optimization since one seeks a series of signs $\{s_i = \pm 1\}_{i=1}^m$, such that the solution to the system of linear equations $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle = s_i \psi_i$, where $\psi_i := \sqrt{y_i}$, obeys the given quadratic system. Evidently, there are a total

of 2^m different combinations of $\{s_i\}_{i=1}^m$, among which only two lead to $\boldsymbol{x}$ up to a global sign. The complex case becomes even more complicated, where instead of a set of signs $\{s_i\}_{i=1}^m$, one must determine a collection of unimodular complex scalars $\{\sigma_i \in \mathbb{C}\}_{i=1}^m$. Special cases with $\boldsymbol{a}_i > \boldsymbol{0}$ (entry-wise inequality), $x_i^2 = 1$, and $y_i = 0$, $1 \leq i \leq m$, correspond to the so-called stone problem [10, Sec. 3.4.1].

1.2.1 Uniqueness of the phase retrieval problem

We first review some results from the literature that give conditions under which the phase retrieval problem (1.2) has a unique solution (up to a global unimodular constant). In this process, the different models for the sampling vectors $\boldsymbol{a}_i$ that have been commonly used will also become clear. To that end, collect all sampling vectors $\boldsymbol{a}_i$ in the $m \times n$ matrix $\boldsymbol{A} := [\boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_m]^{\mathcal{H}}$, and concatenate all modulus squares y_i to form the data vector $\boldsymbol{y} := [y_1 \ \cdots \ y_m]^{\mathcal{T}}$, which collectively lead to the more compact matrix-vector representation $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$, where the modulus operator $|\cdot|$ should be understood entry-wise.

Let us start with the uniqueness of Fourier phase retrieval, in which one takes measurements of the form

$$\boldsymbol{y} = |\boldsymbol{F}\boldsymbol{x}|^2 \tag{1.3}$$

where $\boldsymbol{F} \in \mathbb{C}^{n \times n}$ corresponds to the DFT matrix. It is clear from signal processing that there are trivial ambiguities in the Fourier phase retrieval problem due to e.g.,

i. phase shift ($x[i] \to x[i] \cdot \mathrm{e}^{j\theta}$);

ii. spatial shift ($x[i] \to \overline{x[i + i_0]}$); and

iii. conjugate inversion ($x[i] \to \overline{x[-i]}$)

each or any combination of which conserves Fourier magnitude, namely yields the same modulus data $\boldsymbol{y}$. Beyond trivial ambiguities, it has been shown that unique reconstruction of signals is impossible using Fourier modulus measurements [58]. To see this, assign each magnitude measurement y_i an arbitrary phase θ_i, and a direct inverse Fourier transform would yield a solution that also adheres to the data (1.3). For the Fourier phase retrieval problem to be uniquely solvable (up to trivial ambiguities), a number of approaches have been suggested, a sample of which are outlined next [108, 61].

- *Additional constraints on the signal vector*. One common assumption is that the signal vector $\boldsymbol{x}$ has bounded support. However, this assumption does not overcome the ill-poseness of the 1D Fourier phase retrieval, but it is helpful when it comes to retrieving the phase of 2D or 3D signals. Another structural assumption is that $\boldsymbol{x}$ is sparse in the sense that only a small fraction of entries are nonzero. In this case one may also be able to recover $\boldsymbol{x}$ using only a few Fourier measurements, namely a number $m < n$ of equations as in (1.3) [56]. Another possibility is when nonnegative entries of the signal vector are constrained to be nonnegative, which clearly holds true in imaging applications [47, 107].

- *Oversampled DFT*. An alternative is to have additional redundant measurements. A common strategy is to sample the signal in the continuous domain at a finer scale. That is, instead of using frequencies $\omega_i = 2\pi i/n$, one can use $\omega_i' = 2\pi i/m$ for $i = 1, 2, \ldots, m$, and $m \geq n$. Using the argument that the autocorrelation function of any signal is twice the size of the signal itself, it has been proved that phases of real signals can be uniquely retrieved from the Fourier measurements if and only if one oversamples by a factor of two [85].

Let us now turn our attention to the generalized phase retrieval, where one considers general measurement vectors $\boldsymbol{a}_i$ that do not necessarily correspond to rows of DFT matrices. To this end, let us first introduce the notion of generic measurements. If $\boldsymbol{A}$ corresponds to a point in a non-empty Zariski open subset of $\mathbb{C}^{m\times n}/\mathbb{R}^{m\times n}$, the measurements $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2$ are said to be generic [37]. When both $\boldsymbol{x}$ and $\boldsymbol{a}_i$ are real, it has been shown in [8] that $m = 2n - 1$ generic measurements are necessary for ensuring injectivity of the mapping $\boldsymbol{z} \to |\boldsymbol{A}\boldsymbol{z}|^2$, and $m \geq 2n-1$ generic measurements are also sufficient for injectivity. For example, $m \geq 2n - 1$ Gaussian random measurements are injective with probability one. In the complex case, a line of research has established that the minimum number of generic measurements required for injectivity is smaller than $4n - 4$, and likewise, when $m \geq 4n - 4$ generic measurements are available, the mapping is injective with probability one [37].

1.2.2 Algorithmic developments

Adopting the maximum likelihood criterion, the task of recovering $\boldsymbol{x}$ from data y_i observed in additive white Gaussian noise (AWGN) can be recast as that of minimizing the intensity-based

empirical loss [22, 44]

$$\underset{\boldsymbol{z}\in\mathbb{C}^n}{\text{minimize}}\ \ f(\boldsymbol{z}) := \frac{1}{2m}\sum_{i=1}^{m}\left(y_i - |\langle \boldsymbol{a}_i, \boldsymbol{z}\rangle|^2\right)^2. \tag{1.4}$$

An alternative is to consider the Poisson likelihood [32]

$$\underset{\boldsymbol{z}\in\mathbb{C}^n}{\text{minimize}}\ \ h(\boldsymbol{z}) := -\frac{1}{2m}\sum_{i=1}^{m} y_i \log(|\langle \boldsymbol{a}_i, \boldsymbol{z}\rangle|^2) + |\langle \boldsymbol{a}_i, \boldsymbol{z}\rangle|^2. \tag{1.5}$$

Unfortunately, both (1.4) and (1.5) are non-convex. Minimizing non-convex cost functions, which may exhibit many stationary points, is in general NP-hard [89]. In fact, even checking whether a given point is a local minimum or establishing convergence to a local minimum turns out to be NP-complete [89].

Previous approaches to solving (1.4) or (1.5) fall under two categories: non-convex and convex ones. Popular non-convex solvers include alternating projection such as Gerchberg-Saxton [48] and Fineup [47], AltMinPhase [91], (Truncated) Wirtinger flow (WF/TWF) [22, 32, 111], and trust-region methods [117]. Convex counterparts, on the other hand, rely on the so-called matrix-lifting technique or Shor's semidefinite relaxation to obtain the solvers abbreviated as PhaseLift [20], PhaseCut [121], and CoRK [59].

Another line of convex relaxation [51, 49, 6, 54, 41] reformulates phase retrieval as a sparse signal recovery problem, and solves a linear program in the natural parameter vector domain. Although exact signal recovery can be established assuming an accurate enough anchor vector, its empirical performance is often not competitive with state-of-the-art phase retrieval approaches. Additional related approaches can be found in [59, 34, 146, 27, 11, 77, 5, 43, 141, 83, 33, 95, 63, 120, 40, 82, 118, 11, 31, 150, 103, 78, 87]. Interested readers can also access Matlab implementations in [26] for a sample of the aforementioned as well as our proposed phase retrieval solvers.

In terms of sample complexity, it has been proven that[1] $\mathcal{O}(n)$ noise-free measurements suffice for uniquely determining a general signal [44]. It is also self-evident that recovering a general n-dimensional $\boldsymbol{x}$ requires at least $\mathcal{O}(n)$ measurements. Convex approaches enable exact recovery from the optimal bound $\mathcal{O}(n)$ of noiseless measurements [20]; they are based

[1]The notation $\phi(n) = \mathcal{O}(g(n))$ or $\phi(n) \gtrsim g(n)$ (respectively, $\phi(n) \lesssim g(n)$) means there exists a numerical constant $c > 0$ such that $\phi(n) \leq cg(n)$, while $\phi(n) \asymp g(n)$ means $\phi(n)$ and $g(n)$ are orderwise equivalent.

on solving a semidefinite program with a matrix variable of size $n \times n$, thus incurring worst-case computational complexity on the order of $\mathcal{O}(n^{4.5})$ [121] that does not scale well with n. Upon exploiting the problem structure, $\mathcal{O}(n^{4.5})$ can be reduced to $\mathcal{O}(n^3)$ [121]. Solving for vector variables, non-convex approaches on the other hand, enjoy significantly improved computational performance. Adopting a spectral initialization commonly employed in matrix completion [69], AltMinPhase establishes exact recovery with sample complexity $\mathcal{O}(n \log^3 n)$ with resampling [91].

The WF iteratively refines the spectral initial estimate by means of gradient-type updates, which can be approximately interpreted as a variant of stochastic gradient descent [22], [111]. The follow-up TWF improves upon WF through a truncation procedure to separate gradient components of excessively extreme (large or small) sizes. Likewise, due to the heavy tails present in the initialization stage, data $\{y_i\}_{i=1}^m$ are pre-screened to yield improved initial estimates in the so-termed truncated spectral initialization method [32]. The WF allows exact recovery from $\mathcal{O}(n \log n)$ measurements in $\mathcal{O}(mn^2 \log(1/\epsilon))$ time/flops to yield an ϵ-accurate solution for any given $\epsilon > 0$ [22], while TWF advances these to $\mathcal{O}(n)$ measurements and $\mathcal{O}(mn \log(1/\epsilon))$ time [32]. It is also worth mentioning that when $m \geq Cn \log^3 n$ for some sufficiently large positive constant C, the objective function in (1.4) has been shown to admit benign geometric structure that allows certain iterative algorithms (e.g., trust-region methods) to efficiently find a global minimizer with random initializations [117].

1.2.3 Applications of phase retrieval

In this subsection, we discuss two applications of phase retrieval. The goal here is to demonstrate how the Fourier phase retrieval emerges naturally in diverse imaging applications, and how the generalized phase retrieval (with general measurements) can be instrumental to real-world applications.

Optical imaging

In optical imaging, researchers constantly wish to increase resolution so that details can be imaged at a finer scale. Eventually, successful imaging of large protein complexes and biological specimens would further advance the understanding of bio-chemical activities at the molecular level. Phaseless imaging has also become critical in emerging national security applications

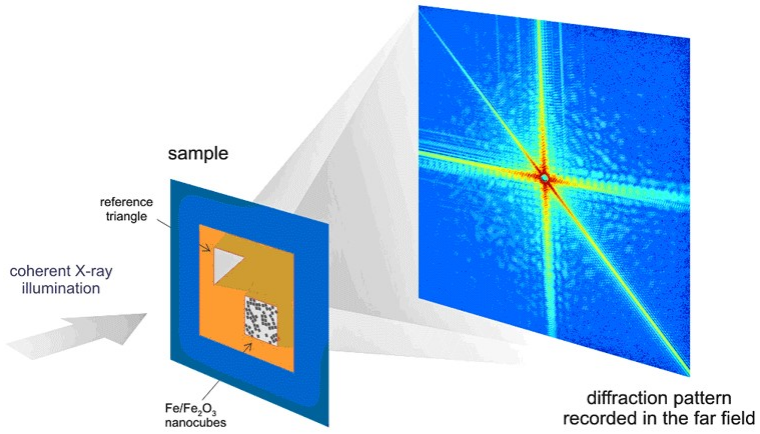

Figure 1.1: Schematic diagram of the experimental setup for coherent diffraction imaging: A coherent wave diffracts from a sample of Fe/Fe_2O_3, and generates a far-field diffraction pattern which corresponds to the modulus of the Fourier transform of the sample.

such as monitoring electronic products intended for military or infrastructure use, where the goal is to ensure such products are clear of secret backdoors granting foreign governments cyber access to vital US infrastructure. The limitation of lens-type imaging devices is that the required optical components such as mirrors and lenses are in general very difficult to construct at short wavelengths, so increasing resolution becomes rather difficult for lens-type imaging devices. On the other hand, phaseless imaging that is based on the Fourier phase retrieval offers a promising alternative for recovery of phase structure because it does not require such optical components. Below, we briefly overview the history of algorithmic phase retrieval in optical imaging.

The use of algorithmic phase retrieval was first suggested in X-ray diffraction microscopy by Sayre in a seminal contribution back in 1952 [104]. Later in 1978, Fienup developed an approach to empirically recover the phase information of 2D images from the modulus of their Fourier transform leveraging prior knowledge such as non-negativity and known support of signal values [47]. The revival of algorithmic phase retrieval around 2000 was due mainly to the successful experimental recording and reconstruction of a diffraction pattern of a non-crystalline object using the so-called coherent diffraction imaging [84]. Figure 1.1 depicts an illustrative experimental setup[2], where the diffraction patterns are (proportional to) the intensities of the Fourier transform of the object [108]. There techniques have recently been employed to image

[2]This figure is adapted from `https://www6.slac.stanford.edu/`.

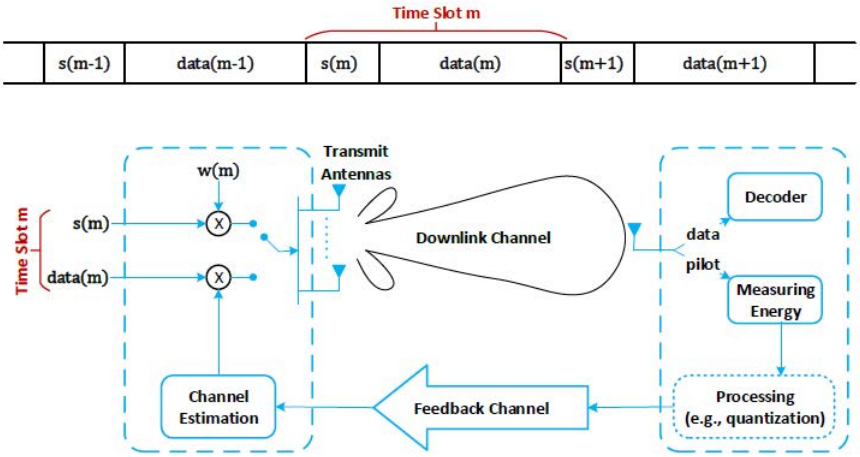

Figure 1.2: The time-slotted frame diagram of the RSS/CQI feedback system adapted from [96].

small particles in diverse applications, including nano-particles, biomaterials, specimen, and electronic circuits [108, 112].

Wireless channel estimation

Another interesting application of algorithmic phase retrieval lies in the estimation of a communication channel. Channel estimation is a critical task in any wireless communication system, and 5G massive multiple-input multiple-output (MIMO) systems, in particular [72] - because the receiver must estimate and feed back to the transmitter a high-dimensional multiple-input single-output (MISO) channel vector for each receiving antenna element, posing critical challenges in terms of mobile computation and power resources, as well as communication overhead [96, 143]. To compensate for temporal channel variations, existing and emerging wireless communication systems provide access to basic received signal strength (RSS)/channel quality indicator (CQI) feedback, which has prompted recent works to address the channel estimation problem using RSS/CQI feedback only [96].

Such a channel estimation setup is depicted in Fig. 1.2 for a time-slotted frame structure of the RSS/CQI feedback system. Per time slot k, the transmitter sends a constant-modulus symbol $s(k) \in \mathbb{C}$ to the receiver with beamforming vector $\boldsymbol{w}(k) \in \mathbb{C}^n$, and subsequently transmits data symbols using different beamforming vectors. The MISO channel vector between the n transmitting antennas and the single receiving antenna at time slot k is denoted by the vector

$\boldsymbol{h}(k) \in \mathbb{C}^n$. Although time-varying in general, for ease of exposition we shall consider static channels over several slots, which means $\boldsymbol{h}(k) = \boldsymbol{h}$ for $k = 1, 2, \ldots, m$. Therefore, the received signal corresponding to $s(k)$ in the presence of noise, is given by

$$z(k) = \boldsymbol{w}^{\mathcal{H}}(k)\boldsymbol{h}s(k) + \eta(k) \tag{1.6}$$

where $\eta(k) \in \mathbb{C}$ models the additive Gaussian white noise. The goal of channel estimation is to estimate (track in the case of time-varying channels) the channel vector $\boldsymbol{h}$ from a collection of RSS measurements $\{|z(k)|^2\}_{k=1}^m$. Again, we arrive at an instance of the phase retrieval problem stated as follows [96]

$$\underset{\boldsymbol{h}\in\mathbb{C}^n}{\text{minimize}} \ \frac{1}{2m}\sum_{k=1}^{m}\left(\psi(k) - \left|\boldsymbol{w}^{\mathcal{H}}(k)\boldsymbol{h}\right|\right)^2 \tag{1.7}$$

with $\psi(k) := |z(k)|$ for all k.

Interestingly, in wireless communications, the choice of transmit beamforming vectors $\{\boldsymbol{w}(k)\}$ is completely up to the communication system designer. In other words, the transmitter has the freedom of designing the beamforming vectors to suit its own purposes [96]. One choice is the random Gaussian design, in which $\{\boldsymbol{w}(k)\}$ are (pseudo)-random Gaussian vectors that are i.i.d. across time and space.

1.3 Thesis Outline and Contributions

The research in this thesis contributes to the advancement of phase retrieval theory and methods, by putting forth a comprehensive algorithmic framework of amplitude flows. The amplitude data are leveraged to form non-convex optimization models, which are different from the intensity-based ones in the existing literature. Our models trade off smoothness for 'low-order' non-convexity, but also smoothness for for (near-)optimal computational and statistical efficiency. To solve the resultant non-convex and non-smooth optimization, the present thesis develops a number of algorithms that provably reconstruct the signal vector using an optimal-order number of Gaussian random measurements, and incur also minimal computational complexity. Algorithms of lower per-iteration complexity are realized based on stochastic non-convex optimization techniques. Phase retrieval theory and methods for sparse signals are also presented. The major difference between the work in this thesis and related non-convex phase retrieval approaches

[91, 22, 32] is the use of the amplitude based least-squares cost function as well as novel gradient regularization techniques. Beyond phase retrieval, the amplitude flow algorithms have potential to tackle tasks such as matrix sensing, blind deconvolution, noncoherent channel estimation, power system state estimation, and deep learning.

Chapter 2 starts with the amplitude-based least-squares formulation, for which the amplitude flows are introduced. Specifically, a novel initialization procedure that is termed orthogonality-promoting initialization is first presented, followed by the truncated amplitude flow iterations for refining the initialization. The developed gradient regularization (truncation or trimming) helps improve the exact recovery performance considerably, which may be of independent interest to related non-convex optimization tasks. On the theoretical side, we establish that for Gaussian random designs, the proposed (truncated) amplitude flow algorithms recover the true signals exactly with high probability. This holds true with no assumption on the signal, and as soon as the number of measurements is larger than some constant times the number of unknowns, namely $m \geq c_0 n$ for some large enough constant $c_0 > 0$.

Relative to existing non-convex phase retrieval approaches [91, 22, 32], leveraging the amplitude data for loss minimization and carefully designed gradient regularization is among the main contributions of this present thesis, and Chapter 2 in particular. Extensive numerical experiments using computer generated data and real images are presented to corroborate the merits of the proposed amplitude flow approaches and validate the associated theoretical claims. The material of Chapter 2 has been reported in [124, 129, 130, 131].

In the amplitude flow algorithm in Chapter 2, all data samples are retained after regularization, and treated equivalently in terms of how they affect initialization and the corresponding search direction. This may lead to suboptimal performance since data samples in the context of non-convex optimization may behave differently. Building on this observation, a novel iterative reweighting technique is invoked in the amplitude flow algorithm, which leads to our so-called reweighted amplitude flow algorithm in Chapter 3. Substantial tests are presented to corroborate the merits of the reweighted amplitude flow algorithm.

To improve the scalability of phase retrieval approaches, Chapter 4 contributes computationally efficient phase retrieval solvers, that are well suited for large-scale imaging tasks typically in the order of millions. Instead of resorting to the gradient-type approaches in Chapters 2 and 3, Chapter 4 advocates inexpensive stochastic iterations based on non-convex schemes for phase retrieval of high-dimensional signals. Specifically, scalable solvers are put forth for carrying out

both the initialization and the refinement stages, each of which only incurs near optimal-order per-iteration complexity. Furthermore, Kaczmarz variants that have been studied in the context of phase retrieval [140] turn out to be special cases of our stochastic amplitude flow algorithms. We also prove that the stochastic non-convex optimization schemes also reconstruct the true signals exactly, and exponentially fast when $m \geq c_0 n$. Our theory also establishes the exact recovery of Kaczmarz variants. Simulated tests using synthetic data and real images corroborate the scalability and effectiveness of the proposed stochastic non-convex optimization procedures in phase retrieval of signals of millions of unknowns relative to competing approaches. The results of Chapter 4 have been reported in [126, 127].

In real-world applications, especially those related to imaging, the signals are naturally sparse or admit a sparse representation after some known and deterministic linear transformation. Chapter 5 deals with phase retrieval of such sparse signals. To generalize the orthogonality-promoting initialization of Chapter 2 to phase retrieval of sparse signals, a novel method is developed first for estimating the support of the sparse signal, and it is followed by power iterations solving an eigenvalue problem over the dimensionality-reduced data. To enable sparse recovery, hard-thresholding based gradient iterations are invoked for the amplitude based least-squares formulation. We demonstrate that when the number of measurements becomes large enough, the support of the underlying signal can be estimated exactly, relying on which exact reconstruction of the underlying sparse signal is also guaranteed with high probability. The results of Chapter 5 are included in [135, 128, 149, 136].

The thesis is summarized, and interesting open problems are included in Chapter 6.

1.4 Notational Conventions

The following notation will be used throughout the subsequent chapters. Lower- (upper-) case boldface letters denote vectors (matrices). Calligraphic letters are reserved for sets, e.g., $\mathcal{S}$. Symbol $\mathcal{T}$ ($\mathcal{H}$) as superscript stands for matrix/vector transposition (conjugate transposition). For vectors, $\|\cdot\|_2$ or $\|\cdot\|$ represents the Euclidean norm, while $\|\cdot\|_0$ denotes the ℓ_0 pseudo-norm counting the number of nonzero entries. The floor (ceiling) operation $\lfloor c \rfloor$ ($\lceil c \rceil$) denotes the largest integer no greater (the smallest integer but no smaller) than the given number $c > 0$; $|\mathcal{S}|$ counts the number of entries in $\mathcal{S}$. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the vector Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; and the Gauss error function $\mathrm{erf}(x)$ is defined as $\mathrm{erf}(x) :=$

$(1/\sqrt{\pi}) \int_{-x}^{x} e^{-\tilde{x}^2} d\tilde{x}$. For any integer $m > 0$, we use $[m]$ to denote the set $\{1, 2, \ldots, m\}$. Finally, the symbol $\succeq$ means the positive semi-definiteness of matrices, while the ordered eigenvalues of matrix $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ are given as $\lambda_1(\boldsymbol{X}) \geq \lambda_2(\boldsymbol{X}) \geq \cdots \geq \lambda_n(\boldsymbol{X})$.

Chapter 2

Phase Retrieval via Amplitude Flow

Consider a system of m quadratic equations

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2, \quad i = 1, 2, \ldots, m \tag{2.1}$$

where the data vector $\boldsymbol{y} := [y_1 \ \cdots \ y_m]^{\mathcal{T}}$ and feature vectors $\boldsymbol{a}_i \in \mathbb{R}^n$ or $\mathbb{C}^n$ are known, whereas the vector $\boldsymbol{x} \in \mathbb{R}^n$ or $\mathbb{C}^n$ is the wanted unknown. When $\boldsymbol{a}_i$ and/or $\boldsymbol{x}$ are complex-valued, the magnitudes of their inner-products $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ are given but phase information is lacking; in the real case, only the signs of $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ are unknown. Assuming that the system of quadratic equations in (2.1) admits a unique solution $\boldsymbol{x}$ (up to a global unimodular constant), our goal here is to recover $\boldsymbol{x}$ from m quadratic equations, or equivalently, to recover the missing phases of $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$.

2.1 Non-convex Optimization Models

Different than existing models in (1.4) and (1.5) that are based on the intensity data, this thesis considers directly the following amplitude data-based empirical loss minimization

$$\underset{\boldsymbol{z} \in \mathbb{C}^n}{\text{minimize}} \quad \ell(\boldsymbol{z}) := \frac{1}{2m} \sum_{i=1}^{m} \big(\psi_i - |\boldsymbol{a}_i^{\mathcal{H}} \boldsymbol{z}|\big)^2 \tag{2.2}$$

which is not only non-convex, but also non-smooth due to the modulus terms.

Along the lines of suitably initialized non-convex procedures [22, 32], the present chapter develops a linear-time (i.e., the computational time linearly in both dimensions m and n)

algorithm, referred to as *truncated amplitude flow* (TAF). Our approach provably recovers an n-dimensional unknown signal $\boldsymbol{x}$ exactly from a near-optimal number of noiseless random measurements, while also featuring near-perfect statistical performance in the noisy setting. TAF operates in two stages: In the first stage, we introduce an orthogonality-promoting initialization that is computable with a few power iterations. Stage two refines the initial estimate by successive updates of truncated generalized gradient iterations.

Our initialization is built upon the hidden orthogonality characteristics of high-dimensional random vectors [17], which is in contrast to spectral alternatives originating from the strong law of large numbers (SLLN) [91, 22, 32]. Furthermore, one challenge of phase retrieval lies in reconstructing the signs/phases of $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$ in the real/complex settings. Our TAF's refinement stage leverages a simple yet effective regularization rule to eliminate the erroneously estimated phases in the generalized gradient components with high probability. Numerical experiments corroborate that the proposed initialization returns more accurate and robust initial estimates than its spectral counterparts in the noiseless and noisy settings. In addition, our TAF (with gradient truncation) markedly improves upon its "plain-vanilla" version AF.

2.2 Truncated Amplitude Flow

In this section, the two stages of TAF are presented. First, the challenge of handling the non-convex and non-smooth amplitude-based cost function is analyzed, and it is addressed by a carefully designed gradient regularization rule. Limitations of (truncated) spectral initializations are then pointed out, followed by a simple motivating example to inspire our orthogonality-promoting initialization method. For concreteness, the analysis will focus on the real Gaussian model with $\boldsymbol{x} \in \mathbb{R}^n$, and independent and identically distributed (i.i.d.) sampling vectors $\boldsymbol{a}_i \in \mathbb{R}^n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Numerical experiments using the complex Gaussian model with $\boldsymbol{x} \in \mathbb{C}^n$, and i.i.d. $\boldsymbol{a}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n) := \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n/2) + j\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n/2)$ will be discussed briefly.

To start, let us define the Euclidean distance of any estimate $\boldsymbol{z}$ to the solution set: $\mathrm{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min\{\|\boldsymbol{z}+\boldsymbol{x}\|, \|\boldsymbol{z}-\boldsymbol{x}\|\}$ for real signals, and $\mathrm{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min_{\phi\in[0,2\pi)} \|\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{i\phi}\|$ for complex ones [22]. Define also the indistinguishable global phase constant in the real setting as

$$\phi(\boldsymbol{z}) := \begin{cases} 0, & \|\boldsymbol{z}-\boldsymbol{x}\| \le \|\boldsymbol{z}+\boldsymbol{x}\|, \\ \pi, & \text{otherwise.} \end{cases} \tag{2.3}$$

Henceforth, fixing $\boldsymbol{x}$ to be any solution of the quadratic system (2.1), we always assume that $\phi(\boldsymbol{z}) = 0$; otherwise, $\boldsymbol{z}$ is replaced by $\mathrm{e}^{-j\phi(\boldsymbol{z})}\boldsymbol{z}$, but for simplicity of presentation, the constant phase adaptation term $\mathrm{e}^{-j\phi(\boldsymbol{z})}$ will be dropped whenever it is clear from the context.

2.2.1 Truncated gradient iterations

For brevity, collect all amplitudes $\{\psi_i\}_{i=1}^m$ to form the data vector $\boldsymbol{\psi} := [\psi_1 \ \cdots \ \psi_m]^{\mathcal{T}}$. One can rewrite the amplitude-based cost function in matrix-vector representation as

$$\underset{\boldsymbol{z}\in\mathbb{R}^n}{\text{minimize}} \ \ell(\boldsymbol{z}) := \frac{1}{m}\sum_{i=1}^{m}\ell_i(\boldsymbol{z}) = \frac{1}{2m}\big\|\boldsymbol{\psi} - |\boldsymbol{A}\boldsymbol{z}|\,\big\|^2 \tag{2.4}$$

where $\ell_i(\boldsymbol{z}) := \frac{1}{2}(\psi_i - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|)^2$, and with a slight abuse of notation, $|\boldsymbol{A}\boldsymbol{z}| := [|\boldsymbol{a}_1^{\mathcal{T}}\boldsymbol{z}| \ \cdots \ |\boldsymbol{a}_m^{\mathcal{T}}\boldsymbol{z}|]^{\mathcal{T}}$. Apart from being non-convex, another challenging aspect of the amplitude-based loss function $\ell(\boldsymbol{z})$ is that it is non-differentiable, and that it is not clear how to run gradient-type algorithms.

In the presence of smoothness or convexity, convergence analysis of iterative algorithms relies either on continuity of the gradient (ordinary gradient methods) [109], or, on the convexity of the objective functional (subgradient methods) [99]. Although subgradient methods have found widespread applicability in non-smooth optimization, they are limited to the class of convex functions [110, Page 4]. Nevertheless, as the loss function is differentiable except for at isolated points rendering each or some of the least-squares zero; and one can use the notion of generalized gradients in such non-convex and non-smooth optimization settings, which define the gradient at a non-differentiable point as one of the limit points of the gradient in a local neighborhood of the non-differentiable point, and considerably broadens the scope of the (sub)gradient to the class of almost everywhere differentiable functions [36].

Formally, consider a continuous but not necessarily differentiable function $h(\boldsymbol{z}) \in \mathbb{R}$ defined over an open region $\mathcal{S} \subseteq \mathbb{R}^n$. We then have the following definition.

Definition 1. *[35, Def. 1.1] The generalized gradient of a function h at $\boldsymbol{z}$, denoted by ∂h, is the convex hull of the set of limits of the form $\lim \nabla h(\boldsymbol{z}^k)$, where $\boldsymbol{z}^k \to \boldsymbol{z}$ as $k \to +\infty$, i.e.,*

$$\partial h(\boldsymbol{z}) := \mathrm{conv}\Big\{\lim_{k\to+\infty}\nabla h(\boldsymbol{z}^k) : \boldsymbol{z}^k \to \boldsymbol{z},\ \boldsymbol{z}^k \notin \mathcal{G}_\ell\Big\}$$

*where the symbol '*conv*' signifies the convex hull of a set, and $\mathcal{G}_\ell$ denotes the set of points in $\mathcal{S}$ at which h fails to be differentiable.*

Having introduced the notion of a generalized gradient, and with t denoting the iteration count, our approach to solving (2.4) amounts to iteratively refining the initial guess $\boldsymbol{z}^0$ (returned by the orthogonality-promoting initialization method to be detailed shortly) by means of the ensuing *truncated* generalized gradient iterations

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \mu^t \, \partial \ell_{\rm tr}(\boldsymbol{z}^t). \tag{2.5}$$

Here, $\mu^t > 0$ is the step size, and the (truncated) generalized gradient $\partial \ell_{\rm tr}(\boldsymbol{z}^t)$ is given by

$$\partial \ell_{\rm tr}(\boldsymbol{z}^t) := \frac{1}{m} \sum_{i \in \mathcal{I}^{t+1}} \left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t - \psi_i \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t|} \right) \boldsymbol{a}_i \tag{2.6}$$

for some index set $\mathcal{I}^{t+1} \subseteq \{1, 2, \ldots, m\}$ to be designed next. The convention $\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t|} := 0$ is adopted, if $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t = 0$. It is easy to verify that the update in (2.5) with a full generalized gradient in (2.6) monotonically decreases the objective function value in (2.4).

Any stationary point $\boldsymbol{z}^*$ of $\ell(\boldsymbol{z})$ can be characterized by the following fixed-point equation

$$\boldsymbol{A}^{\mathcal{T}} \left(\boldsymbol{A}\boldsymbol{z}^* - \boldsymbol{\psi} \odot \frac{\boldsymbol{A}\boldsymbol{z}^*}{|\boldsymbol{A}\boldsymbol{z}^*|} \right) = \boldsymbol{0} \tag{2.7}$$

for entry-wise product $\odot$, which may have many solutions. Clearly, if $\boldsymbol{z}^*$ is a solution, then so is $-\boldsymbol{z}^*$. Furthermore, both solutions/global minimizers $\boldsymbol{x}$ and $-\boldsymbol{x}$ satisfy (2.7) due to the fact that

$$\boldsymbol{A}\boldsymbol{x} - \boldsymbol{\psi} \odot \frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} = \boldsymbol{0}.$$

Considering any stationary point $\boldsymbol{z}^* \neq \pm \boldsymbol{x}$ that has been adapted such that $\phi(\boldsymbol{z}^*) = 0$, write

$$\boldsymbol{z}^* = \boldsymbol{x} + (\boldsymbol{A}^{\mathcal{T}} \boldsymbol{A})^{-1} \boldsymbol{A}^{\mathcal{T}} \left[\boldsymbol{\psi} \odot \left(\tfrac{\boldsymbol{A}\boldsymbol{z}^*}{|\boldsymbol{A}\boldsymbol{z}^*|} - \tfrac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|} \right) \right]. \tag{2.8}$$

Thus, a necessary condition for $\boldsymbol{z}^* \neq \boldsymbol{x}$ in (2.8) is $\frac{\boldsymbol{A}\boldsymbol{z}^*}{|\boldsymbol{A}\boldsymbol{z}^*|} \neq \frac{\boldsymbol{A}\boldsymbol{x}}{|\boldsymbol{A}\boldsymbol{x}|}$. Expressed differently, there must be sign differences between $\boldsymbol{A}\boldsymbol{z}^*$ and $\boldsymbol{A}\boldsymbol{x}$ whenever one gets stuck with an undesirable stationary point $\boldsymbol{z}^*$. Inspired by this observation, it is reasonable to devise solvers that can detect and separate out the generalized gradient components corresponding to mistakenly estimated signs $\left\{ \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t|} \right\}$ along the iterates $\{\boldsymbol{z}^t\}$.

Precisely, if $\boldsymbol{z}^t$ and $\boldsymbol{x}$ lie at different sides of the hyperplane $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z} = 0$, then the sign of $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t$

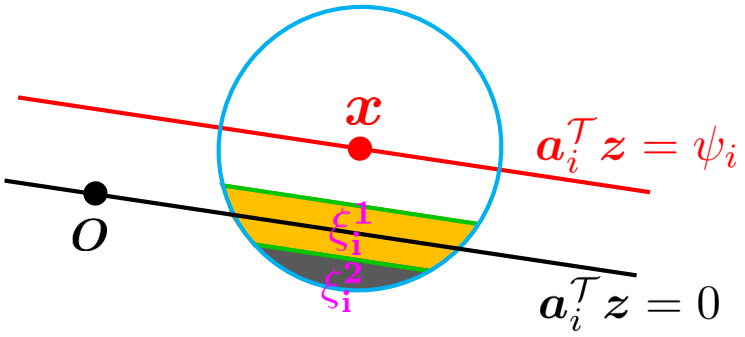

Figure 2.1: Geometric description of the proposed truncation rule on the i-th gradient component involving $\boldsymbol{a}_i^T \boldsymbol{x} = \psi_i$, where the red dot denotes the solution $\boldsymbol{x}$ and the black one is the origin. Hyperplanes $\boldsymbol{a}_i^T \boldsymbol{z} = \psi_i$ and $\boldsymbol{a}_i^T \boldsymbol{z} = 0$ (of $\boldsymbol{z} \in \mathbb{R}^n$) passing through points $\boldsymbol{z} = \boldsymbol{x}$ and $\boldsymbol{z} = \boldsymbol{0}$, respectively, are shown.

will be different than that of $\boldsymbol{a}_i^T \boldsymbol{x}$; that is, $\frac{\boldsymbol{a}_i^T \boldsymbol{x}}{|\boldsymbol{a}_i^T \boldsymbol{x}|} \neq \frac{\boldsymbol{a}_i^T \boldsymbol{z}}{|\boldsymbol{a}_i^T \boldsymbol{z}|}$. Specifically, one can re-write the i-th generalized gradient component as

$$\begin{aligned}
\partial \ell_i(\boldsymbol{z}) &= \Big(\boldsymbol{a}_i^T \boldsymbol{z} - \psi_i \frac{\boldsymbol{a}_i^T \boldsymbol{z}}{|\boldsymbol{a}_i^T \boldsymbol{z}|}\Big)\boldsymbol{a}_i \\
&= \Big(\boldsymbol{a}_i^T \boldsymbol{z} - |\boldsymbol{a}_i^T \boldsymbol{x}| \frac{\boldsymbol{a}_i^T \boldsymbol{x}}{|\boldsymbol{a}_i^T \boldsymbol{x}|}\Big)\boldsymbol{a}_i + \Big(\frac{\boldsymbol{a}_i^T \boldsymbol{x}}{|\boldsymbol{a}_i^T \boldsymbol{x}|} - \frac{\boldsymbol{a}_i^T \boldsymbol{z}}{|\boldsymbol{a}_i^T \boldsymbol{z}|}\Big)\psi_i \boldsymbol{a}_i \\
&= \boldsymbol{a}_i \boldsymbol{a}_i^T (\boldsymbol{z} - \boldsymbol{x}) + \Big(\frac{\boldsymbol{a}_i^T \boldsymbol{x}}{|\boldsymbol{a}_i^T \boldsymbol{x}|} - \frac{\boldsymbol{a}_i^T \boldsymbol{z}}{|\boldsymbol{a}_i^T \boldsymbol{z}|}\Big)\psi_i \boldsymbol{a}_i \\
&= \boldsymbol{a}_i \boldsymbol{a}_i^T \boldsymbol{h} + \underbrace{\Big(\frac{\boldsymbol{a}_i^T \boldsymbol{x}}{|\boldsymbol{a}_i^T \boldsymbol{x}|} - \frac{\boldsymbol{a}_i^T \boldsymbol{z}}{|\boldsymbol{a}_i^T \boldsymbol{z}|}\Big)\psi_i \boldsymbol{a}_i}_{\triangleq \boldsymbol{r}_i},
\end{aligned} \tag{2.9}$$

where $\boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$. Intuitively, the SLLN asserts that averaging the first term $\boldsymbol{a}_i \boldsymbol{a}_i^T \boldsymbol{h}$ over m instances approaches $\boldsymbol{h}$, which qualifies it as a desirable search direction. However, certain generalized gradient entries involve erroneously estimated signs of $\boldsymbol{a}_i^T \boldsymbol{x}$; hence, nonzero $\boldsymbol{r}_i$ terms exert a negative influence on the search direction $\boldsymbol{h}$ by dragging the iterate away from $\boldsymbol{x}$, and they typically have sizable magnitudes as will be further elaborated in Rmk. 2 shortly.

Figure 2.1 demonstrates this from a geometric perspective, where the black dot is the origin,

and the red dot the solution $\boldsymbol{x}$; here, $-\boldsymbol{x}$ is omitted for ease of exposition. Assume without loss of generality that the i-th missing sign is positive, i.e., $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x} = \psi_i$. As will be demonstrated in Thm. 2, with high probability, the initial estimate returned by our orthogonality-promoting method obeys $\|\boldsymbol{h}\| \le \rho\|\boldsymbol{x}\|$ for some sufficiently small constant $\rho > 0$. Therefore, all points lying on or within the circle (or sphere in high-dimensional spaces) in Fig. 2.1 satisfy $\|\boldsymbol{h}\| \le \rho\|\boldsymbol{x}\|$. If $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} = 0$ does not intersect with the circle, then all points within the circle satisfy $\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|} = \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}$ qualifying the i-th generalized gradient as a desirable search (descent) direction in (2.9). If, on the other hand, $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} = 0$ intersects the circle, then points lying on the same side of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} = 0$ with $\boldsymbol{x}$ in Fig. 2.1 admit correctly estimated signs, while points lying on different sides of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} = 0$ with $\boldsymbol{x}$ would have $\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|} \neq \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}$. This gives rise to a corrupted search direction in (2.9), implying that the corresponding generalized gradient component should be eliminated.

However, it is difficult or even impossible to check whether the sign of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t$ equals that of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}$. Fortunately, as demonstrated in Fig. 2.1, most spurious generalized gradient components (those corrupted by nonzero $\boldsymbol{r}_i$ terms) hover around the watershed hyperplane $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t = 0$. For this reason, TAF includes only components having $\boldsymbol{z}^t$ sufficiently away from its watershed, i.e.,

$$\mathcal{I}^{t+1} := \left\{ 1 \le i \le m \,\middle|\, \frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|} \ge \frac{1}{1+\gamma} \right\}, \quad t \ge 0 \tag{2.10}$$

for an appropriately selected threshold $\gamma > 0$. To be specific, the light yellow color-coded area denoted by ξ_i^1 in Fig. 2.1 signifies the truncation region of $\boldsymbol{z}$: if $\boldsymbol{z} \in \xi_i^1$ satisfies the condition in (2.10), then the corresponding generalized gradient component $\partial\ell_i(\boldsymbol{z}; \psi_i)$ will be thrown out. However, the truncation rule may mis-reject certain 'good' gradients if $\boldsymbol{z}^t$ lies in the upper part of ξ_i^1; 'bad' gradients may be missed as well if $\boldsymbol{z}^t$ belongs to the spherical cap ξ_i^2. Fortunately, as we will show in Lemmas 5 and 6, the probabilities of misses and mis-rejections are provably very small, hence precluding a noticeable influence on the descent direction. Although not perfect, it turns out that such a regularization rule succeeds in detecting and eliminating most corrupted generalized gradient components with high probability, therefore maintaining a well-behaved search direction. Further from our numerical experiments, the developed truncation procedure turns out to be useful in avoiding spurious stationary points in the context of nonconvex optimization, as will be justified in Sec. 2.4 by the numerical comparison between our amplitude flow (AF) algorithms with or without the judiciously designed truncation rule. Interestingly, similar ideas including censoring have been developed for large-scale linear

regressions [123, 14, 137, 13, 30].

Regarding our gradient regularization rule in (2.10), two observations come in order.

Remark 1. The truncation rule in (2.10) includes only relatively sizable $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t$'s, hence enforcing the smoothness of the (truncated) objective function $\ell_{\rm tr}(\boldsymbol{z}^t)$ at $\boldsymbol{z}^t$. Therefore, the truncated generalized gradient $\partial\ell_{\rm tr}(\boldsymbol{z})$ employed in (2.5) and (2.6) boils down to the ordinary gradient/Wirtinger derivative $\nabla\ell_{\rm tr}(\boldsymbol{z}^t)$ in the real/complex case.

Remark 2. As will be elaborated in (A.20) and (A.22), the quantities $(1/m)\sum_{i=1}^m \psi_i$ and $\max_{i\in[m]}\psi_i$ in (2.9) have magnitudes on the order of $\sqrt{\pi/2}\|\boldsymbol{x}\|$ and $\sqrt{m}\|\boldsymbol{x}\|$, respectively. In contrast, Prop. 1 asserts that the first term in (2.9) obeys $\|\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\| \approx \|\boldsymbol{h}\| \le \rho\|\boldsymbol{x}\|$ for a sufficiently small $\rho \ll \sqrt{\pi/2}$. Thus, spurious generalized gradient components typically have large magnitudes. It turns out that our gradient regularization rule in (2.10) also throws out gradient components of large sizes. To see this, for all $\boldsymbol{z}\in\mathbb{R}^n$ such that $\|\boldsymbol{h}\| \le \rho\|\boldsymbol{x}\|$ in (2.27), one can re-express

$$\sum_{i=1}^m \partial\ell_i(\boldsymbol{z}) = \sum_{i=1}^m \underbrace{\left(1-\frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|}\right)}_{\triangleq\beta_i}\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} \tag{2.11}$$

for some weight $\beta_i \in [-\infty,\, 1)$ assigned to the direction $\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} \approx \boldsymbol{z}$ due to $\mathbb{E}[\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}] = \boldsymbol{I}_n$. Then $\partial\ell_i(\boldsymbol{z})$ of an excessively large size corresponds to a large $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|$ in (2.11), or equivalently a small $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ in (2.10), thus causing the corresponding $\partial\ell_i(\boldsymbol{z})$ to be eliminated according to the truncation rule in (2.10).

Our truncation rule deviates from the intuition behind TWF, which throws away gradient components corresponding to large-size $\{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\}$ in (2.10). As demonstrated by our analysis in Appendix A.5, it rarely happens that a gradient component having large $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ yields an incorrect sign of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}$ under a sufficiently accurate initialization. Moreover, discarding too many samples (those for which $i \notin \mathcal{T}^{t+1}$ in TWF [32, Sec. 2.1]) introduces large bias into $(1/m)\sum_{i\in\mathcal{T}^{t+1}}^m \boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}$, so that TWF does not work well when m/n is close to the information-limit of $m/n \approx 2$. In sharp contrast, the motivation and objective of our truncation rule in (2.10) is to directly sense and eliminate gradient components that involve mistakenly estimated signs with high probability.

To demonstrate the power of TAF, numerical tests comparing all stages of (T)AF and (T)WF will be presented throughout our analysis. The basic test settings used are described next. For

fairness, all pertinent parameters involved in all compared schemes were set to their default values. Simulated estimates are averaged over 100 independent Monte Carlo (MC) realizations without mentioning this explicitly each time. Performance of different schemes is evaluated in terms of the relative root mean-square error (MSE), i.e.,

$$\text{Relative error} := \frac{\text{dist}(\boldsymbol{z},\, \boldsymbol{x})}{\|\boldsymbol{x}\|} \tag{2.12}$$

and the success rate among 100 trials, where a success is claimed for a trial if the returned estimate incurs a relative error less than 10^{-5} [32]. Simulated tests under both noiseless and noisy Gaussian models are performed, corresponding to

$$\psi_i = |\boldsymbol{a}_i^{\mathcal{H}}\boldsymbol{x}| + \eta_i$$

with $\eta_i = 0$ and $\eta_i \sim \mathcal{N}(0, \sigma^2)$, respectively, with i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ or $\boldsymbol{a}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$.

Numerical comparison depicted in Fig. 2.2 using the noiseless real Gaussian model suggests that even when starting with the same truncated spectral initialization, TAF's refinement outperforms those of TWF and WF, demonstrating the merits of our gradient update rule over TWF/WF. Furthermore, comparing TAF (gradient iterations in (2.5)-(2.6) with truncation in (2.10) initialized by the truncated spectral estimate) and AF (gradient iterations in (2.5)-(2.6) initialized by the truncated spectral estimate) corroborates the power of the truncation rule in (2.10).

2.2.2 Orthogonality-promoting initialization

Leveraging the SLLN, spectral initialization methods estimate $\boldsymbol{x}/\|\boldsymbol{x}\|$ as the leading eigenvector of $\boldsymbol{Y} := \frac{1}{m}\sum_{i\in\mathcal{T}^0} y_i \boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}$, where $\mathcal{T}^0$ is an index set accounting for possible data truncation. As asserted in [32], each summand $(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x})^2\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}$ follows a heavy-tail probability density function lacking a moment generating function. This causes major performance degradation especially when the number of measurements is small. Instead of spectral initializations, we shall take another route to bypass this hurdle. To gain intuition into our approach, a motivating example is presented first that reveals fundamental characteristics of high-dimensional vectors.

Fixing any nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$, generate data $\psi_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|$ using i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

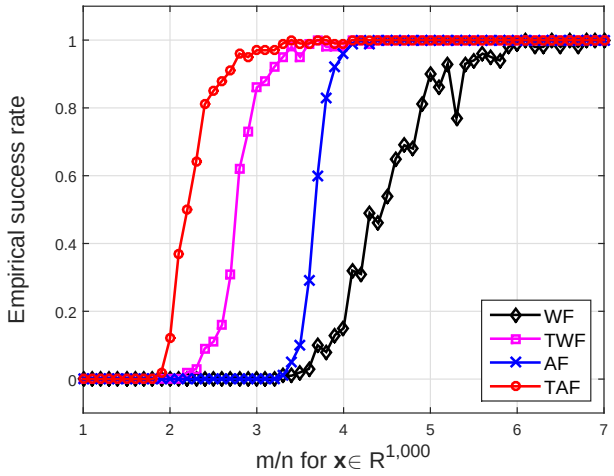

Figure 2.2: Empirical success rate from the same truncated spectral initialization under the real Gaussian model.

for $1 \le i \le m$. Evaluate the following squared normalized inner-product

$$\cos^2\theta_i := \frac{|\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2}{\|\boldsymbol{a}_i\|^2\|\boldsymbol{x}\|^2} = \frac{\psi_i^2}{\|\boldsymbol{a}_i\|^2\|\boldsymbol{x}\|^2}, \quad 1 \le i \le m \tag{2.13}$$

where θ_i is the angle between vectors $\boldsymbol{a}_i$ and $\boldsymbol{x}$. Consider ordering all $\{\cos^2\theta_i\}$ in an ascending fashion, and collectively denote them as $\boldsymbol{\xi} := [\cos^2\theta_{[m]} \ \cdots \ \cos^2\theta_{[1]}]^{\mathcal{T}}$ with $\cos^2\theta_{[1]} \ge \cdots \ge \cos^2\theta_{[m]}$. Figure 2.3 plots the ordered entries in $\boldsymbol{\xi}$ for m/n varying by 2 from 2 to 10 with $n = 1,000$. Observe that almost all $\{\boldsymbol{a}_i\}$ vectors have a squared normalized inner-product with $\boldsymbol{x}$ smaller than 10^{-2}, while half of the inner-products are less than 10^{-3}, which implies that $\boldsymbol{x}$ is nearly orthogonal to a large number of $\boldsymbol{a}_i$'s.

This example corroborates the folklore that random vectors in high-dimensional spaces are almost always nearly orthogonal to each other [17]. This inspired us to pursue an *orthogonality-promoting initialization method*. Our key idea is to approximate $\boldsymbol{x}$ by a vector that is most orthogonal to a subset of vectors $\{\boldsymbol{a}_i\}_{i\in\mathcal{I}^0}$, where $\mathcal{I}^0$ is an index set with cardinality $|\mathcal{I}^0| < m$ that includes indices of the smallest squared normalized inner-products $\cos^2\theta_i$. Since $\|\boldsymbol{x}\|$ appears in all inner-products, its exact value does not influence their ordering. Henceforth, we assume with no loss of generality that $\|\boldsymbol{x}\| = 1$.

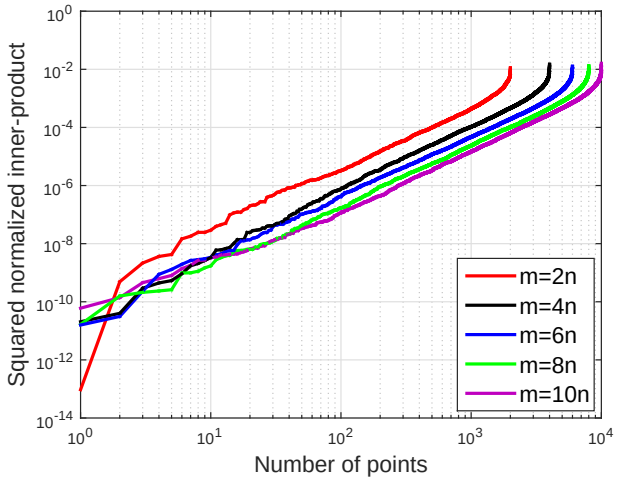

Figure 2.3: Ordered squared normalized inner-product for pairs $\boldsymbol{x}$ and $\boldsymbol{a}_i$.

Using data $\{(\boldsymbol{a}_i;\, \psi_i)\}$, evaluate $\cos^2\theta_i$ according to (2.13) for each pair $\boldsymbol{x}$ and $\boldsymbol{a}_i$. Instrumental for the ensuing derivations is noticing from the inherent near-orthogonal property of high-dimensional random vectors that the summation of $\cos^2\theta_i$ over all indices $i \in \mathcal{I}^0$ should be very small; rigorous justification is deferred to Sec. 2.5. Therefore, the sum $\sum_{i\in\mathcal{I}^0}\cos^2\theta_i$ is also small, or according to (2.13), equivalently,

$$\sum_{i\in\mathcal{I}^0}\frac{|\langle \boldsymbol{a}_i,\boldsymbol{x}\rangle|^2}{\|\boldsymbol{a}_i\|^2\|\boldsymbol{x}\|^2}=\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\Big(\sum_{i\in\mathcal{I}^0}\frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}\Big)\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \tag{2.14}$$

is small. Therefore, a meaningful approximation of $\boldsymbol{x}$ can be obtained by minimizing the former with $\boldsymbol{x}$ replaced by the optimization variable $\boldsymbol{z}$, namely

$$\underset{\|\boldsymbol{z}\|=1}{\text{minimize}}\;\; \boldsymbol{z}^{\mathcal{T}}\left(\frac{1}{|\mathcal{I}^0|}\sum_{i\in\mathcal{I}^0}\frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}\right)\boldsymbol{z}. \tag{2.15}$$

This amounts to finding the smallest eigenvalue and the associated eigenvector of $\boldsymbol{Y}_0 := \frac{1}{|\mathcal{I}^0|}\sum_{i\in\mathcal{I}^0}\frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} \succeq \boldsymbol{0}$. Finding the smallest eigenvalue calls for eigen-decomposition or matrix inversion, each typically requiring computational complexity on the order of $\mathcal{O}(n^3)$. Such a computational burden may be intractable when n grows large. Applying a standard concentration result, we show how the computation can be significantly reduced.

Since $\boldsymbol{a}_i/\|\boldsymbol{a}_i\|$ has unit norm and is uniformly distributed on the unit sphere, it is uniformly spherically distributed.[1] Spherical symmetry implies that $\boldsymbol{a}_i/\|\boldsymbol{a}_i\|$ has zero mean and covariance matrix $\boldsymbol{I}_n/n$ [119]. Appealing again to the SLLN, the sample covariance matrix $\frac{1}{m}\sum_{i=1}^m \frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}$ approaches $\boldsymbol{I}_n/n$ as m grows. Simple derivations lead to

$$\sum_{i\in\mathcal{I}^0} \frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} = \sum_{i=1}^m \frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} - \sum_{i\in\bar{\mathcal{I}}^0} \frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} \approxeq \frac{m}{n}\boldsymbol{I}_n - \sum_{i\in\bar{\mathcal{I}}^0} \frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2} \tag{2.16}$$

where $\bar{\mathcal{I}}^0$ is the complement of $\mathcal{I}^0$ in the set $[m]$. Define $\boldsymbol{S} := [\boldsymbol{a}_1/\|\boldsymbol{a}_1\| \ \cdots \ \boldsymbol{a}_m/\|\boldsymbol{a}_m\|]^{\mathcal{T}} \in \mathbb{R}^{m\times n}$, and form $\bar{\boldsymbol{S}}_0$ by removing the rows of $\boldsymbol{S}$ whose indices belong to $\mathcal{I}^0$. Seeking the smallest eigenvalue of $\boldsymbol{Y}_0 = \frac{1}{|\bar{\mathcal{I}}^0|}\boldsymbol{S}_0^{\mathcal{T}}\boldsymbol{S}_0$ then reduces to computing the largest eigenvalue of the matrix

$$\bar{\boldsymbol{Y}}_0 := \frac{1}{|\bar{\mathcal{I}}^0|}\bar{\boldsymbol{S}}_0^{\mathcal{T}}\bar{\boldsymbol{S}}_0, \tag{2.17}$$

namely,

$$\tilde{\boldsymbol{z}}^0 := \arg\max_{\|\boldsymbol{z}\|=1} \ \boldsymbol{z}^{\mathcal{T}}\bar{\boldsymbol{Y}}_0\boldsymbol{z} \tag{2.18}$$

which can be efficiently solved via simple power iterations.

When $\|\boldsymbol{x}\| \neq 1$, the estimate $\tilde{\boldsymbol{z}}^0$ from (2.18) is scaled so that its norm matches approximately that of $\boldsymbol{x}$, which is estimated as $\sqrt{\frac{1}{m}\sum_{i=1}^m y_i}$, or more accurately $\sqrt{\frac{n\sum_{i=1}^m y_i}{\sum_{i=1}^m \|\boldsymbol{a}_i\|^2}}$. To motivate these estimates, using the rotational invariance property of normal distributions, it suffices to consider the case where $\boldsymbol{x} = \|\boldsymbol{x}\|\boldsymbol{e}_1$, with $\boldsymbol{e}_1$ denoting the first canonical vector of $\mathbb{R}^n$. Indeed,

$$\left|\left\langle \boldsymbol{a}_i, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right\rangle\right|^2 = |\langle \boldsymbol{a}_i, \boldsymbol{U}\boldsymbol{e}_1\rangle|^2 = \left|\left\langle \boldsymbol{U}^{\mathcal{T}}\boldsymbol{a}_i, \boldsymbol{e}_1\right\rangle\right|^2 \overset{d}{=} |\langle \boldsymbol{a}_i, \boldsymbol{e}_1\rangle|^2 \tag{2.19}$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ is some unitary matrix, and $\overset{d}{=}$ means that terms on both sides of the equality have the same distribution. It is then easily verified that

$$\frac{1}{m}\sum_{i=1}^m y_i = \frac{1}{m}\sum_{i=1}^m a_{i,1}^2\|\boldsymbol{x}\|^2 \approx \|\boldsymbol{x}\|^2 \tag{2.20}$$

[1] A random vector $\boldsymbol{z} \in \mathbb{R}^n$ is said to be spherical (or spherically symmetric) if its distribution does not change under rotations of the coordinate system; that is, the distribution of $\boldsymbol{P}\boldsymbol{z}$ coincides with that of $\boldsymbol{z}$ for any given orthogonal $n \times n$ matrix $\boldsymbol{P}$.

where the last approximation arises from the following concentration result $(1/m)\sum_{i=1}^{m} a_{i,1}^2 \approx \mathbb{E}[a_{i,1}^2] = 1$ using again the SLLN. Regarding the second estimate, one can rewrite its square as

$$\frac{n\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m}\|\boldsymbol{a}_i\|^2} = \frac{1}{m}\sum_{i=1}^{m} y_i \cdot \frac{n}{(1/m)\sum_{i=1}^{m}\|\boldsymbol{a}_i\|^2}. \tag{2.21}$$

It is clear from (2.20) that the first term on the right hand side of (2.21) approximates $\|\boldsymbol{x}\|^2$. The second term approaches 1 because the denominator $(1/m)\sum_{i=1}^{m}\|\boldsymbol{a}_i\|^2 \approx n$ appealing to the SLLN again and the fact that $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. For brevity, we work with the first norm estimate

$$\boldsymbol{z}^0 = \sqrt{\frac{\sum_{i=1}^{m} y_i}{m}}\tilde{\boldsymbol{z}}^0. \tag{2.22}$$

It is worth highlighting that, compared to $\boldsymbol{Y} := \frac{1}{m}\sum_{i\in\mathcal{T}^0} y_i \boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}$ used in spectral methods, our constructed matrix $\bar{\boldsymbol{Y}}_0$ in (2.17) does not depend on the observed data y_i explicitly; the dependence is only through the choice of the index set $\mathcal{I}^0$. The novel orthogonality-promoting initialization thus enjoys two advantages over its spectral alternatives: a1) it does not suffer from heavy-tails of the fourth-order moments of Gaussian $\boldsymbol{a}_i$ vectors common in spectral initialization schemes; and, a2) it is less sensitive to noisy data.

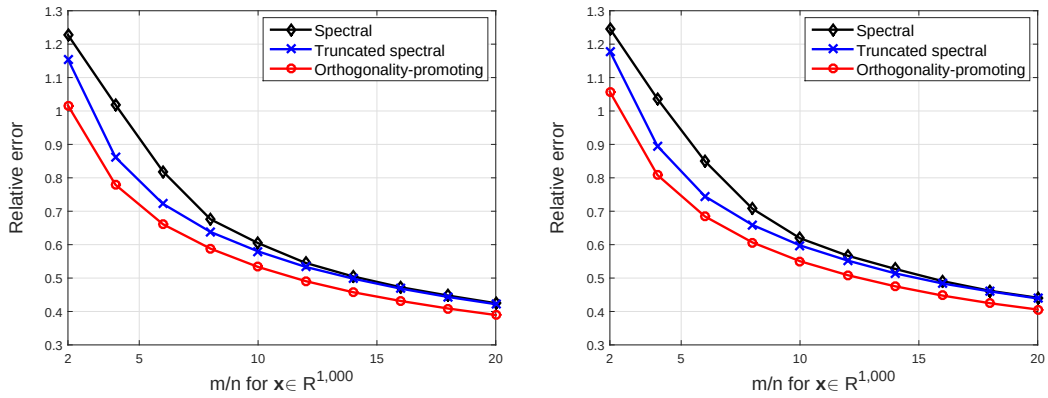

Figure 2.4: Relative initialization error versus m/n. Left: Noiseless real Gaussian model; Right: Noisy real Gaussian model with $\sigma^2 = 0.2^2\|\boldsymbol{x}\|^2$.

Figure 2.4 compares three different initialization schemes including spectral initialization [91, 22], truncated spectral initialization [32], and the proposed orthogonality-promoting initialization.

The relative error of their returned initial estimates versus the measurement/unknown ratio m/n is depicted under the noiseless and noisy real Gaussian models, where $\boldsymbol{x} \in \mathbb{R}^{1,000}$ was randomly generated and m/n increases by 2 from 2 to 20. Clearly, all schemes enjoy improved performance as m/n increases in both noiseless and noisy settings. The orthogonality-promoting initialization achieves consistently superior performance over its competing spectral alternatives under both noiseless and noisy Gaussian data. Interestingly, the spectral and truncated spectral schemes exhibit similar performance when m/n becomes sufficiently large (e.g., $m/n \geq 14$ in the noiseless setup or $m/n \geq 16$ in the noisy one). This confirms that the truncation helps only if m/n is relatively small. Indeed, the truncation discards measurements of excessively large or small sizes emerging from the heavy tails of the data distribution. Hence, its advantage over the non-truncated spectral initialization diminishes as the number of measurements increases, which gradually straightens out the heavy tails.

Algorithm 1 Truncated amplitude flow (TAF)

1: **Input:** Amplitude data $\{\psi_i := |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|\}_{i=1}^m$ and design vectors $\{\boldsymbol{a}_i\}_{i=1}^m$; maximum number of iterations T; by default, take constant step sizes $\mu = 0.6/1$ for the real/complex models, thresholds $|\overline{\mathcal{I}}^0| = \lceil \frac{1}{6}m \rceil$, and $\gamma = 0.7$.

2: **Set** $\overline{\mathcal{I}}^0$ as the set of indices corresponding to the $|\overline{\mathcal{I}}^0|$ largest values of $\{\psi_i/\|\boldsymbol{a}_i\|\}$.

3: **Initialize** $\boldsymbol{z}^0$ to $\sqrt{\frac{\sum_{i=1}^m \psi_i^2}{m}}\tilde{\boldsymbol{z}}^0$, where $\tilde{\boldsymbol{z}}^0$ is the normalized leading eigenvector of

$$\bar{\boldsymbol{Y}}_0 := \frac{1}{|\overline{\mathcal{I}}^0|} \sum_{i \in \overline{\mathcal{I}}^0} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}.$$

4: **Loop: for** $t = 0$ **to** $T - 1$

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \frac{\mu}{m} \sum_{i \in \mathcal{I}^{t+1}} \left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t - \psi_i \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t|} \right) \boldsymbol{a}_i$$

where $\mathcal{I}^{t+1} := \left\{ 1 \leq i \leq m \,\middle|\, |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{1}{1+\gamma}\psi_i \right\}$.

5: **Output:** $\boldsymbol{z}^T$.

2.3 Main Results

The TAF algorithm is summarized in Alg. 1. Default values are set for pertinent algorithmic parameters. Assuming independent data samples $(\boldsymbol{a}_i; \psi_i)$ drawn from the noiseless real Gaussian model, the following result establishes the theoretical performance of TAF.

Theorem 1 (Exact recovery)**.** *Let $\boldsymbol{x} \in \mathbb{R}^n$ be an arbitrary signal vector, and consider (noise-free) measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$, in which $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le i \le m$. Then with probability at least $1 - (m+5)\mathrm{e}^{-n/2} - \mathrm{e}^{-c_0 m} - 1/n^2$ for some universal constant $c_0 > 0$, the initialization $\boldsymbol{z}^0$ returned by the orthogonality-promoting method in Alg. 1 satisfies*

$$\operatorname{dist}(\boldsymbol{z}^0, \boldsymbol{x}) \le \rho \|\boldsymbol{x}\| \tag{2.23}$$

with $\rho = 1/10$ (or any sufficiently small positive constant), provided that $m \ge c_1 |\overline{\mathcal{I}}^0| \ge c_2 n$ for some numerical constants $c_1,\, c_2 > 0$, and sufficiently large n. Furthermore, choosing a constant step size $\mu \le \mu_0$ along with a truncation level $\gamma \ge 1/2$, and starting from any initial guess $\boldsymbol{z}^0$ satisfying (2.23)*, successive estimates of the TAF solver (tabulated in Alg. 1) obey*

$$\operatorname{dist}\big(\boldsymbol{z}^t, \boldsymbol{x}\big) \le \rho\,(1-\nu)^t \|\boldsymbol{x}\|, \quad t = 0,\, 1,\, 2,\, \ldots \tag{2.24}$$

for some constant $0 < \nu < 1$, which holds with probability exceeding $1 - (m+5)\mathrm{e}^{-n/2} - 8\mathrm{e}^{-c_0 m} - 1/n^2$.

Typical parameter values for TAF in Alg. 1 are $\mu = 0.6$, and $\gamma = 0.7$. The proof of Thm. 2 is relegated to Sec. 2.5. Theorem 2 asserts that: i) TAF reconstructs the solution $\boldsymbol{x}$ exactly as soon as the number of equations is about the number of unknowns, which is theoretically order optimal. Our numerical tests demonstrate that for the real Gaussian model, TAF achieves a success rate of 100% when m/n is as small as 3, which is slightly larger than the information limit of $m/n = 2$. This is a significant reduction in the sample complexity ratio, which is 5 for TWF and 7 for WF. Surprisingly, TAF also enjoys a success rate of over 50% when m/n is the information limit 2, which has not yet been presented for any existing algorithms. See further discussion in Sec. 2.4; and, ii) TAF converges exponentially fast with convergence rate independent of n. Specifically, TAF requires at most $\mathcal{O}(\log(1/\epsilon))$ iterations to achieve any given solution accuracy $\epsilon > 0$ (a.k.a., $\operatorname{dist}(\boldsymbol{z}^t, \boldsymbol{x}) \le \epsilon \|\boldsymbol{x}\|$), with iteration cost $\mathcal{O}(mn)$. Since the truncation takes time on the order of $\mathcal{O}(m)$, the computational burden of TAF per iteration is

dominated by the evaluation of the gradient components. The latter involves two matrix-vector multiplications that are computable in $\mathcal{O}(mn)$ flops, namely, $\boldsymbol{A}\boldsymbol{z}^t$ yields $\boldsymbol{u}^t$, and $\boldsymbol{A}^{\mathcal{T}}\boldsymbol{v}^t$ the gradient, where $\boldsymbol{v}^t := \boldsymbol{u}^t - \boldsymbol{\psi} \odot \frac{\boldsymbol{u}^t}{|\boldsymbol{u}^t|}$. Hence, the total running time of TAF is $\mathcal{O}(mn \log(1/\epsilon))$, which is proportional to the time taken to read the data $\mathcal{O}(mn)$.

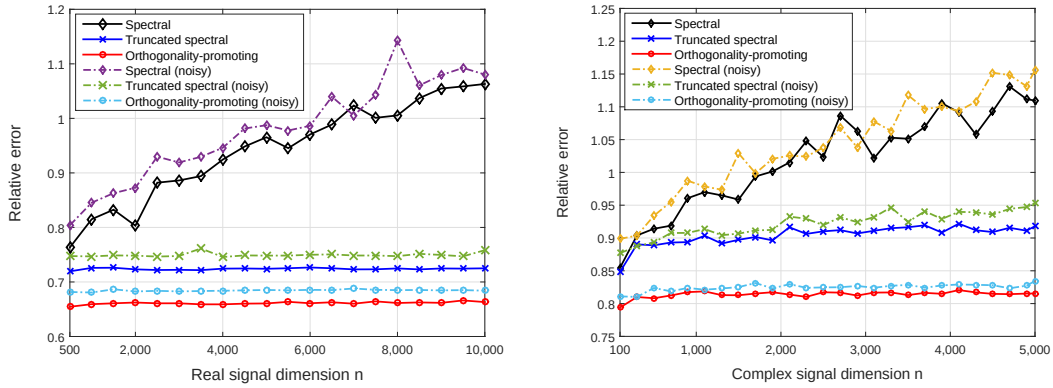

Figure 2.5: Relative initialization error using noise-free (solid) and noisy (dotted) data. Left: Real Gaussian model with $\sigma^2 = 0.2^2 \|\boldsymbol{x}\|^2$; Right: Complex Gaussian model with $\sigma^2 = 0.2^2 \|\boldsymbol{x}\|^2$.

In the noisy setting, TAF is stable under additive noise. Consider the amplitude-based noisy data model $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}| + \eta_i$. It can be shown that the TAF estimates in Alg. 1 satisfy

$$\operatorname{dist}(\boldsymbol{z}^t, \boldsymbol{x}) \lesssim (1-\nu)^t \|\boldsymbol{x}\| + \frac{1}{\sqrt{m}} \|\boldsymbol{\eta}\|, \quad t = 0, 1, \ldots \tag{2.25}$$

with high probability for all $\boldsymbol{x} \in \mathbb{R}^n$, provided that $m \geq c_1 |\overline{\mathcal{I}}^0| \geq c_2 n$ for sufficiently large n and the noise is bounded $\|\boldsymbol{\eta}\|_\infty \leq c_3 \|\boldsymbol{x}\|$ with $\boldsymbol{\eta} := [\eta_1 \ \cdots \ \eta_n]^{\mathcal{T}}$, where $0 < \nu < 1$, and $c_1, c_2, c_3 > 0$ are some universal constants. The proof can be adapted from those of Thm. 2 above and Thm. 2 in [32].

2.4 Numerical Experiments

In this section, we provide additional numerical tests evaluating performance of the proposed schemes relative to (T)WF [2]. The initial estimate was found using 50 power iterations, and

[2]Matlab codes directly downloaded from the authors' websites: `http://statweb.stanford.edu/~candes/TWF/algorithm.html`; and `http://www-bcf.usc.edu/~soltanol/WFcode.html`.

was subsequently refined by $T = 1,000$ gradient-type iterations in each scheme. The Matlab implementations of TAF are available at `https://gangwg.github.io/TAF/` for reproducibility.

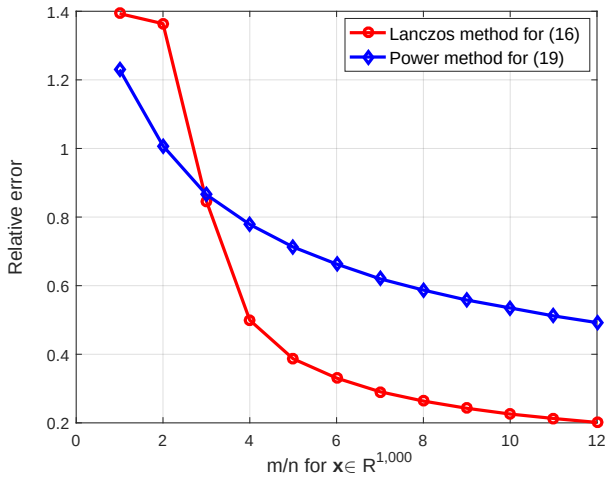

Figure 2.6: Relative initialization errors of solving (2.15) via the Lanczos method and solving (2.18) via the power method.

Top panel in Fig. 2.5 presents the average relative error of three initialization methods on a series of noiseless/noisy real Gaussian problems with $m/n = 6$ fixed, and n varying from 500 to 10^4, while those for the corresponding complex Gaussian instances are shown in the bottom panel. Clearly, the proposed initialization method returns more accurate and robust estimates than the spectral ones. Under the same condition for the real Gaussian model, Fig. 2.6 compares the initialization implemented in Alg. 1 obtained by solving the maximum eigenvalue problem in (2.18) with the one obtained by tackling the minimum eigenvalue problem in (2.15) via the Lanczos method [100]. When the number of equations is relatively small (less than about $3n$), the former performs better than the latter. Interestingly though, the latter works remarkably well and almost halves the error incurred by the implemented initialization of Alg. 1 as soon as the number of equations becomes larger than 4.

To demonstrate the power of TAF, Fig. 2.8 plots the relative error of recovering a real signal in logarithmic scale versus the iteration count under the information-limit of $m = 2n - 1$ noiseless i.i.d. Gaussian measurements [8]. In this case, since the returned initial estimate is relatively far from the optimal solution (see Fig. 2.4), TAF converges slowly for the first 200 iterations

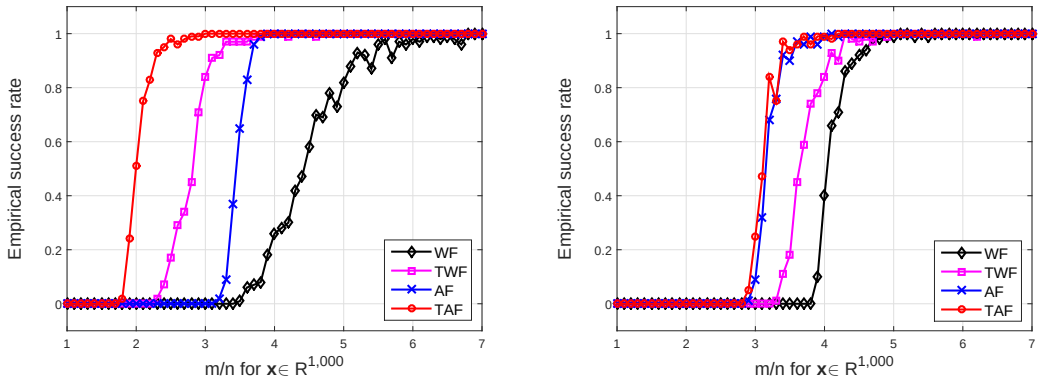

Figure 2.7: Empirical success rate. Left: Real Gaussian model; Right: Complex Gaussian model.

or so due to elimination of a significant amount of 'bad' generalized gradient components. As the iterate gets more accurate and lands within a small-size neighborhood of $\boldsymbol{x}$, TAF converges exponentially fast to the globally optimal solution. It is worth emphasizing that no existing method succeeds in this case. Figure 2.7 compares the empirical success rate of three schemes under both real and complex Gaussian models with $n = 10^3$ and m/n varying by 0.1 from 1 to 7, where a success is claimed if the estimate has a relative error less than 10^{-5}. For real vectors, TAF achieves a success rate of over 50% when $m/n = 2$, and guarantees perfect recovery from about $3n$ measurements; while for complex ones, TAF enjoys a success rate of 95% when $m/n = 3.4$, and ensures perfect recovery from about $4.5n$ measurements.

To demonstrate the stability of TAF, the relative MSE

$$\text{Relative MSE} := \frac{\text{dist}^2(\boldsymbol{z}^T,\, \boldsymbol{x})}{\|\boldsymbol{x}\|^2}$$

as a function of the signal-to-noise ratio (SNR) is plotted for different m/n values. We consider the noisy model $\psi_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle| + \eta_i$ with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{1,000})$ and real independent Gaussian sensing vectors $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{1,000})$, in which m/n takes values $\{6,\, 8,\, 10\}$, and the SNR in dB, given by

$$\text{SNR} := 10 \log_{10} \frac{\sum_{i=1}^m |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2}{\sum_{i=1}^m \eta_i^2}$$

is varied from 10 dB to 50 dB. Averaging over 100 independent trials, Fig. 2.9 demonstrates that the relative MSE for all m/n values scales inversely proportional to SNR, hence justifying the

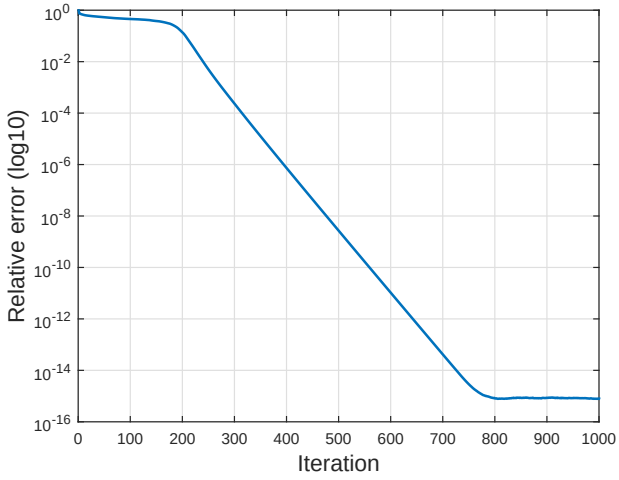

Figure 2.8: Relative error versus iteration for TAF with $m = 2n - 1$.

stability of TAF under bounded additive noise.

The next experiment evaluates the efficacy of the proposed initialization method, simulating all schemes initialized by the truncated spectral estimate [32] and the orthogonality-promoting estimate. Evidently, all solvers except for WF admit a significant performance improvement when initialized by the proposed orthogonality-promoting initialization relative to the truncated spectral initialization. Nonetheless, TAF with our developed orthogonality-promoting initialization enjoys superior performance over all simulated approaches.

Finally, to examine the effectiveness and scalability of TAF in real-world conditions, we simulate recovery of the Milky Way Galaxy image [3] $\boldsymbol{X} \in \mathbb{R}^{1080\times 1920\times 3}$ shown in Fig. 2.11. The first two indices encode the pixel locations, and the third the RGB (red, green, blue) color bands. Consider the coded diffraction pattern (CDP) measurements with random masks [21, 22, 32]. Letting $\boldsymbol{x} \in \mathbb{R}^n$ be a vectorization of a certain band of $\boldsymbol{X}$ and postulating a number K of random masks, one can further write

$$\boldsymbol{\psi}^{(k)} = \left|\boldsymbol{F}\boldsymbol{D}^{(k)}\boldsymbol{x}\right|, \quad k = 1,\, 2,\, \ldots,\, K \tag{2.26}$$

where $\boldsymbol{F}$ denotes the $n \times n$ discrete Fourier transform matrix, and $\boldsymbol{D}^{(k)}$ is a diagonal matrix holding entries sampled uniformly at random from $\{1,\, -1,\, j,\, -j\}$ (also known as phase delays)

[3] Downloaded from `http://pics-about-space.com/milky-way-galaxy`.

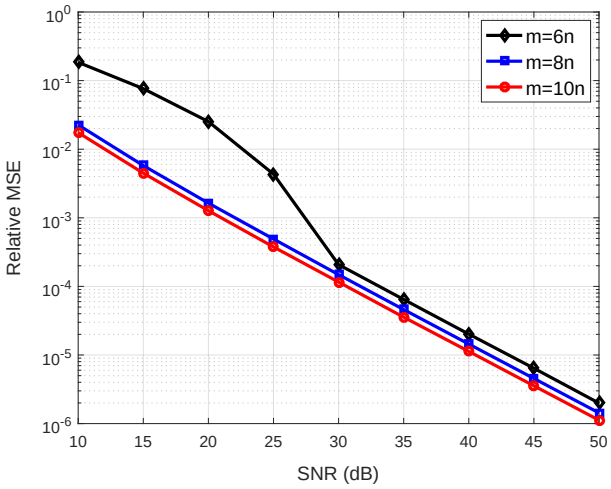

Figure 2.9: Relative MSE versus SNR for TAF under the amplitude-based noisy data model.

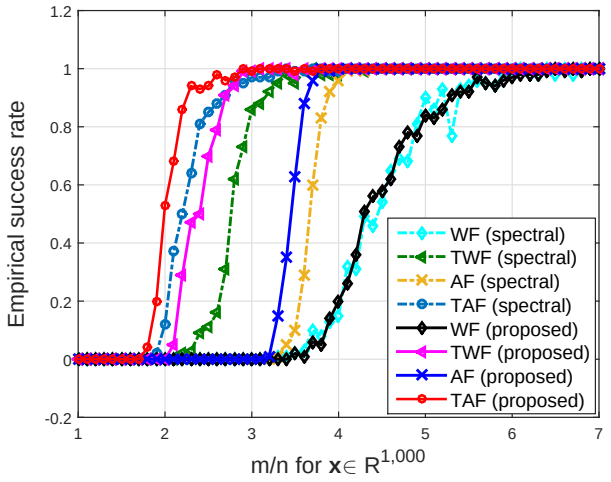

Figure 2.10: Empirical success rate using the truncated spectral and the orthogonality-promoting initializations.

on its diagonal, with j denoting the imaginary unit. Each $\boldsymbol{D}^{(k)}$ represents a random mask placed after the object [21]. With $K = 6$ masks implemented in our experiment, the total number of quadratic measurements is $m = 6n$. Per algorithm was run independently on each of the three bands. A number 100 of power iterations were used to obtain an initialization, which was refined by 100 gradient-type iterations. The relative errors after our orthogonality-promoting

initialization and after 100 TAF iterations are 0.6807 and 9.8631×10^{-5}, respectively, and the recovered images are displayed in Fig. 2.11. In sharp contrast, TWF returns images of corresponding relative errors 1.3801 and 1.3409, which are far away from the ground truth.

Regarding runtimes in our reported experiments, TAF converges slightly faster than TWF, while both are markedly faster than WF. All experiments in this chapter and subsequent chapters were implemented using MATLAB on an Intel CPU @ 3.4 GHz (32 GB RAM) desktop computer.

2.5 Proofs

This section presents the main ideas behind the proof of Thm. 2, and establishes a few necessary lemmas. Technical details are deferred to the Appendix. Relative to WF and TWF, our objective function involves non-smoothness and non-convexity, rendering the proof of exact recovery of TAF nontrivial. In addition, our initialization method starts from a rather different perspective than spectral alternatives, so that the tools involved in proving performance of our initialization deviate from those of spectral methods [91, 22, 32].

The proof of Thm. 2 consists of two parts: Sec. 2.5.1 justifies the performance of the proposed orthogonality-promoting initialization, which essentially achieves any given constant relative error as soon as the number of equations is on the order of the number of unknowns, namely, $m \asymp n$. Section 2.5.2 demonstrates theoretical convergence of TAF to the solution of the quadratic system in (2.1) at a geometric rate, provided that the initial estimate has a sufficiently small constant relative error.

2.5.1 Constant relative error by initialization

This section concentrates on proving guaranteed performance of the proposed orthogonality-promoting initialization method, as asserted in the following proposition. An alternative approach may be found in [43].

Proposition 1. *Fix $\boldsymbol{x} \in \mathbb{R}^n$ arbitrarily, and consider the noiseless case $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$, where $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0},\, \boldsymbol{I}_n)$ for $i = 1,\, 2,\, \ldots,\, m$. With probability at least $1-(m+5)\mathrm{e}^{-n/2}-\mathrm{e}^{-c_0 m}-1/n^2$ for some universal constant $c_0 > 0$, the initialization $\boldsymbol{z}^0$ returned by the orthogonality-promoting method satisfies*

$$\operatorname{dist}(\boldsymbol{z}^0,\, \boldsymbol{x}) \le \rho \, \|\boldsymbol{x}\| \tag{2.27}$$

for $\rho = 1/10$ or any positive constant, with the proviso that $m \geq c_1|\overline{\mathcal{I}}^0| \geq c_2 n$ for some numerical constants $c_1,\, c_2 > 0$ and sufficiently large n.

Due to homogeneity in (2.27), it suffices to consider the case $\|\boldsymbol{x}\| = 1$. Assume for the moment that $\|\boldsymbol{x}\| = 1$ is known and $\boldsymbol{z}^0$ has been scaled such that $\|\boldsymbol{z}^0\| = 1$ in (2.22). The error between the employed $\boldsymbol{x}$'s norm estimate $\sqrt{\frac{1}{m}\sum_{i=1}^m y_i}$ and the unknown norm $\|\boldsymbol{x}\| = 1$ will be accounted for at the end of this section. Instrumental in proving Prop. 1 is the following result, whose proof is provided in Appendix A.1.

Lemma 1. *Consider the noiseless data $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$, where $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0},\, \boldsymbol{I}_n)$ for $i = 1,\, 2,\, \ldots,\, m$. For any unit vector $\boldsymbol{x} \in \mathbb{R}^n$, there exists a vector $\boldsymbol{u} \in \mathbb{R}^n$ with $\boldsymbol{u}^{\mathcal{T}}\boldsymbol{x} = 0$ and $\|\boldsymbol{u}\| = 1$ such that*

$$\frac{1}{2}\left\|\boldsymbol{x}\boldsymbol{x}^{\mathcal{T}} - \boldsymbol{z}^0(\boldsymbol{z}^0)^{\mathcal{T}}\right\|_F^2 \leq \frac{\|\overline{\boldsymbol{S}}_0\boldsymbol{u}\|^2}{\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\|^2} \tag{2.28}$$

for $\boldsymbol{z}^0 = \tilde{\boldsymbol{z}}^0$, where the unit vector $\tilde{\boldsymbol{z}}^0$ is given in (2.18)*, and $\overline{\boldsymbol{S}}_0$ is formed by removing the rows of $\boldsymbol{S} := [\boldsymbol{a}_1/\|\boldsymbol{a}_1\| \;\cdots\; \boldsymbol{a}_m/\|\boldsymbol{a}_m\|]^{\mathcal{T}} \in \mathbb{R}^{m\times n}$ if their indices do not belong to the set $\overline{\mathcal{I}}^0$ specified in Alg. 1.*

We now turn to prove Prop. 1. The first step consists in upper-bounding the term on the right-hand-side of (2.28). Specifically, its numerator is upper bounded, and the denominator lower bounded, as summarized in Lemmas 2 and 3 next; their proofs are provided in Appendix A.2 and Appendix A.3, respectively.

Lemma 2. *In the setup of Lemma 1, if $|\overline{\mathcal{I}}^0| \geq c_1' n$, then*

$$\left\|\overline{\boldsymbol{S}}_0\boldsymbol{u}\right\|^2 \leq 1.01|\overline{\mathcal{I}}^0|/n \tag{2.29}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_K n}$, where c_2' and c_K are some universal constants.

Lemma 3. *In the setup of Lemma 1, the following holds with probability at least $1 - (m+1)\mathrm{e}^{-n/2} - \mathrm{e}^{-c_0 m} - 1/n^2$:*

$$\left\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\right\|^2 \geq \frac{0.99|\overline{\mathcal{I}}^0|}{2.3n}\left[1 + \log(m/|\overline{\mathcal{I}}^0|)\right] \tag{2.30}$$

provided that $|\overline{\mathcal{I}}^0| \geq c_1' n$, $m \geq c_2'|\overline{\mathcal{I}}^0|$, and $m \geq c_3' n$ for some absolute constants $c_1',\, c_2',\, c_3' > 0$, and sufficiently large n.

Leveraging the upper and lower bounds in (2.29) and (2.30), one arrives at

$$\frac{\left\|\bar{\boldsymbol{S}}_0\boldsymbol{u}\right\|^2}{\left\|\bar{\boldsymbol{S}}_0\boldsymbol{x}\right\|^2} \leq \frac{2.4}{1+\log(m/|\bar{\mathcal{I}}^0|)} \stackrel{\triangle}{=} \kappa \tag{2.31}$$

which holds with probability at least $1-(m+3)\mathrm{e}^{-n/2}-\mathrm{e}^{-c_0 m}-1/n^2$, assuming that $m \geq c_1'|\bar{\mathcal{I}}^0|$, and $m \geq c_2' n$, $|\bar{\mathcal{I}}^0| \geq c_3' n$ for some absolute constants c_1', c_2', $c_3' > 0$, and sufficiently large n.

The bound κ in (2.31) is meaningful only when the ratio $\log(m/|\bar{\mathcal{I}}^0|) > 1.4$, i.e., $m/|\bar{\mathcal{I}}^0| > 4$, because the left hand side is expressible in terms of $\sin^2\theta$, and therefore, enjoys a trivial upper bound of 1. Henceforth, we will assume $m/|\bar{\mathcal{I}}^0| > 4$. Empirically, $\lfloor m/|\bar{\mathcal{I}}^0| \rfloor = 6$ or equivalently $|\bar{\mathcal{I}}^0| = \lceil \frac{1}{6} m \rceil$ in Alg. 1 works well when m/n is relatively small. Note further that the bound κ can be made arbitrarily small by letting $m/|\bar{\mathcal{I}}^0|$ be large enough. Without any loss of generality, let us take $\kappa := 0.001$. An additional step leads to the wanted bound on the distance between $\tilde{\boldsymbol{z}}^0$ and $\boldsymbol{x}$; similar arguments are found in [22, Sec. 7.8]. Recall that

$$|\boldsymbol{x}^{\mathcal{T}}\tilde{\boldsymbol{z}}^0|^2 = \cos^2\theta = 1-\sin^2\theta \geq 1-\kappa. \tag{2.32}$$

Therefore,

$$\begin{aligned} \mathrm{dist}^2(\tilde{\boldsymbol{z}}^0, \boldsymbol{x}) &\leq \|\tilde{\boldsymbol{z}}^0\|^2 + \|\boldsymbol{x}\|^2 - 2|\boldsymbol{x}^{\mathcal{T}}\tilde{\boldsymbol{z}}^0| \\ &\leq \left(2-2\sqrt{1-\kappa}\right)\|\boldsymbol{x}\|^2 \\ &\approx \kappa\|\boldsymbol{x}\|^2. \end{aligned} \tag{2.33}$$

Coming back to the case in which $\|\boldsymbol{x}\|$ is unknown stated prior to Lemma 1, the unit eigenvector $\tilde{\boldsymbol{z}}^0$ is scaled by an estimate of $\|\boldsymbol{x}\|$ to yield the initial guess $\boldsymbol{z}^0 = \sqrt{\frac{1}{m}\sum_{i=1}^m y_i}\tilde{\boldsymbol{z}}^0$. Using the results in Lemma 7.8 in [22], the following holds with high probability

$$\|\boldsymbol{z}^0 - \tilde{\boldsymbol{z}}^0\| = |\|\boldsymbol{z}^0\| - 1| \leq (1/20)\|\boldsymbol{x}\|. \tag{2.34}$$

Summarizing the two inequalities, we conclude that

$$\mathrm{dist}(\boldsymbol{z}^0, \boldsymbol{x}) \leq \|\boldsymbol{z}^0 - \tilde{\boldsymbol{z}}^0\| + \mathrm{dist}(\tilde{\boldsymbol{z}}^0, \boldsymbol{x}) \leq (1/10)\|\boldsymbol{x}\|. \tag{2.35}$$

The initialization thus obeys $\mathrm{dist}(\boldsymbol{z}^0, \boldsymbol{x})/\|\boldsymbol{x}\| \leq 1/10$ for any $\boldsymbol{x} \in \mathbb{R}^n$ with high probability

provided that $m \geq c_1|\overline{\mathcal{I}}^0| \geq c_2 n$ holds for some universal constants $c_1,\, c_2 > 0$ and sufficiently large n.

2.5.2 Exact recovery from noiseless data

We now prove that with accurate enough initial estimates, TAF converges at a geometric rate to $\boldsymbol{x}$ with high probability (i.e., the second part of Thm. 2). To be specific, with initialization obeying (2.27) in Prop. 1, TAF reconstructs the solution $\boldsymbol{x}$ exactly in linear time. To start, it suffices to demonstrate that the TAF's update rule (i.e., Step 4 in Alg. 1) is locally contractive within a sufficiently small neighborhood of $\boldsymbol{x}$, as asserted in the following proposition.

Proposition 2 (Local error contraction)**.** *Consider the noise-free measurements $\psi_i = \left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}\right|$ with i.i.d. Gaussian design vectors $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0},\, \boldsymbol{I}_n)$, $1 \leq i \leq m$, and fix any $\gamma \geq 1/2$. There exist universal constants $c_0,\, c_1 > 0$ and $0 < \nu < 1$ such that with probability at least $1 - 7\mathrm{e}^{-c_0 m}$, the following holds*

$$\mathrm{dist}^2\left(\boldsymbol{z} - \frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}),\, \boldsymbol{x}\right) \leq (1-\nu)\mathrm{dist}^2\left(\boldsymbol{z},\, \boldsymbol{x}\right) \tag{2.36}$$

for all $\boldsymbol{x},\, \boldsymbol{z} \in \mathbb{R}^n$ obeying (2.27) *with the proviso that $m \geq c_1 n$ and that the constant step size μ satisfies $0 < \mu \leq \mu_0$ for some $\mu_0 > 0$.*

Proposition 8 asserts that the distance of TAF's successive iterates to $\boldsymbol{x}$ is monotonically decreasing once the algorithm enters a small-size neighborhood around $\boldsymbol{x}$. This neighborhood is commonly referred to as the basin of attraction; see related discussions in [22, 111, 32]. In other words, as soon as one lands within the basin of attraction, TAF's iterates remain in this region and will be attracted to $\boldsymbol{x}$ exponentially fast. To substantiate Prop. 8, recall the local regularity condition, which plays a fundamental role in establishing linear convergence to global optimum of non-convex optimization approaches such as WF/TWF [22, 111, 32].

Consider the update rule of TAF

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}^t), \quad t = 0,\, 1,\, 2,\, \ldots \tag{2.37}$$

where the truncated gradient $\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}^t)$ (as elaborated in Rmk. 1) evaluated at some point $\boldsymbol{z}^t \in \mathbb{R}^n$

is given by

$$\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z}^t) \triangleq \frac{1}{m}\sum_{i\in\mathcal{I}^{t+1}}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t - \psi_i\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}\right)\boldsymbol{a}_i.$$

The truncated gradient $\nabla\ell_{\rm tr}(\boldsymbol{z})$ is said to satisfy the local regularity condition, or $\mathrm{LRC}(\mu,\lambda,\epsilon)$ for some constant $\lambda>0$, provided that

$$\left\langle \frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z}),\, \boldsymbol{h}\right\rangle \geq \frac{\mu}{2}\left\|\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})\right\|^2 + \frac{\lambda}{2}\|\boldsymbol{h}\|^2 \tag{2.38}$$

holds for all $\boldsymbol{z}\in\mathbb{R}^n$ such that $\|\boldsymbol{h}\| = \|\boldsymbol{z}-\boldsymbol{x}\| \leq \epsilon\|\boldsymbol{x}\|$ for some constant $0<\epsilon<1$, where the ball $\|\boldsymbol{z}-\boldsymbol{x}\|\leq\epsilon\|\boldsymbol{x}\|$ is the basin of attraction. Simple linear algebra along with the regularity condition in (2.38) leads to

$$\begin{aligned}
\mathrm{dist}^2\left(\boldsymbol{z}-\frac{\mu}{m}\nabla\ell_{\rm tr}(\boldsymbol{z}),\boldsymbol{x}\right) &= \left\|\boldsymbol{z}-\frac{\mu}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})-\boldsymbol{x}\right\|^2 \\
&= \|\boldsymbol{h}\|^2 - 2\mu\left\langle \boldsymbol{h}, \frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})\right\rangle + \left\|\frac{\mu}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})\right\|^2 \qquad (2.39)\\
&\leq \|\boldsymbol{h}\|^2 - 2\mu\left(\frac{\mu}{2}\left\|\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})\right\|^2 + \frac{\lambda}{2}\|\boldsymbol{h}\|^2\right) + \left\|\frac{\mu}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})\right\|^2 \\
&= (1-\lambda\mu)\|\boldsymbol{h}\|^2 \\
&= (1-\lambda\mu)\,\mathrm{dist}^2(\boldsymbol{z},\,\boldsymbol{x}) \qquad (2.40)
\end{aligned}$$

for all $\boldsymbol{z}$ obeying $\|\boldsymbol{h}\|\leq\epsilon\|\boldsymbol{x}\|$. Evidently, if the $\mathrm{LRC}(\mu,\lambda,\epsilon)$ is proved for TAF, then (2.36) follows upon letting $\nu := \lambda\mu$.

Proof of the local regularity condition in (2.38)

By definition, justifying the local regularity condition in (2.38) entails controlling the norm of the truncated gradient $\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})$, i.e., bounding the last term in (2.39). Recall that

$$\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z}) = \frac{1}{m}\sum_{i\in\mathcal{I}}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} - \psi_i\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|}\right)\boldsymbol{a}_i \triangleq \frac{1}{m}\boldsymbol{A}\boldsymbol{v} \tag{2.41}$$

where $\mathcal{I} := \{1 \le i \le m | |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}| \ge |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|/(1+\gamma)\}$, and $\boldsymbol{v} := [v_1 \ \cdots \ v_m]^{\mathcal{T}} \in \mathbb{R}^m$ with $v_i := \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|} \left(|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}| - \psi_i\right) \mathbb{1}_{\left\{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}| \ge |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|/(1+\gamma)\right\}}$. Now, consider

$$\begin{aligned} |v_i|^2 &= \left| (|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}| - |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|) \mathbb{1}_{\left\{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}| \ge |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|/(1+\gamma)\right\}} \right|^2 \\ &\le \left| |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}| - |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}| \right|^2 \\ &\le |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}|^2 \end{aligned} \tag{2.42}$$

where $\boldsymbol{h} = \boldsymbol{z} - \boldsymbol{x}$. Appealing to [23, Lemma 3.1], fixing any $\delta' > 0$, the following holds for any $\boldsymbol{h} \in \mathbb{R}^n$ with probability at least $1 - \mathrm{e}^{-m\delta'^2/2}$:

$$\|\boldsymbol{v}\|^2 = \sum_{i=1}^m v_i^2 \le \sum_{i=1}^m |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}|^2 \le (1+\delta') m \|\boldsymbol{h}\|^2. \tag{2.43}$$

On the other hand, standard matrix concentration results confirm that the largest singular value of $\boldsymbol{A} = [\boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_m]^{\mathcal{T}}$ with i.i.d. Gaussian $\{\boldsymbol{a}_i\}$ satisfies $\sigma_1 := \|\boldsymbol{A}\| \le (1+\delta'')\sqrt{m}$ for some $\delta'' > 0$ with probability exceeding $1 - 2\mathrm{e}^{-c_0 m}$ as soon as $m \ge c_1 n$ for sufficiently large $c_1 > 0$, where $c_1 > 0$ is a universal constant depending on δ'' [119, Rmk. 5.25]. Combining (2.41), (2.42), and (2.43) yields

$$\begin{aligned} \left\| \frac{1}{m} \nabla \ell_{\mathrm{tr}}(\boldsymbol{z}) \right\| &\le \frac{1}{m} \|\boldsymbol{A}\| \, \|\boldsymbol{v}\| \\ &\le (1+\delta')(1+\delta'') \|\boldsymbol{h}\| \\ &\le (1+\delta)^2 \|\boldsymbol{h}\|, \quad \delta := \max\{\delta', \delta''\} \end{aligned} \tag{2.44}$$

which holds with high probability. This condition essentially asserts that the truncated gradient of the objective function $\ell(\boldsymbol{z})$ or the search direction is well behaved (the function value does not vary too much).

We have related $\|\nabla \ell_{\mathrm{tr}}(\boldsymbol{z})\|^2$ to $\|\boldsymbol{h}\|^2$ through (2.44). Therefore, a more conservative lower bound for $\langle \frac{1}{m} \nabla \ell_{\mathrm{tr}}(\boldsymbol{z}), \, \boldsymbol{h} \rangle$ in LRC can be given in terms of $\|\boldsymbol{h}\|^2$. It is equivalent to show that the truncated gradient $\frac{1}{m} \nabla \ell_{\mathrm{tr}}(\boldsymbol{z})$ ensures sufficient descent, i.e., it obeys a uniform lower bound along the search direction $\boldsymbol{h}$ taking the form

$$\left\langle \frac{1}{m} \nabla \ell_{\mathrm{tr}}(\boldsymbol{z}), \, \boldsymbol{h} \right\rangle \gtrsim \|\boldsymbol{h}\|^2 \tag{2.45}$$

which occupies the remaining of this section. Formally, this can be stated as follows.

Proposition 3. *Consider the noiseless measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$, and fix any sufficiently small constant $\epsilon > 0$. There exist universal constants $c_0,\, c_1 > 0$ such that if $m > c_1 n$, then the following holds with probability exceeding $1 - 4\mathrm{e}^{-c_0 m}$:*

$$\left\langle \frac{1}{m} \nabla \ell_{\mathrm{tr}}(\boldsymbol{z}), \boldsymbol{h} \right\rangle \geq 2\left(1 - \zeta_1 - \zeta_2 - 2\epsilon\right) \|\boldsymbol{h}\|^2 \tag{2.46}$$

for all $\boldsymbol{x},\, \boldsymbol{z} \in \mathbb{R}^n$ such that $\|\boldsymbol{h}\| \,/\, \|\boldsymbol{x}\| \leq \rho$ for $0 < \rho \leq 1/10$ and any fixed $\gamma \geq 1/2$, where the estimates $\zeta_1 \approx 0.0782$, and $\zeta_2 \approx 0.3894$.

Before justifying Prop. 3, we introduce the following events.

Lemma 4. *Fix any $\gamma > 0$. For each $i \in [m]$, define*

$$\mathcal{E}_i := \left\{ \frac{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \geq \frac{1}{1+\gamma} \right\}, \tag{2.47}$$

$$\mathcal{D}_i := \left\{ \frac{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}|}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \geq \frac{2+\gamma}{1+\gamma} \right\}, \tag{2.48}$$

$$\text{and} \quad \mathcal{K}_i := \left\{ \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|} \neq \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \right\} \tag{2.49}$$

where $\boldsymbol{h} = \boldsymbol{z} - \boldsymbol{x}$. Under the condition $\|\boldsymbol{h}\| \,/\, \|\boldsymbol{x}\| \leq \rho$, the following inclusion holds for all nonzero $\boldsymbol{z},\, \boldsymbol{h} \in \mathbb{R}^n$

$$\mathcal{E}_i \cap \mathcal{K}_i \subseteq \mathcal{D}_i \cap \mathcal{K}_i. \tag{2.50}$$

Proof. From Fig. 2.1, it is clear that if $\boldsymbol{z} \in \xi_i^2$, then the sign of $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}$ will be different than that of $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}$. The region ξ_i^2 can be readily specified by the conditions that

$$\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|} \neq \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|}$$

and

$$\frac{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}|}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \geq 1 + \frac{1}{1+\gamma} = \frac{2+\gamma}{1+\gamma}.$$

Under our initialization condition $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| \leq \rho$, it is self-evident that $\mathcal{D}_i$ describes two symmetric spherical caps over $\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x} = \psi_i$ with one being ξ_i^2. Hence, it holds that $\mathcal{E}_i \cap \mathcal{K}_i = \xi_i^2 \subseteq \mathcal{D}_i \cap \mathcal{K}_i$. □

To prove (2.46), consider rewriting the truncated gradient in terms of the events defined in Lemma 4:

$$\begin{aligned}\frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|}\right)\boldsymbol{a}_i\mathbb{1}_{\mathcal{E}_i} \\ &= \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\mathbb{1}_{\mathcal{E}_i} - \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|} - \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\right)|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\boldsymbol{a}_i\mathbb{1}_{\mathcal{E}_i}. \end{aligned} \tag{2.51}$$

Using the definitions and properties in Lemma 4, one further arrives at

$$\begin{aligned}\left\langle\frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}),\,\boldsymbol{h}\right\rangle &\geq \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^2\mathbb{1}_{\mathcal{E}_i} - \frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}||\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}|\mathbb{1}_{\mathcal{E}_i\cap\mathcal{K}_i} \\ &\geq \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^2\mathbb{1}_{\mathcal{E}_i} - \frac{2}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}||\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}|\mathbb{1}_{\mathcal{D}_i\cap\mathcal{K}_i} \\ &\geq \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^2\mathbb{1}_{\mathcal{E}_i} - \frac{1+\gamma}{2+\gamma}\cdot\frac{2}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^2\mathbb{1}_{\mathcal{D}_i\cap\mathcal{K}_i} \end{aligned} \tag{2.52}$$

where the last inequality arises from the property $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}| \leq \frac{1+\gamma}{2+\gamma}|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}|$ by the definition of $\mathcal{D}_i$.

Establishing the regularity condition or Prop. 3, boils down to lower bounding the right-hand side of (2.52), namely, to lower bounding the first term and to upper bounding the second one. By the SLLN, the first term in (2.52) approximately gives $\|\boldsymbol{h}\|^2$ as long as our truncation procedure does not eliminate too many generalized gradient components (i.e., summands in the first term). Regarding the second, one would expect its contribution to be small under our initialization condition in (2.27) and as the relative error $\|\boldsymbol{h}\|/\|\boldsymbol{x}\|$ decreases. Specifically, under our initialization, $\mathcal{D}_i$ is provably a rare event, thus eliminating the possibility of the second term exerting a noticeable influence on the first term. Rigorous analyses concerning the two terms are elaborated in Lemma 5 and Lemma 6, whose proofs are provided in Appendix A.4 and Appendix A.5, respectively.

Lemma 5. *Fix $\gamma \geq 1/2$ and $\rho \leq 1/10$, and let $\mathcal{E}_i$ be defined in* (2.47)*. For independent random variables $W \sim \mathcal{N}(0,\,1)$ and $Z \sim \mathcal{N}(0,\,1)$, set*

$$\zeta_1 := 1 - \min\left\{\mathbb{E}\left[\mathbb{1}_{\left\{\left|\frac{1-\rho}{\rho}+\frac{W}{Z}\right|\geq\frac{\sqrt{1.01}}{\rho(1+\gamma)}\right\}}\right],\ \mathbb{E}\left[Z^2\mathbb{1}_{\left\{\left|\frac{1-\rho}{\rho}+\frac{W}{Z}\right|\geq\frac{\sqrt{1.01}}{\rho(1+\gamma)}\right\}}\right]\right\}. \tag{2.53}$$

Then for any $\epsilon > 0$ and any vector $\boldsymbol{h}$ obeying $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| \leq \rho$, the following holds with probability exceeding $1 - 2\mathrm{e}^{-c_5\epsilon^2 m}$:

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^2\mathbb{1}_{\mathcal{E}_i} \geq (1-\zeta_1-\epsilon)\,\|\boldsymbol{h}\|^2 \tag{2.54}$$

provided that $m > (c_6 \cdot \epsilon^{-2}\log\epsilon^{-1})n$ for some universal constants $c_5,\ c_6 > 0$.

To have a sense of how large the quantities involved in Lemma 5 are, when $\gamma = 0.7$ and $\rho = 1/10$, it holds that

$$\mathbb{E}\Big[\mathbb{1}_{\left\{\left|\frac{1-\rho}{\rho}+\frac{W}{Z}\right|\geq\frac{\sqrt{1.01}}{\rho(1+\gamma)}\right\}}\Big] \approx 0.92$$

and

$$\mathbb{E}\Big[Z^2\mathbb{1}_{\left\{\left|\frac{1-\rho}{\rho}+\frac{W}{Z}\right|\geq\frac{\sqrt{1.01}}{\rho(1+\gamma)}\right\}}\Big] \approx 0.99$$

hence leading to $\zeta_1 \approx 0.08$.

Having derived a lower bound for the first term in the right-hand side of (2.52), it remains to deal with the second one.

Lemma 6. *Fix $\gamma > 0$ and $\rho \leq 1/10$, and let $\mathcal{D}_i$, $\mathcal{K}_i$ be defined in* (2.48), (2.49)*, respectively. For any constant $\epsilon > 0$, there exists some universal constants $c_5,\ c_6 > 0$ such that*

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^2\mathbb{1}_{\mathcal{D}_i\cap\mathcal{K}_i} \leq (\zeta_2'+\epsilon)\|\boldsymbol{h}\|^2 \tag{2.55}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_5\epsilon^2 m}$ provided that $m/n > c_6 \cdot \epsilon^{-2}\log\epsilon^{-1}$ for universal constants $c_5,\ c_6 > 0$, where $\zeta_2' = 0.9748\sqrt{\rho\tau/(0.99\tau^2-\rho^2)}$ with $\tau = (2+\gamma)/(1+\gamma)$.

With our TAF default parameters $\rho = 1/10$ and $\gamma = 0.7$, we have $\zeta_2' \approx 0.2463$. Using (2.52), (2.54), and (2.55), choosing m/n exceeding some sufficiently large constant such that $c_0 \leq c_5\epsilon^2$, and denoting $\zeta_2 := 2\zeta_2'(1+\gamma)/(2+\gamma)$, the following holds with probability exceeding $1 - 4\mathrm{e}^{-c_0 m}$

$$\left\langle \boldsymbol{h}, \frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z})\right\rangle \geq (1-\zeta_1-\zeta_2-2\epsilon)\|\boldsymbol{h}\|^2 \tag{2.56}$$

for all $\boldsymbol{x}$ and $\boldsymbol{z}$ such that $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| \leq \rho$ for $0 < \rho \leq 1/10$ and any fixed $\gamma \geq 1/2$. This combined with (2.38) and (2.40) proves Prop. 8 for appropriately chosen $\mu > 0$ and $\lambda > 0$.

To conclude this section, an estimate for the working step size is provided next. Plugging the results of (2.44) and (2.46) into (2.39) suggests that

$$\begin{aligned}\operatorname{dist}^2\left(\boldsymbol{z}-\frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}),\boldsymbol{x}\right) &= \|\boldsymbol{h}\|^2-2\mu\left\langle \boldsymbol{h},\frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z})\right\rangle+\left\|\frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z})\right\|^2 \qquad (2.57)\\ &\le \left\{1-\mu\left[2\left(1-\zeta_1-\zeta_2-2\epsilon\right)-\mu(1+\delta)^4\right]\right\}\|\boldsymbol{h}\|^2\\ &\overset{\triangle}{=} (1-\nu)\|\boldsymbol{h}\|^2, \qquad (2.58)\end{aligned}$$

and also that

$$\lambda = 2\left(1-\zeta_1-\zeta_2-2\epsilon\right)-\mu(1+\delta)^4 \overset{\triangle}{=} \lambda_0$$

in the local regularity condition in (2.38). Clearly, it holds that $0<\lambda<2(1-\zeta_1-\zeta_2)$. Taking ϵ and δ to be sufficiently small, one obtains the feasible range of the step size for TAF

$$\mu \le \frac{2\left(0.99-\zeta_1-\zeta_2\right)}{1.05^4} \overset{\triangle}{=} \mu_0. \qquad (2.59)$$

In particular, under default parameters in Alg. 1, $\mu_0 = 0.8388$ and $\lambda_0 = 1.22$, thus concluding the proof of Thm. 2.

Figure 2.11: The recovered Milky Way Galaxy images after i) truncated spectral initialization (top); ii) orthogonality-promoting initialization (middle); and iii) 100 TAF gradient iterations refining the orthogonality-promoting initialization (bottom).

Chapter 3

Phase Retrieval via Iteratively Reweighted Algorithms

Building upon but going well beyond the scope of previous non-convex paradigms, the present chapter puts forth a novel iterative linear-time procedure, that we term *(iteratively) reweighted amplitude flow* (RAF) here. Our methodology is capable of solving noiseless random quadratic equations exactly, and constructing an estimate of (near)-optimal statistical accuracy from noisy modulus observations. Exactness and accuracy hold with high probability and without any extra assumption on the signal $\boldsymbol{x}$ to be recovered, provided that the ratio m/n of the number of measurements to that of the unknowns exceeds some large constant. The new twist here is to leverage judiciously designed yet conceptually simple (iterative) (re)weighting regularization techniques to enhance existing initializations and also gradient refinements. An informal depiction of our RAF methodology is given in two stages below, with rigorous details deferred to Sec. 3.2.

S1) Weighted maximal correlation initialization: Obtain an initialization $\boldsymbol{z}^0$ maximally correlated with a carefully selected subset $\mathcal{S} \subsetneqq \mathcal{M} := \{1, 2, \ldots, m\}$ of feature vectors $\boldsymbol{a}_i$, whose contributions toward constructing $\boldsymbol{z}^0$ are judiciously weighted by suitable parameters $\{w_i^0 > 0\}_{i \in \mathcal{S}}$; and

S2) Iteratively reweighted "gradient-like" iterations: Loop over $0 \leq t \leq T$:

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \frac{\mu^t}{m} \sum_{i=1}^{m} w_i^t \nabla \ell(\boldsymbol{z}^t; \psi_i) \tag{3.1}$$

for some time-varying weights $w_i^t \geq 0$ that are adapted in time, each depending on the current iterate $\boldsymbol{z}_t$ and the datum $(\boldsymbol{a}_i; \psi_i)$.

Two attributes of our novel methodology are worth highlighting. First, albeit being a variant of the orthogonality-promoting initialization [129], the initialization here [cf. S1)] is distinct in the sense that different importance is attached to each selected datum $(\boldsymbol{a}_i; \psi_i)$, or more precisely, to each selected directional vector $\boldsymbol{a}_i$. Likewise, the gradient flow [cf. S2)] weighs judiciously the search direction suggested by each datum $(\boldsymbol{a}_i; \psi_i)$. In this manner, more accurate and robust initializations as well as more stable overall search directions in the gradient flow stage can be obtained even based only on a relatively limited number of data samples. Moreover, with particular choices of weights w_i^t's (for example, when they take 0 or 1 values), our methodology subsumes as special cases TAF [129], and RWF [147].

3.1 Reweighted Amplitude Flow

This section explains the intuition and the basic principles behind each stage of RAF.

3.1.1 Weighted maximal correlation initialization

For general non-convex iterative heuristics to succeed in finding the global optimum is to seed them with an excellent starting point [69]. In fact, several smart initialization strategies have been advocated for iterative phase retrieval algorithms; see e.g., the spectral [91], [22], truncated spectral [32], and orthogonality-promoting [129] initializations. One promising approach among them is the one proposed in [129], which is robust to outliers [43], and also enjoys better phase transitions than the spectral procedures [81]. To hopefully achieve perfect signal recovery at the information-theoretic limit however, its numerical performance may still need further enhancement. On the other hand, it is intuitive that improving the initialization performance (over state-of-the-art schemes) becomes increasingly challenging as the number of acquired data samples approaches the information-theoretic limit of $m = 2n - 1$.

In this context, we develop a more flexible initialization scheme based on the correlation property (as opposed to orthogonality), in which the added benefit relative to the initialization procedure in [129] is the inclusion of a flexible weighting regularization technique to better balance the useful information exploited in all selected data. In words, we introduce carefully designed weights to the initialization procedure developed in [129]. Similar to other approaches, our strategy entails estimating both the norm $\|\boldsymbol{x}\|$ and the direction $\boldsymbol{x}/\|\boldsymbol{x}\|$. Leveraging the SLLN and the rotational invariance of Gaussian $\boldsymbol{a}_i$ sampling vectors (the latter suffices to assume $\boldsymbol{x} = \|\boldsymbol{x}\|\boldsymbol{e}_1$, with $\boldsymbol{e}_1$ being the first canonical vector in $\mathbb{R}^n$), it is clear that

$$\sum_{i=1}^{m} \psi_i^2 = \sum_{i=1}^{m} \big|\langle \boldsymbol{a}_i, \|\boldsymbol{x}\|\boldsymbol{e}_1 \rangle\big|^2 = \sum_{i=1}^{m} a_{i,1}^2 \|\boldsymbol{x}\|^2 \approx m\|\boldsymbol{x}\|^2 \tag{3.2}$$

whereby $\|\boldsymbol{x}\|$ can be estimated as $\sum_{i=1}^{m} \psi_i^2/m$. This estimate proves very accurate even with a very limited number of data samples, because it is unbiased and tightly concentrated.

The challenge thus lies in accurately estimating the direction of $\boldsymbol{x}$, or seeking a unit vector maximally aligned with $\boldsymbol{x}$, which is a bit tricky. To gain intuition for our initialization strategy, let us first present a variant of the initialization in [129], whose robust counterpart to outlying measurements has been recently discussed in [43]. Note that the larger the modulus ψ_i of the inner-product between $\boldsymbol{a}_i$ and $\boldsymbol{x}$ is, the known design vector $\boldsymbol{a}_i$ is deemed more correlated to the unknown solution $\boldsymbol{x}$, hence bearing useful directional information of $\boldsymbol{x}$. Inspired by this fact and based on available data $\{(\boldsymbol{a}_i; \psi_i)\}_{i=1}^m$, one can sort all (absolute) correlation coefficients $\{\psi_i\}_{i=1}^m$ in an ascending order, to yield ordered coefficients denoted by $0 < \psi_{[m]} \leq \cdots \leq \psi_{[2]} \leq \psi_{[1]}$. Sorting m records takes time proportional to $\mathcal{O}(m \log m)$. Let $\mathcal{S} \subsetneqq \mathcal{M}$ represent the set of selected feature vectors $\boldsymbol{a}_i$ to be used for computing the initialization, which is to be designed next. Fix a priori the cardinality $|\mathcal{S}|$ to some integer on the order of m, say $|\mathcal{S}| := \lfloor 3m/13 \rfloor$. It is then natural to define $\mathcal{S}$ to collect the $\boldsymbol{a}_i$ vectors that correspond to one of the largest $|\mathcal{S}|$ correlation coefficients $\{\psi_{[i]}\}_{1 \leq i \leq |\mathcal{S}|}$, each of which can be thought of as pointing to (roughly) the direction of $\boldsymbol{x}$. Approximating the direction of $\boldsymbol{x}$ thus boils down to finding a vector to maximize its correlation with the subset $\mathcal{S}$ of selected directional vectors $\boldsymbol{a}_i$. Succinctly, the wanted approximation vector can be efficiently found as the solution of

$$\underset{\|\boldsymbol{z}\|=1}{\text{maximize}} \ \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \big|\langle \boldsymbol{a}_i, \boldsymbol{z} \rangle\big|^2 = \boldsymbol{z}^{\mathcal{T}} \Big(\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}} \Big) \boldsymbol{z} \tag{3.3}$$

Upon scaling the solution of (3.3) by the norm estimate $\sum_{i=1}^{m}\psi_i^2/m$ in (3.2) to match the size of $\boldsymbol{x}$, we obtain what we will henceforth refer to as maximal correlation initialization.

As long as $|\mathcal{S}|$ is chosen on the order of m, the maximal correlation method outperforms the spectral ones in [22, 91, 32], and has comparable performance to the orthogonality-promoting method [129]. Its empirical performance around the information-theoretic limit however, is still not the best that we can hope for. Observe that all directional vectors $\{\boldsymbol{a}_i\}_{i\in\mathcal{S}}$ selected for forming the matrix $\overline{\boldsymbol{Y}} := (1/|\mathcal{S}|)\sum_{i\in\mathcal{S}}\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}$ in (3.3) are treated the same in terms of their contributions to constructing the (direction of the) initialization. Nevertheless, according to our starting principle, this ordering information carried by the selected $\boldsymbol{a}_i$ vectors has not been exploited by the initialization scheme in (3.3) (see also [129], [43]). In words, if for selected data $i,j\in\mathcal{S}$, the correlation coefficient of ψ_i with $\boldsymbol{a}_i$ is larger than that of ψ_j with $\boldsymbol{a}_j$, then $\boldsymbol{a}_i$ is deemed more correlated (with $\boldsymbol{x}$) than $\boldsymbol{a}_j$ is, hence bearing more useful information about the wanted direction of $\boldsymbol{x}$. This prompts one to weight more (i.e., attach more importance to) the selected $\boldsymbol{a}_i$ vectors corresponding to larger ψ_i values. Given the ordering information $\psi_{[|\mathcal{S}|]}\leq\cdots\leq\psi_{[2]}\leq\psi_{[1]}$ available from the sorting procedure, a natural way to achieve this goal is by weighting each $\boldsymbol{a}_i$ vector with simple functions of ψ_i, say e.g., taking the weights $w_i^0:=\psi_i^\gamma,\ \forall i\in\mathcal{S}$, with the parameter $\gamma\geq 0$ chosen to maintain the wanted ordering $w_{|\mathcal{S}|}^0\leq\cdots\leq w_{[2]}^0\leq w_{[1]}^0$. In a nutshell, a more flexible initialization scheme, that we refer to as *weighted maximal correlation*, can be summarized as follows

$$\tilde{\boldsymbol{z}}_0 := \arg\max_{\|\boldsymbol{z}\|=1}\ \boldsymbol{z}^{\mathcal{T}}\Big(\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\psi_i^\gamma\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}\Big)\boldsymbol{z}. \tag{3.4}$$

The upshot of (3.4) is that the objective can be efficiently minimized in time proportional to $\mathcal{O}(n|\mathcal{S}|)$ by means of the power method or the Lanczos algorithm [101]. The new initialization can be obtained after scaling $\tilde{\boldsymbol{z}}^0$ from (3.4) with the estimate of its norm, to obtain $\boldsymbol{z}^0 := (\sum_{i=1}^{m}\psi_i^2/m)\tilde{\boldsymbol{z}}^0$. By default, we take $\gamma := 1/2$ in all reported numerical implementations, yielding $w_i^0 := \sqrt{|\langle\boldsymbol{a}_i,\boldsymbol{x}\rangle|}$ for all $i\in\mathcal{S}$.

Regarding the initialization procedure in (3.4), we next highlight two features, while details and theoretical performance guarantees are provided in Sec. 3.2:

F1) The weights $\{w_i^0\}$ in the maximal correlation scheme enable leveraging useful information that each feature vector $\boldsymbol{a}_i$ may bear regarding the direction of $\boldsymbol{x}$.

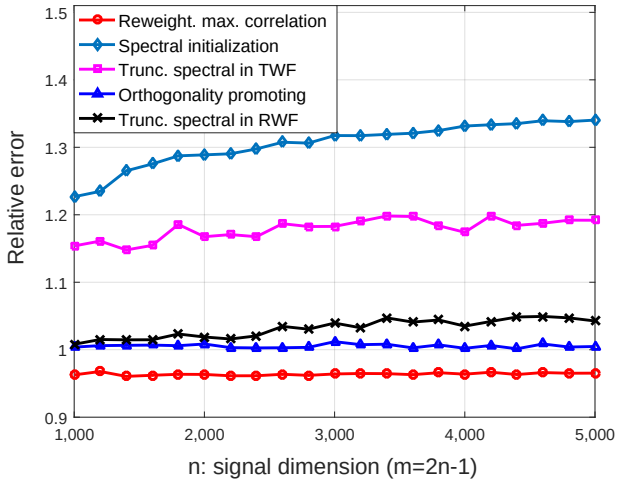

Figure 3.1: Relative initialization error for the real Gaussian model.

F2) Taking $w_i^0 := \psi_i^\gamma$ for all $i \in \mathcal{S}$ and 0 otherwise, (3.4) can be equivalently rewritten as

$$\tilde{\boldsymbol{z}}^0 := \arg\max_{\|\boldsymbol{z}\|=1} \; \boldsymbol{z}^{\mathcal{T}} \Big(\frac{1}{m} \sum_{i=1}^{m} w_i^0 \boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}} \Big) \boldsymbol{z} \tag{3.5}$$

which subsumes existing initialization schemes with particular weight selections; e.g., the "plain-vanilla" spectral initialization in [91, 22] is recovered by choosing $\mathcal{S} := \mathcal{M}$, and $w_i^0 := \psi_i^2$, for $i = 1, \ldots, m$.

Figure 3.1 depicts the performance of the proposed initialization relative to several state-of-the-art strategies. It is clear that our initialization is: i) consistently better than the state-of-the-art; and, ii) stable as the signal dimension n grows, which is in sharp contrast to the instability encountered by the spectral ones [91, 22, 32]. It is also worth stressing that about 5% empirical advantage is shown over the best in [129] at the challenging information-theoretic benchmark, which is nontrivial, and constitutes one of the main advantages of RAF. This numerical advantage becomes increasingly pronounced as m/n of the number of equations to the unknowns grows. This suggests that our proposed initialization procedure may be combined with other iterative phase retrieval approaches to improve their numerical performance.

3.1.2 Adaptively reweighted gradient flow

For independent data adhering to the real Gaussian model, the direction that TAF moves along in stage S2) presented earlier is given by the following (generalized) gradient [129], [35]

$$\frac{1}{m}\sum_{i\in\mathcal{T}}\nabla\ell(\boldsymbol{z};\psi_i)=\frac{1}{m}\sum_{i\in\mathcal{T}}\Big(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}-\psi_i\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|}\Big)\boldsymbol{a}_i \tag{3.6}$$

where the dependence on the iterate count t is neglected for notational brevity.

Unfortunately, the (negative) gradient of the average in (3.6) may not point towards the true $\boldsymbol{x}$, unless the current iterate $\boldsymbol{z}$ is already very close to $\boldsymbol{x}$. As a consequence, moving along such a descent direction may not drag $\boldsymbol{z}$ closer to $\boldsymbol{x}$. To see this, consider an initial guess $\boldsymbol{z}_0$ that has already been in a basin of attraction (i.e., a region within which there is only a unique stationary point) of $\boldsymbol{x}$. Certainly, there are summands $(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}-\psi_i\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|)\boldsymbol{a}_i$ in (3.6), that could give rise to "bad/misleading" search directions due to the erroneously estimated signs $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|\neq\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ in (3.6) [129]. Those gradients as a whole may drag $\boldsymbol{z}$ away from $\boldsymbol{x}$, and hence out of the basin of attraction. Such an effect becomes increasingly severe as the number m of acquired examples approaches the information-theoretic limit of $2n-1$, thus rendering past approaches less effective in this case. Although this issue is somewhat remedied by TAF with a truncation procedure, its efficacy is limited due to misses of bad gradients and mis-rejections of meaningful ones at the information-theoretic limit.

To address this challenge, our reweighted gradient flow effecting suitable search directions from almost all acquired data samples $\{(\boldsymbol{a}_i;\psi_i)\}_{i=1}^m$ will be adopted in a (timely) adaptive fashion; that is,

$$\boldsymbol{z}^{t+1}=\boldsymbol{z}^t-\mu^t\nabla\ell_{\rm rw}(\boldsymbol{z}^t;\psi_i),\quad t=0,\,1,\,\ldots \tag{3.7}$$

The reweighted gradient $\nabla\ell_{\rm rw}(\boldsymbol{z})$ evaluated at the current point $\boldsymbol{z}^t$ is given as

$$\nabla\ell_{\rm rw}(\boldsymbol{z}):=\frac{1}{m}\sum_{i=1}^m w_i\nabla\ell(\boldsymbol{z};\psi_i) \tag{3.8}$$

for suitable weights $\{w_i\}_{i=1}^m$ to be designed shortly.

To that end, we observe that the truncation criterion $\mathcal{T}:=\{1\le i\le m:|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\ge\alpha\}$ with some given parameter $\alpha>0$ suggests to include only gradients associated with $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|$ of relatively large sizes. This is because gradients of sizable $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ offer reliable and

meaningful directions pointing to the true $\boldsymbol{x}$ with large probability [129]. As such, the ratio $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ can be viewed as a confidence score on the reliability or meaningfulness of the corresponding gradient $\nabla\ell(\boldsymbol{z};\psi_i)$. Recognizing that confidence can vary, it is natural to distinguish the contributions that different gradients make to the overall search direction. An easy way is to attach large weights to the reliable gradients, and small weights to the spurious ones. Assume without loss of generality that $0 \leq w_i \leq 1$ for all $1 \leq i \leq m$; otherwise, lump the normalization factor achieving this into the learning rate μ^t. Building upon this observation and leveraging the gradient reliability confidence score $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$, the weight per gradient $\nabla\ell(\boldsymbol{z};\psi_i)$ in our proposed RAF algorithm is

$$w_i := \frac{1}{1+\beta_i/(|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|)}, \quad i = 1,\, 2,\, \ldots,\, m \tag{3.9}$$

where $\{\beta_i > 0\}_{i=1}^m$ are some pre-selected parameters.

Regarding the weighting criterion in (3.9), three remarks are in order.

Remark 3. The weights $\{w_i^t\}_{i=1}^m$ are time adapted to the iterate $\boldsymbol{z}^t$. One can also interpret the reweighted gradient flow $\boldsymbol{z}^{t+1}$ in (3.7) as performing a single gradient step to minimize the smooth reweighted loss $(1/m)\sum_{i=1}^m w_i^t\ell(\boldsymbol{z};\psi_i)$ with starting point $\boldsymbol{z}^t$; see also [29] for related ideas successfully exploited in the iteratively reweighted least-squares approach to compressive sampling.

Remark 4. The larger the confidence score $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ is, the larger the corresponding weight w_i will be. More importance will be then attached to reliable gradients than to spurious ones. Gradients from almost all data are accounted for, which is in contrast to [129], where withdrawn gradients do not contribute the information they carry.

Remark 5. At the points $\{\boldsymbol{z}\}$ where $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} = 0$ for some datum $i \in \mathcal{M}$, the i-th weight will be $w_i = 0$. In other words, the squared losses $\ell(\boldsymbol{z};\psi_i)$ in (2.2) that are non-smooth at points $\boldsymbol{z}$ will be eliminated, to prevent their contribution to the reweighted gradient update in (3.7). This simplifies the convergence analysis of RAF considerably because it does not have to cope with the non-smoothness of the objective function in (2.2).

Having elaborated on the two stages, RAF can be readily summarized in Alg. 2.

Algorithm 2 Reweighted Amplitude Flow (RAF)

1: **Input:** Data $\{(\boldsymbol{a}_i;\psi_i\}_{i=1}^m$; maximum number of iterations T; step sizes $\mu^t = 2/6$ and weighting parameters $\beta_i = 10/5$ for real/complex Gaussian models; subset cardinality $|\mathcal{S}| = \lfloor 3m/13 \rfloor$, and exponent $\gamma = 0.5$.

2: **Construct** $\mathcal{S}$ to include indices associated with the $|\mathcal{S}|$ largest entries among $\{\psi_i\}_{i=1}^m$.

3: **Initialize** $\boldsymbol{z}^0 := \sqrt{\sum_{i=1}^m \psi_i^2/m}\,\tilde{\boldsymbol{z}}^0$ with $\tilde{\boldsymbol{z}}^0$ being the unit-norm principal eigenvector of

$$\frac{1}{m}\sum_{i=1}^m w_i^0 \boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}, \;\; \text{where} \;\; w_i^0 := \begin{cases} \psi_i^\gamma, & i \in \mathcal{S} \subseteq \mathcal{M} \\ 0, & \text{otherwise.} \end{cases}$$

4: **Loop: for** $t = 0$ **to** $T-1$

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \frac{\mu^t}{m}\sum_{i=1}^m w_i^t \Big(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t - \psi_i \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_i \tag{3.10}$$

where $w_i^t := \frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|/\psi_i}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|/\psi_i+\beta_i}$ for all $1 \le i \le m$.

5: **Output:** $\boldsymbol{z}^T$.

3.1.3 Parameters of the algorithm

To optimize the empirical performance and facilitate numerical implementations, the choice of pertinent RAF parameters is outlined here. For the four RAF parameters, our theory and experiments are based on: i) $|\mathcal{S}|/m \le 0.25$; ii) $0 \le \beta_i \le 10$ for all $1 \le i \le m$; and, iii) $0 \le \gamma \le 1$. For convenience, a constant step size $\mu^t \equiv \mu > 0$ is suggested, but other step size rules such as backtracking line search with the reweighted objective would work as well. As will be formalized in Sec. 3.2, RAF converges if the constant μ is not too large, with the upper bound depending in part on the selection of $\{\beta_i\}_{i=1}^m$.

In the numerical tests presented in Secs. 3.1 and 3.3, we take $|\mathcal{S}| := \lfloor 3m/13 \rfloor$, $\beta_i \equiv \beta := 10$, $\gamma := 0.5$, and $\mu := 2$ (larger step sizes can be afforded for larger m/n values).

3.2 Main Results

Our main results stated next establish exact recovery under the real Gaussian model, whose proof is postponed to Sec. 3.4 for readability. Our RAF methodology however, can be generalized readily to the complex Gaussian and CDP models.

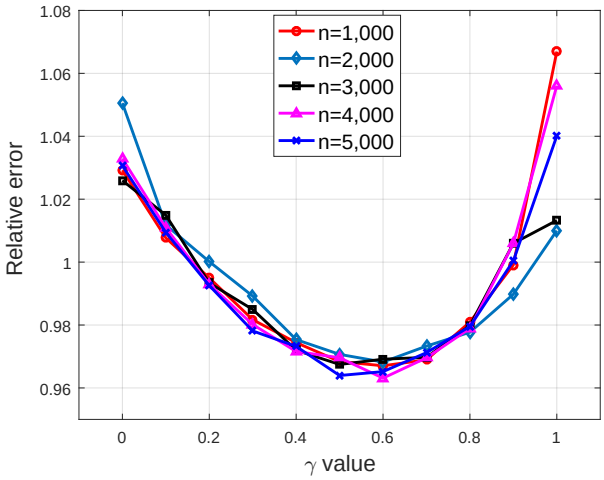

Figure 3.2: Relative error versus γ for the proposed initialization and $m = 2n - 1$ fixed using the real Gaussian model.

Theorem 2 (Exact recovery)**.** *Consider m noiseless measurements $\boldsymbol{\psi} = |\boldsymbol{A}\boldsymbol{x}|$ for an arbitrary signal $\boldsymbol{x} \in \mathbb{R}^n$. If $m \geq c_0|\mathcal{S}| \geq c_1 n$ with $|\mathcal{S}|$ being the pre-selected subset cardinality in the initialization step and the learning rate $\mu \leq \mu_0$, then with probability at least $1 - c_3 \mathrm{e}^{-c_2 m}$, the RAF estimates $\boldsymbol{z}^t$ in Alg. 2 obey*

$$\operatorname{dist}(\boldsymbol{z}^t, \boldsymbol{x}) \leq \frac{1}{10}(1-\nu)^t \|\boldsymbol{x}\|, \quad t = 0, 1, \ldots \tag{3.11}$$

where c_0, c_1, c_2, $c_3 > 0$, $0 < \nu < 1$, and $\mu_0 > 0$ are certain numerical constants depending on the choice of algorithmic parameters $|\mathcal{S}|$, β, γ, and μ.

According to Thm. 2, a few interesting properties of RAF are worth highlighting. To start, RAF recovers the true solution exactly with high probability whenever the ratio m/n of the number of equations to the unknowns exceeds some numerical constant. Expressed differently, RAF achieves the information-theoretic optimal order of sample complexity, which is consistent with the state-of-the-art including TWF [32], TAF [129], and RWF [147]. Notice that the error contraction in (3.11) also holds at $t = 0$, namely $\operatorname{dist}(\boldsymbol{z}^0, \boldsymbol{x}) \leq \|\boldsymbol{x}\|/10$, therefore providing theoretical performance guarantees for the proposed initialization strategy (cf. Step 3 of Alg. 2). Moreover, starting from this initial estimate, RAF converges exponentially fast to the true solution $\boldsymbol{x}$. In other words, to reach any ϵ-relative solution accuracy (i.e., $\operatorname{dist}(\boldsymbol{z}^T, \boldsymbol{x}) \leq \epsilon\|\boldsymbol{x}\|$),

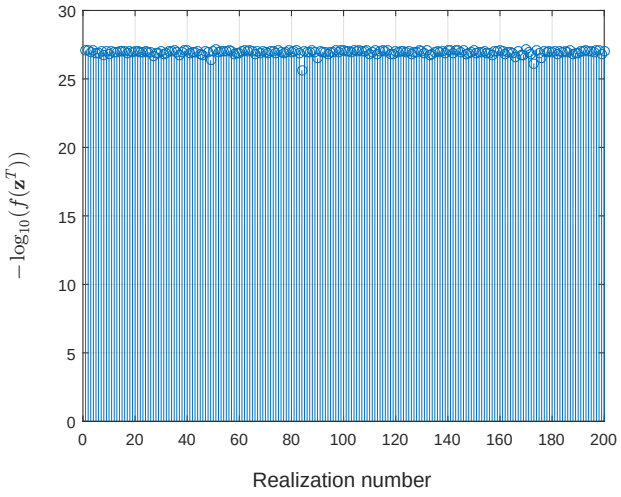

Figure 3.3: Function value $L(\boldsymbol{z}^T)$ evaluated at the returned RAF estimate $\boldsymbol{z}^T$ for 200 trials with $n = 2,000$ and $m = 2n - 1 = 3,999$.

it suffices to run at most $T = \mathcal{O}(\log 1/\epsilon)$ RAF iterations in Step 4 of Alg. 2. This in conjunction with the per-iteration complexity $\mathcal{O}(mn)$ (namely, the complexity of one reweighted gradient update in (B.26)) confirms that RAF solves exactly a quadratic system in time $\mathcal{O}(mn \log 1/\epsilon)$, which is linear in $\mathcal{O}(mn)$, the time required by the processor to read the entire data $\{(\boldsymbol{a}_i; \psi_i)\}_{i=1}^m$. Given the fact that the initialization stage can be performed in time $\mathcal{O}(n|\mathcal{S}|)$ and $|\mathcal{S}| < m$, the overall linear-time complexity of RAF is order-optimal.

3.3 Numerical Experiments

Our theoretical findings about RAF have been corroborated with comprehensive numerical experiments, a sample of which are presented next. Performance of RAF is evaluated relative to (T)WF [22, 32], RWF [147], and TAF [129]. Each scheme obtained its initial guess based on 200 power or Lanczos iterations, followed by a sequence of $T = 2,000$ (which can be set smaller as m/n grows away from the limit of 2) gradient-type iterations. For reproducibility, the Matlab code of RAF is publicly available at `https://gangwg.github.io/RAF/`.

To show the power of RAF in the high-dimensional regime, the function value $L(\boldsymbol{z})$ in (2.2) evaluated at the returned estimate $\boldsymbol{z}^T$ (cf. Step 5 of Alg. 2) after 200 MC realizations is plotted (in negative logarithmic scale) in Fig. 3.3, where the number of simulated noiseless

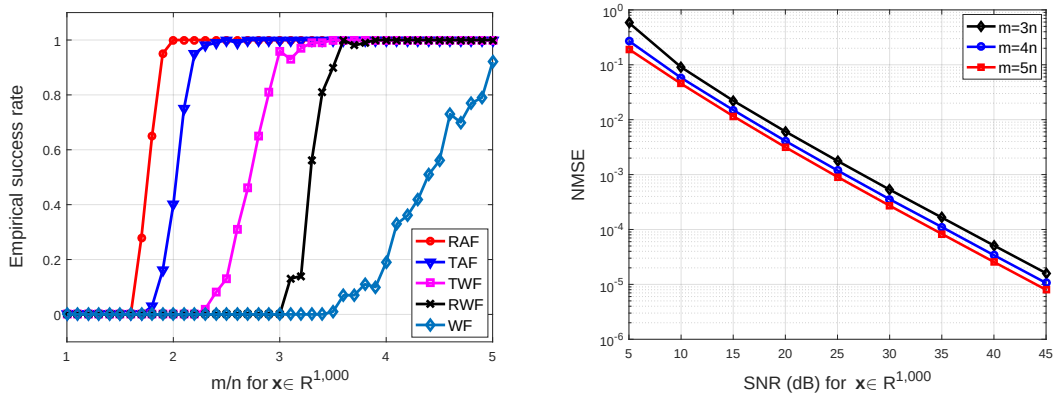

Figure 3.4: Real Gaussian model. Left: Empirical success rate; Right: NMSE vs. SNR.

measurements was set to be the information-theoretic limit, namely $m = 2n - 1 = 3,999$ for $n = 2,000$. It is evident that our proposed RAF approach returns a solution of function value $L(\boldsymbol{z}^T)$ smaller than 10^{-25} in all 200 independent realizations even at this challenging information-theoretic limit condition. To the best of our knowledge, RAF is the first algorithm that empirically reconstructs any high-dimensional (say e.g., $n \geq 1,500$) signals exactly from an optimal number of random quadratic equations.

The left panel in Fig. 3.4 further compares the empirical success rate of five schemes with the signal dimension being fixed at $n = 1,000$ while m/n increasing by 0.1 from 1 to 5. As clearly depicted by the plots, our RAF (color coded red) enjoys markedly improved performance over its competing alternatives. Moreover, it also achieves 100% signal recovery as soon as m is about $2n$, where the others do not show perfect recovery. To numerically demonstrate the stability and robustness of RAF in the presence of additive noise, the right panel in Fig. 3.4 examines $\mathrm{NMSE} := \mathrm{dist}^2(\boldsymbol{z}^T, \boldsymbol{x})/\|\boldsymbol{x}\|^2$ as a function of the SNR for m/n taking values $\{3, 4, 5\}$. The noise model $\psi_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle| + \eta_i$ with $\boldsymbol{\eta} := [\eta_i]_{i=1}^m \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m)$ was simulated, where σ^2 was set such that certain $\mathrm{SNR} := 10\log_{10}(\|\boldsymbol{A}\boldsymbol{x}\|^2/m\sigma^2)$ values were achieved. For all choices of m (as small as $3n$ which is nearly minimal), the numerical experiments illustrate that the NMSE scales inversely proportional to the SNR, which corroborates the stability of our RAF approach.

3.4 Proofs

To prove Thm. 2, this section establishes a few lemmas and the main ideas, whereas technical details are postponed to the Appendix to facilitate readability. It is clear from Alg. 2 that the weighted maximal correlation initialization (cf. Step 3) and the reweighted gradient flow (cf. Step 4) distinguish themselves from those procedures in (T)WF [22, 32], TAF [129], and RWF [147]. Hence, new proof techniques to cope with the weighting in both the initialization and the gradient flow, as well as the non-smoothness and non-convexity of the amplitude-based least-squares functional are required.

The proof of Thm. 2 consists of two parts: Sec. 3.4.1 below asserts guaranteed theoretical performance of the proposed initialization, which essentially achieves any given constant relative error as soon as $m \geq c_1 n$ for some constant $c_1 > 0$. It is worth mentioning that we reserve c and its subscripted versions for absolute constants, even though their values may vary with the context. Under the sample complexity of order $\mathcal{O}(n)$, Sec. 3.4.2 further shows that RAF converges to the true signal $\boldsymbol{x}$ exponentially fast whenever the initial estimate lands within a relatively small-size neighborhood of $\boldsymbol{x}$ defined by $\mathrm{dist}(\boldsymbol{z}^0, \boldsymbol{x}) \leq (1/10)\|\boldsymbol{x}\|$.

3.4.1 Initialization performance

This section is devoted to establishing analytical guarantees for the novel initialization procedure, which is summarized in the following proposition.

Proposition 4. *For an arbitrary $\boldsymbol{x} \in \mathbb{R}^n$, consider the noiseless measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$, $1 \leq i \leq m$. If $m \geq c_0|\mathcal{S}| \geq c_1 n$, then with probability exceeding $1 - c_3 \mathrm{e}^{-c_2 m}$, the initial guess $\boldsymbol{z}^0$ obtained by the weighted maximal correlation method in Step 3 of Alg. 2 satisfies*

$$\mathrm{dist}(\boldsymbol{z}^0, \boldsymbol{x}) \leq \rho \|\boldsymbol{x}\| \tag{3.12}$$

for $\rho = 1/10$ (or any sufficiently small positive number). Here, c_0, c_1, c_2, $c_3 > 0$ are some absolute constants.

Since the norm $\|\boldsymbol{x}\| = 1$ is assumed known, the weighted maximal correlation initialization in Step 3 finds the initial estimate $\boldsymbol{z}^0 = \tilde{\boldsymbol{z}}^0$ (the scaling factor is the exactly known norm 1 in

this case) as the principal eigenvector of

$$\boldsymbol{Y} := \frac{1}{|\mathcal{S}|}\boldsymbol{B}^{\mathcal{T}}\boldsymbol{B} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \psi_i^{\gamma}\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}} \tag{3.13}$$

where $\boldsymbol{B} := \big[\psi_i^{\gamma/2}\boldsymbol{a}_i\big]_{i\in\mathcal{S}}$ is an $|\mathcal{S}| \times n$ matrix, and $\mathcal{S} \subsetneqq \{1, 2, \ldots, m\}$ includes the indices of the $|\mathcal{S}|$ largest entities among all modulus data $\{\psi_i\}_{i=1}^m$. The following result is a modification of [129, Lemma 1], which is key to proving Prop. 4.

Lemma 7. *Consider m noiseless measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$, $1 \le i \le m$. For an arbitrary $\boldsymbol{x} \in \mathbb{R}^n$ of unit norm, the next result holds for all unit-norm vectors $\boldsymbol{u} \in \mathbb{R}^n$ perpendicular to $\boldsymbol{x}$; that is, for all $\boldsymbol{u} \in \mathbb{R}^n$ satisfying $\boldsymbol{u}^{\mathcal{T}}\boldsymbol{x} = 0$ and $\|\boldsymbol{u}\| = 1$, we have*

$$\frac{1}{2}\|\boldsymbol{x}\boldsymbol{x}^{\mathcal{T}} - \boldsymbol{z}^0(\boldsymbol{z}^0)^{\mathcal{T}}\|_F^2 \le \frac{\|\boldsymbol{B}\boldsymbol{u}\|^2}{\|\boldsymbol{B}\boldsymbol{x}\|^2} \tag{3.14}$$

where $\boldsymbol{z}^0 = \tilde{\boldsymbol{z}}^0$ is given by

$$\tilde{\boldsymbol{z}}^0 := \arg\max_{\|\boldsymbol{z}\|=1} \frac{1}{|\mathcal{S}|}\boldsymbol{z}^{\mathcal{T}}\boldsymbol{B}^{\mathcal{T}}\boldsymbol{B}\boldsymbol{z}. \tag{3.15}$$

Let us start with the proof of Prop. 4. The first step consists in upper-bounding the quantity on the right-hand-side of (3.14). This involves upper bounding its numerator, and lower bounding its denominator, tasks summarized in Lemmas 8 and 9, whose proofs are deferred to Appendices B.1 and B.2, accordingly.

Lemma 8. *In the setting of Lemma 7, if $|\mathcal{S}|/n \ge c_4$, then the inequality*

$$\|\boldsymbol{B}\boldsymbol{u}\|^2 \le 1.01\sqrt{2^{\gamma}/\pi}\,\Gamma(\gamma+1/2)|\mathcal{S}| \tag{3.16}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_5 n}$, where $\Gamma(\cdot)$ is the Gamma function, and c_4, c_5 are certain universal constants.

Lemma 9. *In the setting of Lemma 7, the following holds with probability exceeding $1 - \mathrm{e}^{-c_6 m}$*

$$\|\boldsymbol{B}\boldsymbol{x}\|^2 \ge 0.99|\mathcal{S}|\big[1 + \log(m/|\mathcal{S}|)\big] \ge 0.99 \times 1.14^{\gamma}|\mathcal{S}|\big[1 + \log(m/|\mathcal{S}|)\big] \tag{3.17}$$

provided that $m \ge c_0|\mathcal{S}| \ge c_1 n$ for some absolute constants c_0, c_1, $c_6 > 0$.

Taking together, the upper bound in (3.16) and the lower bound in (3.17), one arrives at

$$\frac{\|\boldsymbol{B}\boldsymbol{u}\|^2}{\|\boldsymbol{B}\boldsymbol{x}\|^2} \leq \frac{C}{1+\log(m/|\mathcal{S}|)} \triangleq \kappa \tag{3.18}$$

where $C := 1.02 \times 1.14^{-\gamma}\sqrt{2^{\gamma}/\pi}\Gamma(\gamma+1/2)$, and (3.18) holds with probability at least $1 - 2\mathrm{e}^{-c_5 n} - \mathrm{e}^{-c_6 m}$, with the proviso that $m \geq c_0|\mathcal{S}| \geq c_1 n$. Since $m = \mathcal{O}(n)$, one can rewrite the probability as $1 - c_3\mathrm{e}^{-c_2 m}$ for certain constants $c_2,\, c_3 > 0$. To have a sense of the size of C, taking our default value $\gamma = 0.5$ for instance gives rise to $C = 0.7854$.

3.4.2 Exact Phase Retrieval from Noiseless Data

It has been demonstrated that the initial estimate $\boldsymbol{z}^0$ obtained by means of the weighted maximal correlation initialization strategy has at most a constant relative error to $\boldsymbol{x}$, i.e., $\mathrm{dist}(\boldsymbol{z}^0,\, \boldsymbol{x}) \leq (1/10)\|\boldsymbol{x}\|$. We demonstrate in the following that starting from such an initial estimate, the RAF iterates (in Step 4 of Alg. 2) converge at a linear rate to $\boldsymbol{x}$; that is, $\mathrm{dist}(\boldsymbol{z}^t,\, \boldsymbol{x}) \leq (1/10)c^t\|\boldsymbol{x}\|$ for some constant $0 < c < 1$ depending on the step size $\mu > 0$, the weighting parameter β, and the data $\{(\boldsymbol{a}_i; \psi_i)\}_{i=1}^m$. This constitutes the second part of the proof of Thm. 2. Toward this end, it suffices to show that the iterative update of RAF is locally contractive within a relatively small neighboring region of the true $\boldsymbol{x}$. Instead of directly coping with the moments in the weights, we establish a conservative result based on [129] and [147]. Recall first that our gradient flow uses the reweighted gradient

$$\nabla\ell_{\mathrm{rw}}(\boldsymbol{z}) := \frac{1}{m}\sum_{i=1}^{m} w_i\Big(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|}\Big)\boldsymbol{a}_i \tag{3.19}$$

with

$$w_i = \frac{1}{1+\beta/(|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|)}, \quad 1 \leq i \leq m$$

in which the dependence on the iterate index t is ignored for notational brevity.

Proposition 5 (Local error contraction)**.** *For an arbitrary $\boldsymbol{x} \in \mathbb{R}^n$, consider m noise-free measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$, $1 \leq i \leq m$. There exist some numerical constants $c_1,\, c_2,\, c_3 > 0$, and $0 < \nu < 1$ such that the following holds with probability exceeding $1 - c_3\mathrm{e}^{-c_2 m}$:*

$$\mathrm{dist}^2(\boldsymbol{z} - \mu\nabla\ell_{\mathrm{rw}}(\boldsymbol{z}),\, \boldsymbol{x}) \leq (1-\nu)\mathrm{dist}^2(\boldsymbol{z},\, \boldsymbol{x}) \tag{3.20}$$

for all $\boldsymbol{x},\ \boldsymbol{z} \in \mathbb{R}^n$ *obeying* $\mathrm{dist}(\boldsymbol{z},\ \boldsymbol{x}) \leq (1/10)\|\boldsymbol{x}\|$*, provided that* $m \geq c_1 n$ *and the constant step size* $\mu \leq \mu_0$*, where the numerical constant* μ_0 *depends on the parameter* $\beta > 0$ *and data* $\{(\boldsymbol{a}_i; \psi_i)\}_{i=1}^m$.

Proposition 5 suggests that the distance of RAF's successive iterates to the global optimum $\boldsymbol{x}$ decreases monotonically once the algorithm's iterate $\boldsymbol{z}^t$ enters a small neighboring region around $\boldsymbol{x}$. Expressed differently, RAF's iterates will stay within the region and will be attracted towards $\boldsymbol{x}$ exponentially fast as soon as they land within the basin of attraction. To substantiate Prop. 5, recall the useful analytical tool of the local regularity condition [22], which plays a key role in establishing linear convergence of iterative procedures to the global optimum in [22], [32], [129].

For RAF, the reweighted gradient $\nabla \ell_{\mathrm{rw}}(\boldsymbol{z})$ in (3.19) is said to obey the local regularity condition (LRC), denoted as $\mathrm{LRC}(\mu, \lambda, \epsilon)$ for some constant $\lambda > 0$, if the next inequality

$$\langle \nabla \ell_{\mathrm{rw}}(\boldsymbol{z}),\ \boldsymbol{h} \rangle \geq \frac{\mu}{2} \|\nabla \ell_{\mathrm{rw}}(\boldsymbol{z})\|^2 + \frac{\lambda}{2} \|\boldsymbol{h}\|^2 \tag{3.21}$$

holds for all $\boldsymbol{z} \in \mathbb{R}^n$ such that $\|\boldsymbol{h}\| = \|\boldsymbol{z} - \boldsymbol{x}\| \leq \epsilon \|\boldsymbol{x}\|$ for some constant $0 < \epsilon < 1$.

Letting $\boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$, manipulations in conjunction with (3.21) confirms that

$$\begin{aligned}
\mathrm{dist}^2(\boldsymbol{z} - \mu \nabla \ell_{\mathrm{rw}}(\boldsymbol{z}),\ \boldsymbol{x}) &= \|\boldsymbol{z} - \mu \nabla \ell_{\mathrm{rw}}(\boldsymbol{z}) - \boldsymbol{x}\|^2 \\
&= \|\boldsymbol{h}\|^2 - 2\mu \langle \boldsymbol{h}, \nabla \ell_{\mathrm{rw}}(\boldsymbol{z}) \rangle + \|\mu \nabla \ell_{\mathrm{rw}}(\boldsymbol{z})\|^2 \qquad (3.22) \\
&\leq \|\boldsymbol{h}\|^2 - 2\mu \left(\frac{\mu}{2} \|\nabla \ell_{\mathrm{rw}}(\boldsymbol{z})\|^2 + \frac{\lambda}{2} \|\boldsymbol{h}\|^2 \right) + \|\mu \nabla \ell_{\mathrm{rw}}(\boldsymbol{z})\|^2 \\
&= (1 - \lambda\mu) \|\boldsymbol{h}\|^2 \\
&= (1 - \lambda\mu)\, \mathrm{dist}^2(\boldsymbol{z},\ \boldsymbol{x}) \qquad (3.23)
\end{aligned}$$

for all points $\boldsymbol{z}$ adhering to $\|\boldsymbol{h}\| \leq \epsilon \|\boldsymbol{x}\|$. It is evident that if $\mathrm{LRC}(\mu, \lambda, \epsilon)$ can be established for RAF, our goal of proving the local error contraction in (3.20) follows straightforwardly upon setting $\nu := \lambda\mu$.

Proof of the local regularity condition in (3.21)

The first step to proving the local regularity condition in (3.21) is to control the size of the reweighted gradient $\nabla \ell_{\mathrm{rw}}(\boldsymbol{z})$; that is, to upper bound the last term in (3.22). To start, rewrite the

reweighted gradient in a compact matrix-vector representation

$$\nabla \ell_{\rm rw}(\boldsymbol{z}) = \frac{1}{m}\sum_{i=1}^{m} w_i\Big(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|}\Big)\boldsymbol{a}_i \stackrel{\triangle}{=} \frac{1}{m}{\rm dg}(\boldsymbol{w})\boldsymbol{A}\boldsymbol{v} \tag{3.24}$$

where ${\rm dg}(\boldsymbol{w}) \in \mathbb{R}^{n\times n}$ is a diagonal matrix holding entries of $\boldsymbol{w} := [w_1 \ \cdots \ w_m]^{\mathcal{T}} \in \mathbb{R}^m$ on its main diagonal, and $\boldsymbol{v} := [v_1 \ \cdots \ v_m]^{\mathcal{T}} \in \mathbb{R}^m$ with $v_i := \boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|}$. Based on the definition of the induced matrix 2-norm (namely, the spectral norm), it is easy to check that

$$\|\nabla \ell_{\rm rw}(\boldsymbol{z})\| = \left\|\frac{1}{m}{\rm dg}(\boldsymbol{w})\boldsymbol{A}\boldsymbol{v}\right\| \le \frac{1}{m}\|{\rm dg}(\boldsymbol{w})\|\cdot\|\boldsymbol{A}\|\cdot\|\boldsymbol{v}\| \le \frac{1+\delta'}{\sqrt{m}}\|\boldsymbol{v}\| \tag{3.25}$$

where we have used the inequalities $\|{\rm dg}(\boldsymbol{w})\| \le 1$ due to $w_i \le 1$ for all $1 \le i \le m$, and $\|\boldsymbol{A}\| \le (1+\delta')\sqrt{m}$ for some constant $\delta' > 0$ according to [119, Thm. 5.32], provided that m/n is sufficiently large.

The task therefore remains to bound $\|\boldsymbol{v}\|$ in (3.25), which is addressed next. To this end, notice that

$$\|\boldsymbol{v}\|^2 \le \sum_{i=1}^{m}\left(|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}| - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\right)^2 \le \sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} - \boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}\right)^2 \le (1+\delta'')^2 m\|\boldsymbol{h}\|^2 \tag{3.26}$$

for some numerical constant $\delta'' > 0$, where the last can be obtained using [23, Lemma 3.1], and which holds with probability at least $1 - {\rm e}^{-c_2 m}$ as long as $m > c_1 n$ holds true.

Combing (3.25) with (3.26) and taking $\delta > 0$ larger than the constant $(1+\delta')(1+\delta'') - 1$, the size of $\nabla \ell_{\rm rw}(\boldsymbol{z})$ can be bounded as

$$\|\nabla \ell_{\rm rw}(\boldsymbol{z})\| \le (1+\delta)\|\boldsymbol{h}\| \tag{3.27}$$

which holds with probability $1 - {\rm e}^{-c_2 m}$, with a proviso that m/n exceeds some numerical constant $c_7 > 0$. This result indeed asserts that the reweighted gradient of $L(\boldsymbol{z})$ or the search direction employed in our RAF algorithm is well behaved, implying that the function value along the iterates does not change too much.

In order to prove the LRC, it suffices to show that $\nabla \ell_{\rm rw}(\boldsymbol{z})$ ensures sufficient descent, that is, there exists a numerical constant $c > 0$ such that along the search direction $\nabla \ell_{\rm rw}(\boldsymbol{z})$ the

following uniform lower bound holds

$$\langle \nabla \ell_{\mathrm{rw}}(\boldsymbol{z}),\, \boldsymbol{h} \rangle \geq c \|\boldsymbol{h}\|^2 \tag{3.28}$$

which will be addressed next. Formally, this can be summarized in the following proposition, whose proof is deferred to Appendix B.3.

Proposition 6. *For the noise-free measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$, $1 \leq i \leq m$, and any fixed sufficiently small constant $\epsilon > 0$. There exist some numerical constants $c_1,\, c_2,\, c_3 > 0$ such that the following holds with probability at least $1 - c_3 \mathrm{e}^{-c_2 m}$*

$$\langle \boldsymbol{h}, \nabla \ell_{\mathrm{rw}}(\boldsymbol{z}) \rangle \geq \zeta_3 \|\boldsymbol{h}\|^2 \tag{3.29}$$

for all $\boldsymbol{x},\, \boldsymbol{z} \in \mathbb{R}^n$ obeying $\|\boldsymbol{h}\| \leq (1/10)\|\boldsymbol{x}\|$, provided that $m/n > c_1$, and that $\beta \geq 0$ is small enough. Here, $\zeta_3 := \frac{1-\zeta_1-\epsilon}{1+\beta(1+\eta)} - 2(\zeta_2 + \epsilon) - \frac{2(0.1271-\zeta_2+\epsilon)}{1+\beta/k}$.

Taking the results in (3.29) and (3.27) together back to (3.21), we deduce that the LRC holds for μ and λ obeying the inequality

$$\zeta_3 \geq \frac{\mu}{2}(1+\delta)^2 + \frac{\lambda}{2}. \tag{3.30}$$

For instance, taking $\beta = 2$, $k = 5$, $\eta = 0.5$, and $\epsilon = 0.001$, we have $\zeta_1 = 0.8897$ and $\zeta_2 = 0.0213$, which confirms $\langle \ell_{\mathrm{rw}}(\boldsymbol{z}), \boldsymbol{h} \rangle \geq 0.1065 \|\boldsymbol{h}\|^2$. Setting further $\delta = 0.001$ leads to

$$0.1065 \geq 0.501\mu + 0.5\lambda \tag{3.31}$$

which concludes the proof of the LRC in (3.21). The local error contraction in (3.20) follows directly after substituting the LRC into (3.23), hence validating Prop. 5.

Chapter 4

Phase Retrieval via Stochastic Optimization

Based on the amplitude-based formulation (2.2) again, this chapter puts forth a lightweight algorithm, referred to as *stochastic truncated amplitude flow* (STAF). STAF offers an iterative algorithm that builds upon but considerably broadens the scope of TAF [129]. Specifically, it operates in two stages: Stage one employs a stochastic variance reduced gradient algorithm to obtain an orthogonality-promoting initialization, whereas the second stage applies stochastic truncated amplitude-based iterations to refine the initial estimate. Our approach is shown able to recover any n-dimensional signal $\boldsymbol{x}$ from a nearly minimal number of magnitude-only measurements in linear time. Relative to TAF, STAF is well suited for large-scale applications. Besides achieving order-optimal sample and computational complexities, STAF enjoys $\mathcal{O}(n)$ per-iteration complexity in both initialization and refinement stages, which not only improves upon state-of-the-art alternatives that can afford $\mathcal{O}(n^2)$, but it is also order optimal. This makes STAF applicable and appealing to common large-scale imaging phase retrieval settings. Comparisons between convex and non-convex solvers in terms of sample complexity and computational complexity to acquire an ϵ-accurate solution are presented in Table 4.1.

4.1 Stochastic Truncated Amplitude Flow

In this section, TAF is first reviewed, and its limitations for large-scale applications are highlighted. To cope with these limitations, simple, scalable, and fast stochastic gradient descent

Table 4.1: Computational Complexity of Different Algorithms

Algorithm	Sample complexity m	Computational complexity
PhaseLift [23]	$\mathcal{O}(n)$	$\mathcal{O}\big(n^3/\epsilon\big)$
PhaseCut [121]	$\mathcal{O}(n)$	$\mathcal{O}\big(n^3/\epsilon\big)$
AltMinPhase [91]	$\mathcal{O}\big(n\log n(\log^2 n+\log(1/\epsilon)\big)$	$\mathcal{O}\big(n^2\log n(\log^2 n+\log^2(1/\epsilon)\big)$
WF [22]	$\mathcal{O}(n\log n)$	$\mathcal{O}\big(n^3\log n\log(1/\epsilon)\big)$
TAF [129], TWF [32]	$\mathcal{O}(n)$	$\mathcal{O}\big(n^2\log(1/\epsilon)\big)$
This chapter	$\mathcal{O}(n)$	$\mathcal{O}\big(n^2\log(1/\epsilon)\big)$

(SGD)-type algorithms for both the initialization and gradient refinement stages are developed in this chapter.

The orthogonality-promoting initialization in Chapter 2 amounts to the following principal component analysis (PCA) problem

$$\tilde{\boldsymbol{z}}^0 := \arg\max_{\|\boldsymbol{z}\|=1} \ \boldsymbol{z}^{\mathcal{T}}\bar{\boldsymbol{Y}}_0\boldsymbol{z} \tag{4.1}$$

where

$$\bar{\boldsymbol{Y}}_0 := \frac{1}{|\overline{\mathcal{I}}^0|}\boldsymbol{D}\boldsymbol{D}^{\mathcal{T}} = \frac{1}{|\overline{\mathcal{I}}^0|}\sum_{i\in\overline{\mathcal{I}}^0}\frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}$$

for some index subset $\overline{\mathcal{I}}^0 \subset [m]$. When the signal dimension n is modest, problem (4.1) can be solved exactly by a full singular value decomposition (SVD) of $\boldsymbol{D}$ [52]. Yet it has running time of $\mathcal{O}(\min\{n^2|\overline{\mathcal{I}}^0|, n|\overline{\mathcal{I}}^0|^2\})$ (or simply $\mathcal{O}(n^3)$ because $|\overline{\mathcal{I}}^0|$ is required to be on the order of n), which grows prohibitively in large-scale applications. A common alternative is the power method that is tabulated in Alg. 3, and was also employed by [129, 22, 32] to find an initialization [52]. Power method, on the other hand, involves a matrix-vector multiplication $\bar{\boldsymbol{Y}}_0\boldsymbol{u}^t$ per iteration, thus incurring per-iteration complexity of $\mathcal{O}(n|\overline{\mathcal{I}}^0|)$ or $\mathcal{O}(n^2)$ by passing through the selected data $\{\boldsymbol{a}_i\}_{i\in\overline{\mathcal{I}}^0}$. Furthermore, to produce an ϵ-accurate solution, it incurs runtime of [100]:

$$\mathcal{O}\left(\frac{1}{\delta}n|\overline{\mathcal{I}}^0|\log(1/\epsilon)\right) \tag{4.2}$$

depending on the eigengap $\delta > 0$, which is defined as the gap between the largest and the second largest eigenvalues of $\bar{\boldsymbol{Y}}_0$ normalized by the largest one [52]. It is clear that when the eigengap δ

is small, the runtime of $\mathcal{O}\big(n|\bar{\mathcal{I}}^0|\log(1/\epsilon)/\delta\big)$ required by the power method would be equivalent to many passes over the entire data, and this could be prohibitively for large datasets [106]. Hence, the power method may not be appropriate for computing the initialization in large-scale applications, particularly those involving small eigengaps.

Algorithm 3 Power method

1: **Input:** Matrix $\bar{\boldsymbol{Y}}_0 = \frac{1}{|\bar{\mathcal{I}}^0|}\boldsymbol{D}\boldsymbol{D}^{\mathcal{T}}$.
2: **Initialize** a unit vector $\boldsymbol{u}^0 \in \mathbb{R}^n$ randomly.
3: **For** $t = 0$ **to** $T-1$ **do**
$\boldsymbol{u}^{t+1} = \frac{\bar{\boldsymbol{Y}}_0\boldsymbol{u}^t}{\|\bar{\boldsymbol{Y}}_0\boldsymbol{u}^t\|}$.
4: **End for**
5: **Output:** $\tilde{\boldsymbol{z}}^0 = \boldsymbol{u}^T$.

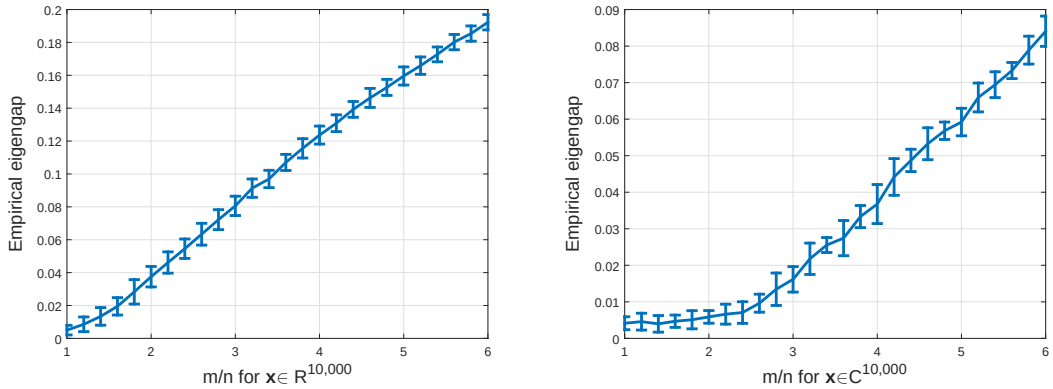

Figure 4.1: Rigengaps δ of $\bar{\boldsymbol{Y}}_0 \in \mathbb{R}^{n\times n}$. Left: Real Gaussian model; Right: Complex Gaussian model.

4.1.1 Variance-reducing orthogonality-promoting initialization

This section first presents some empirical evidence showing that small eigengaps appear commonly in the orthogonality-promoting initialization approach. Figure 4.1 plots empirical eigengaps of $\bar{\boldsymbol{Y}}_0 \in \mathbb{R}^{n\times n}$ under the real and complex Gaussian models over 100 Monte Carlo realizations under default parameters of TAF, where $n = 10,000$ is fixed, and m/n the number of equations and unknowns increases by 0.2 from 1 to 6. As shown in Fig. 4.1, the eigengaps of $\bar{\boldsymbol{Y}}_0$ resulting from the orthogonality-promoting initialization in [129, Alg. 1] are rather small

particularly for small m/n close to the information limit 2. Using power iterations in Alg. 3 of runtime $\mathcal{O}\big(n|\overline{\mathcal{I}}^0|\log(1/\epsilon)/\delta\big)$ in (4.2) thus entails many passes over the entire data due to a small eigengap factor of $1/\delta$, which may not perform well in the presence of large dimensions that are common to imaging applications [106]. On the other hand, instead of using the deterministic power method, stochastic and incremental algorithms have been advocated in [93, 106]. These algorithms perform a much cheaper update per iteration by choosing some $i_t \in \overline{\mathcal{I}}^0$ either uniformly at random or in a cyclic manner, and update the current iterate using only $\boldsymbol{a}_{i_t}$. They are shown to have per-iteration complexity of $\mathcal{O}(n)$, which is very appealing to large-scale applications. Building on recent advances in accelerating stochastic optimization schemes [64], a variance-reducing principal component analysis (VR-PCA) algorithm can be found in [106]. VR-PCA performs cheap stochastic iterations, yet its total runtime is $\mathcal{O}\big(n(\overline{\mathcal{I}}^0 + 1/\delta^2)\log(1/\epsilon)\big)$ which depends only logarithmically on the solution accuracy $\epsilon > 0$. This is in sharp contrast to the standard SGD variant, whose runtime depends on $1/\epsilon$ due to the large variance of stochastic gradients [93].

For the considered large-scale phase retrieval in most imaging applications, this chapter advocates using VR-PCA to solve the orthogonality-promoting initialization problem in (4.1). We refer to the resulting algorithm as the *variance-reducing orthogonality-promoting initialization* (VR-OPI), which is summarized in Alg. 4 next. Specifically, VR-OPI is a double-loop algorithm with a single execution of the inner loop referred to as an iteration and one execution of the outer loop referred to as an epoch. In practice, the algorithm consists of S epochs, while each epoch runs T (typically taken to be the data size $|\overline{\mathcal{I}}^0|$) iterations. Note that the full gradient evaluated per execution of the outer loop combined with the stochastic gradients inside the inner loop can be shown capable of reducing the variance of stochastic gradients [64].

The following results adapted from [106, Thm. 1] establish linear convergence rate of VR-OPI.

Proposition 7 ([106])**.** *Let $\boldsymbol{v}^1 \in \mathbb{R}^n$ be an eigenvector of $\overline{\boldsymbol{Y}}_0$ associated with the largest eigenvalue λ_1. Assume that $\max_{i\in[m]} \|\boldsymbol{a}_i\|^2 \le r := 2.3n$ (which holds with probability at least $1-m\mathrm{e}^{-n/2}$), the two largest eigenvalues of $\overline{\boldsymbol{Y}}_0$ are $\lambda_1 > \lambda_2 > 0$ with eigengap $\delta = (\lambda_1-\lambda_2)/\lambda_1$, and that $\langle \tilde{\boldsymbol{u}}^0, \boldsymbol{v}^1\rangle \ge 1/\sqrt{2}$. With any $0 < \epsilon,\, \xi < 1$, constant step size $\eta > 0$, and epoch length T chosen such that*

$$\eta \le \frac{c_0\xi^2}{r^2}\delta, \quad T \ge \frac{c_1\log(2/\xi)}{\eta\delta}, \quad T\eta^2r^2 + r\eta\sqrt{T\log(2/\xi)} \le c_2 \tag{4.3}$$

Algorithm 4 Variance-reduced orthogonality-promoting initialization (VR-OPI)

Input: Data matrix $\boldsymbol{D} = \{\boldsymbol{a}_i\}_{i\in\bar{\mathcal{I}}^0}$, step size $\eta = 20/m$, as well as the number of epochs $S = 100$, and the epoch length $T = |\bar{\mathcal{I}}^0|$ (by default).
Initialize a unit vector $\tilde{\boldsymbol{u}}^0 \in \mathbb{R}^n$ randomly.
For $s = 0$ **to** $S-1$ **do**
  $\boldsymbol{w} = \frac{1}{|\bar{\mathcal{I}}^0|}\sum_{i\in\bar{\mathcal{I}}^0} \boldsymbol{a}_i(\boldsymbol{a}_i^{\mathcal{T}}\tilde{\boldsymbol{u}}^s)$
  $\boldsymbol{u}^1 = \tilde{\boldsymbol{u}}^s$.
  For $t = 0$ **to** $T-1$ do
    Pick $i_t \in \bar{\mathcal{I}}^0$ uniformly at random
    $\boldsymbol{\nu}^{t+1} = \boldsymbol{u}^t + \eta\left[\boldsymbol{a}_{i_t}\big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{u}^t - \boldsymbol{a}_{i_t}^{\mathcal{T}}\tilde{\boldsymbol{u}}^s\big) + \boldsymbol{w}\right]$
    $\boldsymbol{u}^{t+1} = \frac{\boldsymbol{\nu}^{t+1}}{\|\boldsymbol{\nu}^{t+1}\|}$.
  End For
  $\tilde{\boldsymbol{u}}^{s+1} = \boldsymbol{u}^T$.
End For
Output: $\tilde{\boldsymbol{z}}^0 = \boldsymbol{u}^S$.

for certain universal constants c_0, c_1, $c_2 > 0$, successive estimates of VR-OPI (summarized in Alg. 4) after $S = \lceil \log(1/\epsilon)/\log(2/\xi)\rceil$ epochs satisfy

$$|\langle \tilde{\boldsymbol{u}}_S, \boldsymbol{v}_1\rangle|^2 \geq 1-\epsilon \tag{4.4}$$

with probability exceeding $1 - \lceil \log\epsilon\rceil\xi$. Typical parameter values are $\eta = 20/m$, $S = 100$, and $T = |\bar{\mathcal{I}}^0|$.

The proof of Prop. 7 can be found in [106]. Even though PCA in (4.1) is non-convex, the SGD based VR-OPI algorithm converges to the globally optimal solution under mild conditions [106]. Moreover, fixing any $\xi \in (0,1)$, conditions in (4.3) hold true when T is chosen to be on the order of $1/(\eta\delta)$, and η to be sufficiently smaller than δ/r^2. Expressed differently, if VR-OPI runs $T = \Theta(r^2/\delta^2)$ iterations per epoch for a total number $S = \Theta\big(\log(1/\epsilon)\big)$ of epochs, then the returned VR-OPI estimate is ϵ-accurate with probability at least $1 - \lceil\log_2(1/\epsilon)\rceil\xi$. Since each epoch takes $\mathcal{O}\big(n(T + |\bar{\mathcal{I}}^0|)\big)$ time to implement, the total runtime is of

$$\mathcal{O}\Big(n\Big(|\bar{\mathcal{I}}^0| + \frac{r^2}{\delta^2}\Big)\log(1/\epsilon)\Big) \tag{4.5}$$

which validates the exponential convergence of VR-OPI. Additionally, when $\delta/r \geq \Omega\big(1/\sqrt{|\bar{\mathcal{I}}^0|}\big)$,

the total runtime reduces to $\mathcal{O}\big(n|\overline{\mathcal{I}}^0|\log(1/\epsilon)\big)$ up to log-factors. It is worth emphasizing that the required runtime is proportional to the time required to scan the selected data once, which is in stark contrast to the runtime of $\mathcal{O}\big(n|\overline{\mathcal{I}}^0|\log(1/\epsilon)/\delta\big)$ when using power method [52]. Simulated tests in Sec. 4.3 corroborate the effectiveness of VR-OPI over the popular power method in processing data involving large dimensions m and/or n.

4.1.2 Stochastic truncated gradient iterations

Driven by the need of efficiently processing large-scale phaseless data in imaging applications, a stochastic solution algorithm is put forth for minimizing the amplitude-based cost function in (2.2). To ensure good performance, the gradient regularization rule in (2.10) is also accounted for to lead to our truncated stochastic gradient iterations. It is worth mentioning that the Kaczmarz method [65] was also used for solving a system of phaseless quadratic equations in [140]. However, Kaczmarz variants of block or randomized updates converge to at most a neighborhood of the optimal solution $\boldsymbol{x}$. Distance between the Kaczmarz estimates and $\boldsymbol{x}$ is bounded in terms of the dimension m and the size of the amplitude data vector $\boldsymbol{\psi}$ measured by the ℓ_1- or ℓ_∞-norm. Nevertheless, the obtained bounds of the form $m\|\boldsymbol{\psi}\|_1$ or $m\|\boldsymbol{\psi}\|_\infty$ are rather loose (m typically very large), and less attractive than the geometric convergence to the global solution $\boldsymbol{x}$ to be established for also stochastic iterations based STAF.

Adopting the intensity-based Poisson likelihood function (1.4), an incremental version of TWF was developed in [71], which provably converges to $\boldsymbol{x}$ in linear time. Albeit achieving improved empirical performance and faster convergence over TWF in terms of the number of passes over the entire data to produce an ϵ-accurate solution [32], the number of measurements it requires for exact recovery is still relatively far from the information-theoretic limits. Specifically for the real Gaussian $\boldsymbol{a}_i$ designs, ITWF requires about $m \geq 3.2n$ noiseless measurements to guarantee exact recovery relative to $4.5n$ for TWF [32]. Recall that TAF achieves exact recovery from about $3n$ measurements [129]. Furthermore, gradient iterations can be trapped in saddle points when dealing with non-convex optimization. In contrast, stochastic or perturbed gradient iterations are able to escape saddle points, and converge globally to at least a local minimum [80, 79]. Hence, besides the appealing computational advantage, stochastic counterparts of TAF may further improve the performance over TAF, as also asserted by the comparison between ITWF and TWF. In the following, we present two STAF variants: Starting with an initial estimate $\boldsymbol{z}^0$ found using VR-OPI in Alg. 4, the first variant successively updates $\boldsymbol{z}^0$ through amplitude-based

stochastic gradient iterations with a constant step size $\mu > 0$ chosen on the order of $1/n$, while the second operates much like the Kaczmarz method, yet both suitably account for the truncation rule in (2.10).

Recalling the amplitude-based cost function

$$\underset{\boldsymbol{z}\in\mathbb{R}^n}{\text{minimize}} \;\; \ell(\boldsymbol{z}) = \sum_{i=1}^{m} \ell_i(\boldsymbol{z}) := \frac{1}{2}\sum_{i=1}^{m} \left(\psi_i - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|\right)^2 \tag{4.6}$$

our approach to solving (4.6) amounts to iteratively refining the initial estimate $\boldsymbol{z}^0$ by means of truncated stochastic gradient iterations. This is in contrast to (T)WF and TAF, which rely on (truncated) gradient-type iterations [22, 32, 129]. STAF processes one datum at a time and evaluates the generalized gradient of one component function $\ell_{i_t}(\boldsymbol{z})$ for some index $i_t \in \{1, 2, \ldots, m\}$ per iteration $t \geq 0$. Specifically, STAF successively updates $\boldsymbol{z}^0$ using the following truncated stochastic gradient iterations for all $t \geq 0$:

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \mu^t \nabla \ell_{i_t}(\boldsymbol{z}^t) \mathbb{1}_{\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|/|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{x}| \geq 1/(1+\gamma)\}} \tag{4.7}$$

with

$$\nabla \ell_{i_t}(\boldsymbol{z}^t) = \Big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_{i_t} \tag{4.8}$$

where μ^t is either set to be a constant $\mu > 0$ on the order of $1/n$, or taken as the time-varying one as in Kaczmarz's iteration, namely, $\mu^t = 1/\|\boldsymbol{a}_{i_t}\|^2$ [65]. The index i_t is sampled uniformly at random or with given probabilities from $\{1, 2, \ldots, m\}$, or it simply cycles through the entire set $[m]$. In addition, fixing the truncation threshold to $\gamma = 0.7$, the indicator function $\mathbb{1}_{\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|/|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{x}| \geq 1/(1+\gamma)\}}$ in (4.7) takes the value 1, if $|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|/|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{x}| \geq 1/(1+\gamma)$ holds true; and 0 otherwise. It is worth stressing that this truncation rule provably rejects 'bad' search directions with high probability. Moreover, this regularization maintains only gradient components of large enough $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|$ values, hence saving the objective function (2.2) from being non-differentiable at $\boldsymbol{z}^t$ and simplifying the theoretical analysis. Numerical tests demonstrating the performance improvement using the stochastic truncated iterations will be presented in Sec. 4.3.

Algorithm 5 Stochastic truncated amplitude flow (STAF)

1: **Input:** Data $\{(\boldsymbol{a}_i, \psi_i)\}_{i=1}^m$; maximum number of iterations $T = 500m$; by default, step sizes $\mu = 0.8/n$ or $\mu = 1.2/n$ in the real/complex models, truncation thresholds $|\bar{\mathcal{I}}^0| = \lceil \frac{1}{6}m \rceil$, and $\gamma = 0.7$.

2: **Evaluate** $\bar{\mathcal{I}}^0$ to consist of indices associated with the $|\bar{\mathcal{I}}^0|$ largest values among $\{\psi_i / \|\boldsymbol{a}_i\|\}$.

3: **Initialize** $\boldsymbol{z}^0$ as $\sqrt{\frac{1}{m}\sum_{i=1}^m \psi_i^2}\tilde{\boldsymbol{z}}^0$, where $\tilde{\boldsymbol{z}}^0$ is obtained via Alg. 4 with

$$\bar{\boldsymbol{Y}}_0 := \frac{1}{|\bar{\mathcal{I}}^0|} \sum_{i \in \bar{\mathcal{I}}^0} \frac{\boldsymbol{a}_i \boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}.$$

4: **For** $t = 0$ **to** $T - 1$ do

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \mu \boldsymbol{a}_{i_t} \Big(\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \psi_{i_t} \frac{\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t|} \Big) \mathbb{1}_{\left\{ |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma} \right\}} \tag{4.9}$$

where i_t is sampled uniformly at random from $\{1, 2, \ldots, m\}$, or,

$$\boldsymbol{z}^{t+1} = \boldsymbol{z}^t - \frac{\boldsymbol{a}_{i_t}}{\|\boldsymbol{a}_{i_t}\|^2} \Big(\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \psi_{i_t} \frac{\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t|} \Big) \mathbb{1}_{\left\{ |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma} \right\}} \tag{4.10}$$

where i_t is sampled at random from $\{1,\ 2,\ \ldots,\ m\}$ with probability proportional to $\|\boldsymbol{a}_{i_t}\|^2$.

5: **End for**

6: **Output:** $\boldsymbol{z}^T$.

4.2 Main Results

The proposed STAF scheme is summarized as Alg. 5, with either constant step size $\mu > 0$ in the truncated stochastic gradient iterations in (4.9), or with time-varying step size $\mu_t = 1/\|\boldsymbol{a}_{i_t}\|^2$ in the truncated Kaczmarz iterations in (4.10). Equipped with an initialization obtained using VR-OPI, both STAF variants will be shown to converge at an exponential rate to the globally optimal solution with high probability, as soon as m/n the number of equations and unknowns exceeds some numerical constant.

Assuming m independent data samples $\{(\boldsymbol{a}_i; \psi_i)\}$ drawn from the real Gaussian model, the following establishes theoretical performance of STAF in the absence of noise.

Theorem 3 (**Exact recovery).** *Consider the noiseless measurements* $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$ *with an*

arbitrary signal $\boldsymbol{x} \in \mathbb{R}^n$*, and i.i.d.* $\{\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)\}_{i=1}^m$*. If* μ_t *is either set to be a constant* $\mu > 0$ *as per* (4.9)*, or it is time-varying* $\mu_t = 1/\|\boldsymbol{a}_{i_t}\|^2$ *as per* (4.10) *with the corresponding index sampling scheme, and also*

$$m \geq c_0 n \quad \textit{and} \quad \mu \leq \mu_0 / n \tag{4.11}$$

then with probability at least $1 - c_1 m \exp(-c_2 n)$*, the stochastic truncated amplitude flow (STAF) estimates (tabulated in Alg. 5 with default parameters) satisfy*

$$\mathbb{E}_{\mathcal{P}^t}\left[\mathrm{dist}^2(\boldsymbol{z}^t, \boldsymbol{x})\right] \leq \rho \left(1 - \frac{\nu}{n}\right)^t \|\boldsymbol{x}\|^2, \quad t = 0,\, 1,\, \ldots \tag{4.12}$$

for $\rho = 1/10$ *and some numerical constant* $\nu > 0$*, where the expectation is taken over the path sequence* $\mathcal{P}^t := \{i_0,\, i_1, \ldots, i_{t-1}\}$*, and* $c_0,\, c_1,\, c_2,\, \mu_0 > 0$ *are certain universal constants.*

The mean-square distance between the iterate and the global solution is reduced by a factor of $(1 - \nu/n)^m$ after one pass through the entire data. Heed that the expectation $\mathbb{E}_{\mathcal{P}^t}[\cdot]$ in (4.12) is taken over the algorithmic randomness $\mathcal{P}^t$ rather than the data. This is important since in general the data may be modeled as deterministic. Although only performing stochastic iterations in (4.9) and (4.10), STAF still enjoys linear convergence rate. This is in sharp contrast to typical SGD methods, where variance reduction techniques controlling the variance of the stochastic gradients are required to achieve linear convergence rate [64, 106], as in Alg. 4. Moreover, the largest constant step size that STAF can afford is estimated to be $\mu_0 = 0.8469$, giving rise to a convergence factor of $\nu_0 = 0.0696$ in (4.12). When truncated Kaczmarz iterations are implemented, ν is estimated to be 1.0091 much larger than the one in the constant step size case. Our experience with numerical experiments also confirm that the Kaczmarz-based STAF in (4.10) converges faster than the constant step-size based one in (4.9), yet it is slightly more sensitive when additive noise is present in the data.

4.3 Numerical Experiments

This section presents extensive numerical experiments evaluating performance of STAF using synthetic data and real images. STAF was thoroughly compared with existing alternatives including TAF [129], (T)WF [22, 32], and ITWF [71]. The initialization in each scheme was found based on a number of (power/stochastic) iterations equivalent to 100 passes over

the entire data, which was subsequently refined by a number of iterations corresponding to $1,000$ passes; unless otherwise stated. Two performance evaluation metrics were used: the relative root MSE and the empirical successful recovery rate among 100 independent runs. The Matlab implementations of STAF can be downloaded from `https://gangwg.github.io/STAF/`.

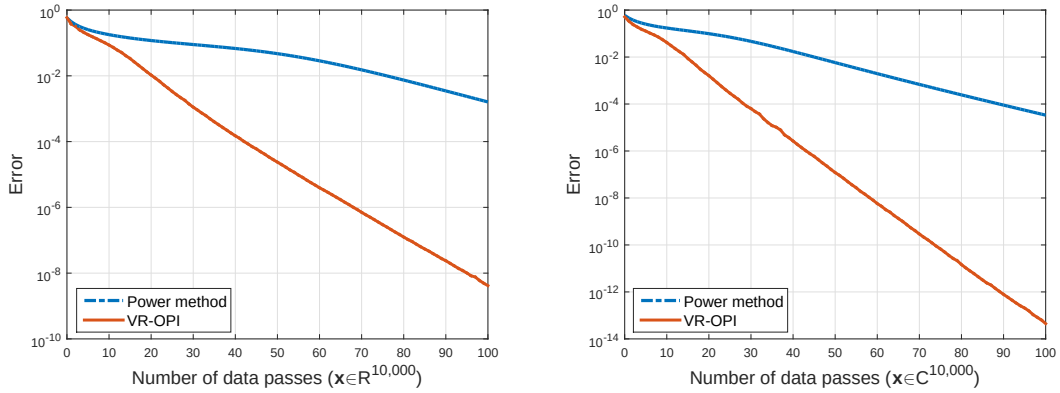

Figure 4.2: Error evolution of the iterates for solving problem (4.1) with step size $\eta = 1$. Left: Noiseless real Gaussian model with $m = 2n - 1$; Right: Noiseless complex Gaussian model with and $m = 4n - 4$.

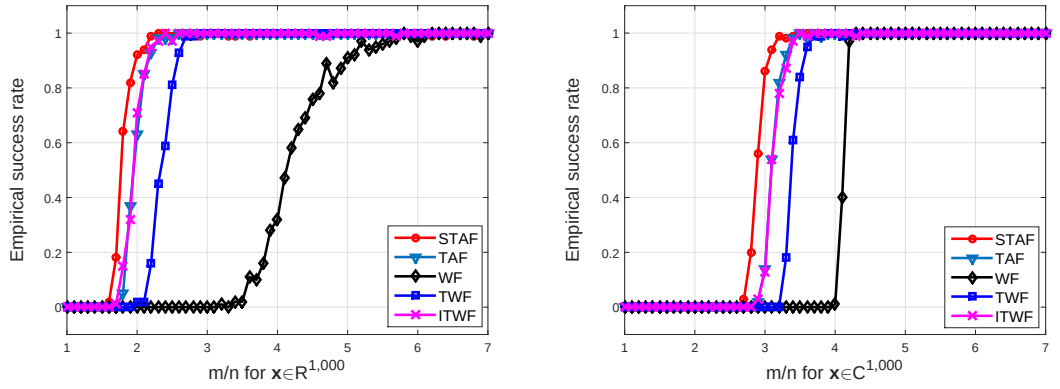

Figure 4.3: Empirical success rate under the same orthogonality-promoting initialization. Left: Noiseless real Gaussian model; Right: Noiseless complex Gaussian model.

The first experiment compares VR-OPI in Alg. 4 with the power method in Alg. 3 to solve

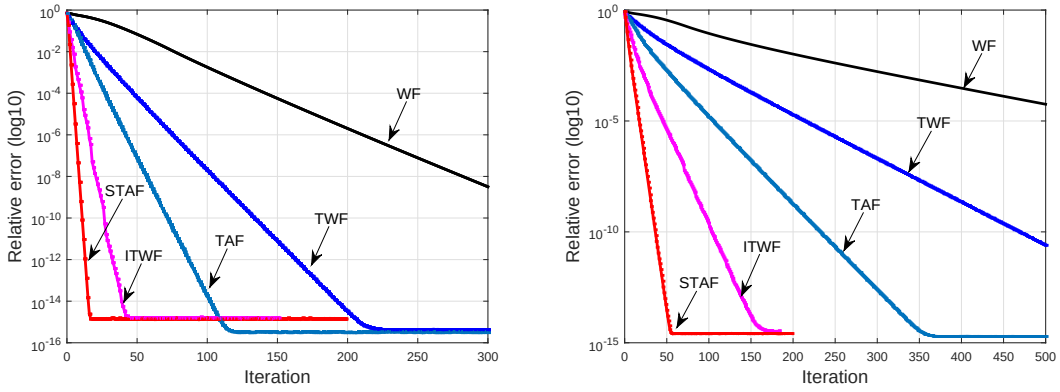

Figure 4.4: Relative error versus iterations under the same orthogonality-promoting initialization. Left: Noiseless real Gaussian model; Right: Noiseless complex Gaussian model.

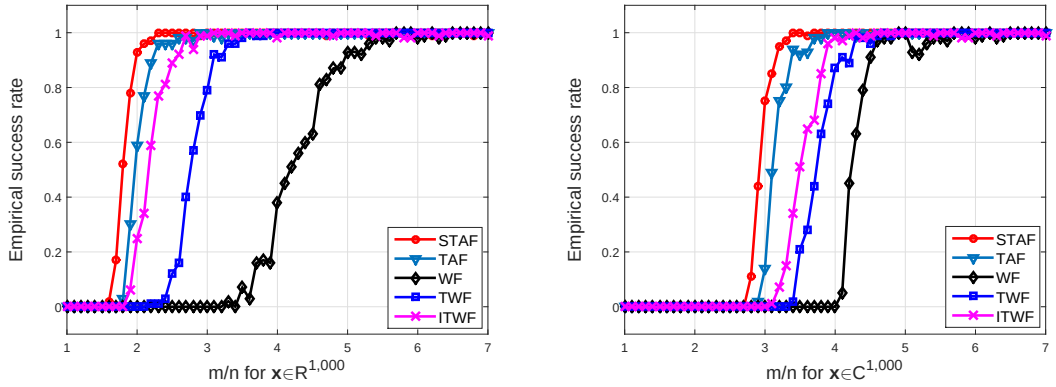

Figure 4.5: Empirical success rate. Left: Noiseless real Gaussian model; Right: Noiseless Gaussian model.

the orthogonality-promoting initialization optimization in (4.1). The comparison is carried out in terms of the number of data passes to achieve the same solution accuracy, in which one pass through the selected data amounts to a number $|\overline{\mathcal{I}}^0|$ of gradient evaluations of component functions. First, synthetic data based experiments are conducted using the real/complex Gaussian models with $n = 10,000$ under the known sufficient conditions for uniqueness, i.e., $m = 2n - 1$ in the real case, and $m = 4n - 4$ in the complex case. Figure 4.2 plots the error evolution of the iterates $\boldsymbol{u}^t$ for the power method and VR-OPI, where the error in logarithmic scale is defined

as $\log_{10}(1 - \|\boldsymbol{D}^{\mathcal{T}}\boldsymbol{u}^t\|^2/\|\boldsymbol{D}^{\mathcal{T}}\boldsymbol{v}^0\|^2)$ with the exact principal eigenvector $\boldsymbol{v}^0$ computed from the SVD of $\bar{\boldsymbol{Y}}_0 = \boldsymbol{D}\boldsymbol{D}^{\mathcal{T}}$ in (4.1). Apparently, the inexpensive stochastic iterations of VR-OPI achieve certain solution accuracy with considerably fewer gradient evaluations or data passes in both real and complex settings. This is important for tasks of large $|\bar{\mathcal{I}}^0|$, or equivalently large dimension m (since $|\bar{\mathcal{I}}^0| = 5m/6$ by default), because one less data pass implies $|\bar{\mathcal{I}}^0|$ fewer gradient evaluations and thus results in considerable savings in computational resources.

The second experiment evaluates the refinement stage of STAF relative to its competing alternatives including those of (T)WF, TAF, and ITWF in a variety of settings. For fairness, all schemes were here initialized using the same orthogonality-promoting initialization found using 100 power iterations, and subsequently applied a number of iterations corresponding to $T = 1,000$ data passes. Figure 4.3 depicts the empirical success rate of all considered schemes with m/n varying by 0.1 from 1 to 7. Figure 4.4 compares the convergence speed of various schemes in terms of the number of data passes to produce solutions of a given accuracy. Starting with the same initialization, STAF outperforms its competing alternatives under both real/complex Gaussian models. In particular, SGD-based STAF improves in terms of exact recovery and convergence speed over the state-of-the-art gradient-type TAF, corroborating the benefit of using SGD-type solvers to cope with saddle points and local minima of non-convex optimization [71].

The previous experiment showed improved performance of STAF under the same initialization. Now, we present numerical results comparing different schemes equipped with their own initialization, namely, WF with spectral initialization [22], (I)TWF with truncated spectral initialization [32], as well as TAF with orthogonality-promoting initialization using power iterations [129], and STAF with VR-OPI. Figure 4.5 demonstrates merits of STAF over its competing alternatives in exact recovery performance on the noiseless real (left) and complex (right) Gaussian model. Specifically in the real case, STAF guarantees exact recovery from about $2.3n$ magnitude-only measurements, which is close to the information-theoretic limit of $m = 2n - 1$. In comparison, existing alternatives require a few times more measurements to achieve exact recovery. STAF also performs well in the complex case.

To demonstrate the robustness of STAF against additive noise, we perform stable phase retrieval under the noisy real/complex Gaussian model $\psi_i = |\boldsymbol{a}_i^{\mathcal{H}}\boldsymbol{x}| + \eta_i$, with $\eta_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ i.i.d., and $\sigma^2 = 0.1^2\|\boldsymbol{x}\|^2$. The noisy data for magnitude-square based algorithms were generated as $y_i = \psi_i^2$. Curves in Fig. 4.6 clearly show near-perfect statistical performance and fast

convergence of STAF.

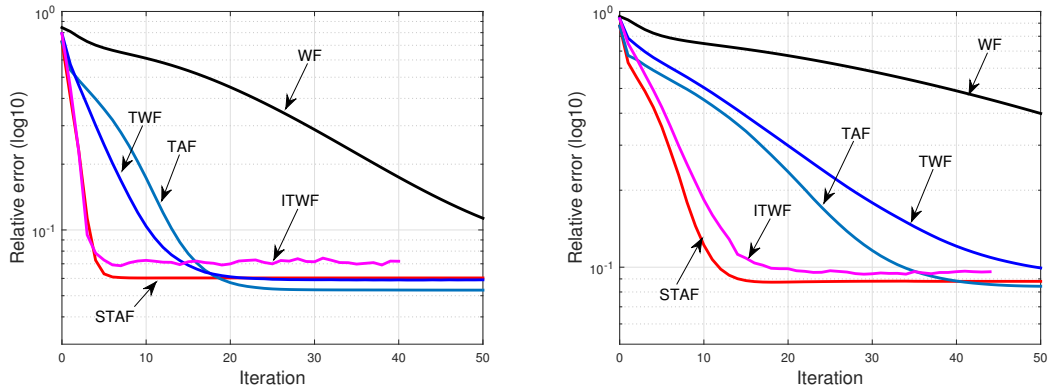

Figure 4.6: Relative error versus iterations with $n = 1,000$ and $m/n = 5$. Left: Noisy real Gaussian model; Right: Noisy complex Gaussian model.

Finally, to demonstrate the effectiveness and scalability of STAF on real data, the Milky Way Galaxy image in Fig. 2.11 is considered again. We collected the physically realizable measurements called coded diffraction patterns (CDP) using random masks [21]. CDP measurements in this experiment were generated using $K = 8$ random masks for a total of $m = nK$ measurements. In this part, since the FFT can be implemented in $\mathcal{O}(n \log n)$ instead of $\mathcal{O}(n^2)$ operations, the advantage of using STAF with optimal per-iteration complexity is less pronounced. Hence, instead of processing one quadratic measurement per iteration, a block STAF version processes per iteration n^2 measurements associated with one random mask. That is, STAF samples randomly the index $k \in \{1, 2, \ldots, K\}$ of masks in (2.26), and updates the iterate using all diffraction patterns corresponding to the k-th mask. In this case, STAF is able to leverage the efficient implementation of FFT, and converges fast. Figure 4.7 displays the recovered images, where the top is obtained after 100 data passes of VR-OPI iterations, and the bottom is produced by 100 data passes of STAF iterations refining the initialization. Apparently, the recovered images corroborate the effectiveness of STAF in real-world conditions.

4.4 Proofs

In this section, we provide the proofs for the main theorem and propositions of this chapter.

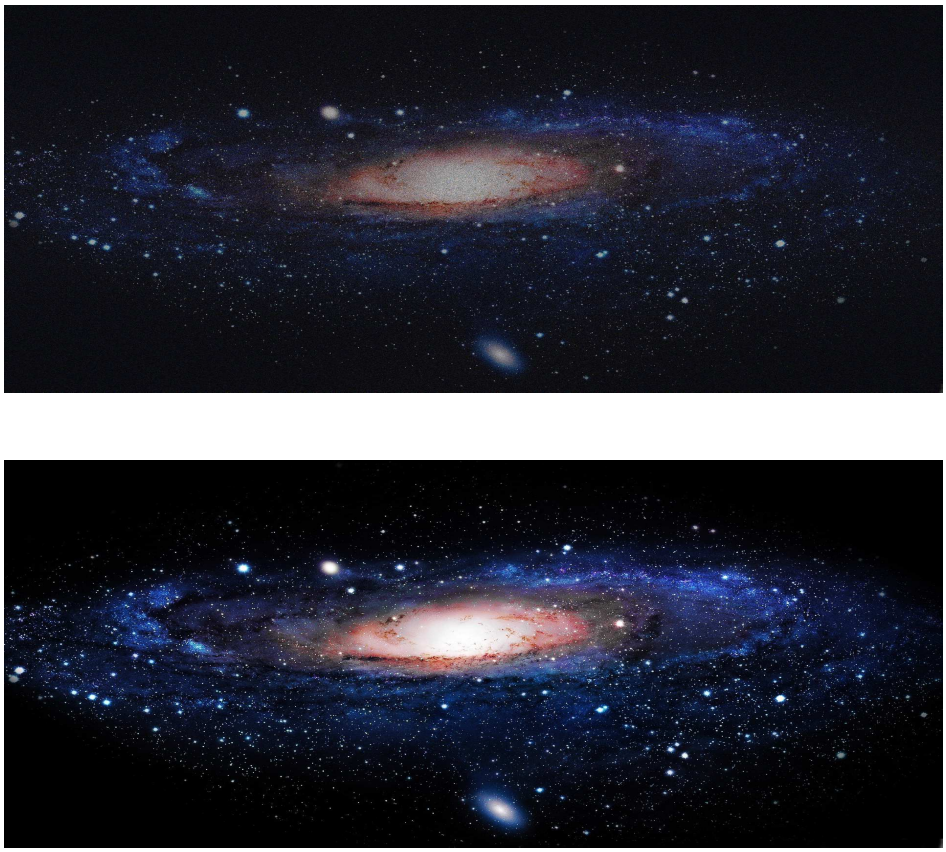

Figure 4.7: Recovered images after: the variance-reducing orthogonality-promoting initialization stage (top panel), and the STAF refinement stage (bottom panel) on the Milky Way Galaxy image using $K = 8$ random masks.

Proof for Theorem 3

Recall from [129, Thm.1] that when m/n exceeds some universal constant $c_0 > 0$, the estimate $\boldsymbol{z}^0$ returned by the orthogonality-promoting initialization obeys the following with high probability

$$\operatorname{dist}(\boldsymbol{z}^0, \boldsymbol{x}) \leq (1/10)\|\boldsymbol{x}\|. \tag{4.13}$$

Along the lines of (T)WF and TAF, to prove Thm. 3, it suffices to show that successive STAF iterates $\boldsymbol{z}^t$ are on average locally contractive around the planted solution $\boldsymbol{x}$, as asserted in the following proposition. See the Appendix for proof details.

Proposition 8 (Local error contraction). *Consider the noiseless measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ with an arbitrary signal $\boldsymbol{x} \in \mathbb{R}^n$, and i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ for $i = 1, 2, \ldots, m$. Under the default algorithmic parameters given in Alg. 5, there exist universal constants $c_0', c_1', c_2' > 0$, and $\nu > 0$, such that with probability at least $1 - c_2' m \exp(-c_1' n)$, the following holds simultaneously for all $\boldsymbol{z}^t$ satisfying* (4.13)

$$\mathbb{E}_{i_t}\big[\mathrm{dist}^2(\boldsymbol{z}^{t+1}, \boldsymbol{x})\big] \le \Big(1 - \frac{\nu}{n}\Big)\,\mathrm{dist}^2(\boldsymbol{z}^t, \boldsymbol{x}) \tag{4.14}$$

provided that $m \ge c_0' n$.

Proposition 8 demonstrates monotonic decrease of the MSE: Once entering a reasonably small-size neighborhood of $\boldsymbol{x}$, successive iterates of STAF will be dragged toward $\boldsymbol{x}$ at a linear rate. Upon establishing the local error contraction property in (4.14), taking expectation on both sides of (4.14) over i_{t-1}, and applying Prop. 8 again, yields a similar relation for the previous iteration. Continuing this process to reach the initialization $\boldsymbol{z}^0$ and appealing to the initialization result in (4.13) collectively, leads to (4.12), hence completes the proof of Thm. 3.

Proof of Proposition 8

To prove Prop. 8, let us first define the truncated gradient of $\ell(\boldsymbol{z})$ as follows

$$\nabla \ell_{\mathrm{tr}}(\boldsymbol{z}) = \sum_{i=1}^{m} \left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} - \psi_i \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|}\right)\boldsymbol{a}_i \mathbb{1}_{\left\{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t| \ge \frac{1}{1+\gamma}\psi_i\right\}} \tag{4.15}$$

which corresponds to the truncated gradient employed by TAF [129]. Instrumental in proving the local error contraction in Prop. 8, the following lemma adopts a sufficient decrease result from 3. The sufficient decrease is a key step in establishing the local regularity condition [22, 32, 129], which suffices to prove linear convergence of iterative optimization algorithms.

Now let us turn to the term on the left hand side of (4.14), which after plugging in the update of $\boldsymbol{z}^{t+1}$ in (4.9) or (4.10), boils down to

$$\begin{aligned}
\mathrm{dist}^2(\boldsymbol{z}^{t+1}, \boldsymbol{x}) &= \Big\|\boldsymbol{h}^t - \mu_t\Big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_{i_t}\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t| \ge \frac{\psi_{i_t}}{1+\gamma}\right\}}\Big\|^2 \\
&= \|\boldsymbol{h}^t\|^2 - 2\mu_t\Big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{h}^t\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t| \ge \frac{\psi_{i_t}}{1+\gamma}\right\}}
\end{aligned}$$

$$+ \mu_t^2 \Big(\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \psi_{i_t} \frac{\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t|} \Big)^2 \|\boldsymbol{a}_{i_t}\|^2 \mathbb{1}_{\left\{ |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma} \right\}} \tag{4.16}$$

where $\mu_t = \mu > 0$ with $i_t \in [m]$ sampled uniformly at random in (4.9), or $\mu_t = 1/\|\boldsymbol{a}_{i_t}\|^2$ with $i_t \in [m]$ selected with probability proportional to $\|\boldsymbol{a}_{i_t}\|^2$ in (4.10).

Consider first the constant step size case in (4.9). Take the expectation of both sides in (4.16) with respect to the selection of index i_t (rather than the data randomness) to obtain

$$\begin{aligned} \mathbb{E}_{i_t} \big[\mathrm{dist}^2(\boldsymbol{z}^{t+1}, \boldsymbol{x}) \big] = \|\boldsymbol{h}^t\|^2 - \frac{2\mu}{m} \sum_{i_t=1}^{m} \Big(\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \psi_{i_t} \frac{\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t|} \Big) \boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{h}^t \mathbb{1}_{\left\{ |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma} \right\}} \\ + \frac{\mu^2}{m} \sum_{i_t=1}^{m} \Big(\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \psi_{i_t} \frac{\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t|} \Big)^2 \|\boldsymbol{a}_{i_t}\|^2 \mathbb{1}_{\left\{ |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma} \right\}}. \end{aligned} \tag{4.17}$$

Now the task reduces to upper bounding the terms on the right hand side of (4.17). Note from (4.15) that by means of $\nabla \ell_{\mathrm{tr}}(\boldsymbol{z}^t)$, the second term in (4.17) can be re-expressed as follows

$$\begin{aligned} -\frac{2\mu}{m} \sum_{i_t=1}^{m} \Big(\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \psi_{i_t} \frac{\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t|} \Big) \boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{h}^t \mathbb{1}_{\left\{ |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma} \right\}} &= -\frac{2\mu}{m} \left\langle \nabla \ell_{\mathrm{tr}}(\boldsymbol{z}^t), \boldsymbol{h}^t \right\rangle \\ &\leq -4\mu \left(1 - \zeta_1 - \zeta_2 - 2\epsilon\right) \|\boldsymbol{h}\|^2 \end{aligned} \tag{4.18}$$

where the inequality follows from Prop. 3. Regarding the last term in (4.17), since for the i.i.d. real Gaussian $\boldsymbol{a}_i$'s, $\max_{i_t \in [m]} \|\boldsymbol{a}_{i_t}\| \leq 2.3n$ holds with probability at least $1 - m\mathrm{e}^{-n/2}$ [129], and also $\mathbb{1}_{\left\{ |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma} \right\}} \leq 1$, then the next holds with high probability

$$\begin{aligned} \frac{\mu^2}{m} \sum_{i_t=1}^{m} \Big(\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \psi_{i_t} \frac{\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t|} \Big)^2 \|\boldsymbol{a}_{i_t}\|^2 \mathbb{1}_{\left\{ |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma} \right\}} &\leq \frac{2.3n\mu^2}{m} \sum_{i_t=1}^{m} \big(|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| - |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{x}| \big)^2 \\ &\leq \frac{2.3n\mu^2}{m} \sum_{i_t=1}^{m} \big(\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{x} \big)^2 \\ &\leq \frac{2.3n\mu^2}{m} (\boldsymbol{h}^t)^{\mathcal{T}} \boldsymbol{A}^{\mathcal{T}} \boldsymbol{A} \boldsymbol{h}^t \\ &\leq 2.3(1+\delta)\mu^2 n \|\boldsymbol{h}^t\|^2 \end{aligned} \tag{4.19}$$

in which the second inequality comes from $(|\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t| - |\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{x}|)^2 \leq (\boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{z}^t - \boldsymbol{a}_{i_t}^{\mathcal{T}} \boldsymbol{x})^2$, and the

last arises due to the fact that $\lambda_{\max}(\boldsymbol{A}^{\mathcal{T}}\boldsymbol{A}) \leq (1+\delta)m$ holds with probability at least $1 - c_2' \exp(-c_1' n\delta^2)$, provided that $m \geq c_0' n\delta^{-2}$ for some universal constant c_0', c_1', $c_2' > 0$ [119, Thm. 5.39].

Substituting (4.18) and (4.19) into (4.17) establishes that

$$\mathbb{E}_{i_t}\big[\mathrm{dist}^2(\boldsymbol{z}^{t+1}, \boldsymbol{x})\big] \leq \big[1 - 4\mu\,(1 - \zeta_1 - \zeta_2 - 2\epsilon) + 2.3(1+\delta)\mu^2 n\big]\|\boldsymbol{h}^t\|^2 \tag{4.20}$$

holds with probability exceeding $1 - c_2 m \exp(-c_1 n)$ provided that $m \geq c_0 n$, where $c_0 \geq c_0'\delta^{-2}$. To obtain legitimate estimates for the step size, fixing ϵ, $\delta > 0$ to be sufficiently small constants, say e.g., 0.01, then using (4.20), μ can be chosen such that

$$4(0.98 - \zeta_1 - \zeta_2) - 2.42\mu n > 0$$

yielding

$$0 < \mu < \frac{4(0.98 - \zeta_1 - \zeta_2)}{2.42n} \approx \frac{0.8469}{n} := \frac{\mu^0}{n}. \tag{4.21}$$

Plugging $\mu = c_3/n$ for some $0 < c_3 \leq \mu^0$ into (4.20), gives rise to

$$\mathbb{E}_{i_t}\big[\mathrm{dist}^2(\boldsymbol{z}^{t+1}, \boldsymbol{x})\big] \leq \Big(1 - \frac{\nu}{n}\Big)\,\mathrm{dist}^2(\boldsymbol{z}^t, \boldsymbol{x}) \tag{4.22}$$

for

$$\nu := 4c_3(1 - \zeta_1 - \zeta_2 - 2\epsilon) - 2.3c_3^2(1+\delta) \leq \nu_0 := 0.0697$$

where the equality holds at the maximum step size $\mu = \mu^0$, hence concluding the proof of Prop. 8 for the constant step size case.

Now let us turn to the case of a time-varying step size. Specifically, let $\mu^t = 1/\|\boldsymbol{a}_{i_t}\|^2$, and i_t be sampled at random from the set $[m]$ with probability $\|\boldsymbol{a}_{i_t}\|^2 / \sum_{i_t=1}^m \|\boldsymbol{a}_{i_t}\|^2 = \|\boldsymbol{a}_{i_t}\|^2/\|\boldsymbol{A}\|_F^2$ [116]. Taking the expectation of both sides in (4.16) over i_t gives rise to

$$\begin{aligned}\mathbb{E}_{i_t}\big[\mathrm{dist}^2(\boldsymbol{z}^{t+1}, \boldsymbol{x})\big] =& \|\boldsymbol{h}^t\|^2 - 2\sum_{i_t=1}^m \frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{h}^t}{\|\boldsymbol{A}\|_F^2}\Big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma}\right\}} \\ &+ \sum_{i_t=1}^m \frac{1}{\|\boldsymbol{A}\|_F^2}\Big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|}\Big)^2\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t| \geq \frac{\psi_{i_t}}{1+\gamma}\right\}}.\end{aligned} \tag{4.23}$$

Consider random $\boldsymbol{A} := [\boldsymbol{a}_1 \ \cdots \ \boldsymbol{a}_m]^{\mathcal{T}}$ with i.i.d. rows $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and any fixed $\sigma > 0$.

Then, by means of Bernstein-type inequality [119, Prop. 5.16]

$$\Big|\frac{1}{mn}\|\boldsymbol{A}\|_F^2 - 1\Big| = \Big|\frac{1}{mn}\sum_{i,j} a_{i,j}^2 - 1\Big| \leq \sigma$$

holds with probability at least $1 - 2\exp(-mn\sigma^2/8)$. Therefore, the second term on the right hand side of (4.23) can be bounded as follows

$$\begin{aligned}
&-\frac{2}{\|\boldsymbol{A}\|_F^2}\sum_{i_t=1}^{m}\Big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{h}^t\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|\geq\frac{\psi_{i_t}}{1+\gamma}\right\}}\\
&\leq -\frac{2}{(1+\sigma)mn}\sum_{i_t=1}^{m}\Big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{h}^t\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|\geq\frac{\psi_{i_t}}{1+\gamma}\right\}}\\
&\leq -\frac{4m}{(1+\sigma)mn}\left(1-\zeta_1-\zeta_2-2\epsilon\right)\|\boldsymbol{h}\|^2\\
&\leq -\frac{4}{(1+\sigma)n}\left(1-\zeta_1-\zeta_2-2\epsilon\right)\|\boldsymbol{h}\|^2
\end{aligned} \tag{4.24}$$

where the second inequality follows from Prop. 3, and the last inequality from the fact that $m \geq c_0 n$. Concerning the last term on the right hand side of (4.23), one obtains that

$$\begin{aligned}
&\sum_{i_t=1}^{m}\frac{\|\boldsymbol{a}_{i_t}\|^2}{\|\boldsymbol{A}\|_F^2}\frac{1}{\|\boldsymbol{a}_{i_t}\|^2}\Big(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \psi_{i_t}\frac{\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|}\Big)^2\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|\geq\frac{\psi_{i_t}}{1+\gamma}\right\}}\\
&= \frac{1}{\|\boldsymbol{A}\|_F^2}\sum_{i_t=1}^{m}\left(|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t| - |\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{x}|\right)^2\mathbb{1}_{\left\{|\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t|\geq\frac{\psi_{i_t}}{1+\gamma}\right\}}\\
&\leq \frac{1}{\|\boldsymbol{A}\|_F^2}\sum_{i_t=1}^{m}\left(\boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{z}^t - \boldsymbol{a}_{i_t}^{\mathcal{T}}\boldsymbol{x}\right)^2\\
&\leq \frac{1}{\|\boldsymbol{A}\|_F^2}(\boldsymbol{h}^t)^{\mathcal{T}}\boldsymbol{A}^{\mathcal{T}}\boldsymbol{A}\boldsymbol{h}^t\\
&\leq \frac{(1+\delta)m}{(1-\sigma)mn}\|\boldsymbol{h}^t\|^2\\
&\leq \frac{(1+\delta)}{(1-\sigma)n}\|\boldsymbol{h}^t\|^2
\end{aligned} \tag{4.25}$$

which holds with high probability as soon as $m \geq c_0 n \geq c_0'\delta^{-2}n$.

Putting results in (4.23), (4.24), and (4.25) together, one establishes that the following holds

$$\mathbb{E}_{i_t}\big[\mathrm{dist}^2(\boldsymbol{z}^{t+1},\, \boldsymbol{x})\big] \leq \Big[1 - \frac{4}{(1+\sigma)n}\,(1-\zeta_1-\zeta_2-2\epsilon) + \frac{(1+\delta)}{(1-\sigma)n}\Big]\|\boldsymbol{h}^t\|^2 \tag{4.26}$$

with probability at least $1 - c_2 m \exp(-c_1 n)$ provided that $m \geq c_0 n$. Hence, one can set in this case

$$\nu := \frac{4}{(1+\sigma)n}\,(1-\zeta_1-\zeta_2-2\epsilon) - \frac{(1+\delta)}{(1-\sigma)n}.$$

Taking without loss of generality $\delta,\, \sigma,\, \epsilon$ to be 0.01, and substituting the estimates of $\zeta_1,\, \zeta_2$ into (4.26), one arrives at $\nu = 1.0091$ to deduce that

$$\mathbb{E}_{i_t}\big[\mathrm{dist}^2(\boldsymbol{z}^{t+1},\, \boldsymbol{x})\big] \leq \left(1 - \frac{1.0091}{n}\right)\mathrm{dist}^2(\boldsymbol{z}^t, \boldsymbol{x}) \tag{4.27}$$

which holds with high probability as soon as $m \geq c_0 n$, establishing the local error contraction property of the truncated Kaczmarz iterations in (4.10), as claimed in Prop. 8.

Combining the results in (4.22) and (4.27), we proved the local error contraction property in Prop. 8 of the two STAF variants under both constant and time-varying step sizes.

Chapter 5

Phase Retrieval of Sparse Signals

In diverse applications, especially those related to imaging, the signal of interest is naturally sparse or admits a sparse representation after some known and deterministic linear transformation [61]. For example, astronomical imaging centers around sparsely distributed stars, while electron microscopy deals with sparsely distributed atoms or molecules. As phase retrieval of sparse signals is of practical relevance, SDP, AltMinPhase, and WF recovery methods have been generalized to sparse phase retrieval producing solvers termed compressive phase retrieval via lifting (CPRL) [92], sparse AltMinPhase [91], thresholded Wirtinger flow (ThWF) [18], SparsePhaseMax [53]. CPRL in particular, accounts for the sparsity by adding an ℓ_1-regularization term on the wanted signal to the original PhaseLift formulation. The other two approaches are two-stage iterative counterparts consisting of a (sparse) initialization, and a series of refinements of the initialization with gradient-type iterations. The greedy sparse phase retrieval (GESPAR) algorithm is based on a fast 2-opt local search [107]. A probabilistic approach is developed based on the generalized approximate message passing (GAMP) algorithm [105]. Assuming noise-free Gaussian random measurements, CPRL recovers any k-sparse n-dimensional ($k \ll n$) signal exactly from $\mathcal{O}(k^2 \log n)$ measurements at computational complexity $\mathcal{O}(n^3)$ [76]. Sparse AltMinPhase and ThWF, on the other hand, require $\mathcal{O}(k^2 \log n)$ measurements [91, 18], and SparseAltMinPhase incurs complexity $\mathcal{O}(k^2 n \log n)$ [91].

In this chapter, we propose here a novel sparse phase retrieval algorithm, which we call *SPARse Truncated Amplitude flow* (SPARTA). Adopting an amplitude-based nonconvex formulation of the sparse phase retrieval , SPARTA emerges as a two-stage iterative solver: In stage one, the support of the underlying signal is estimated first using a well-justified rule, and subsequently

power iterations are employed to obtain an initialization restricted on the recovered support; while the second stage successively refines the initialization with a series of hard thresholding based truncated gradient iterations. Both stages are conceptually simple, scalable, and fast. Moreover, we demonstrate that SPARTA recovers any k-sparse n-dimensional signal $\boldsymbol{x}$ ($k \ll n$) with minimum nonzero entries (in modulus) on the order of $(1/\sqrt{k})\|\boldsymbol{x}\|_2$ from $\mathcal{O}(k^2 \log n)$ measurements. Further, to reach any given solution accuracy $\epsilon > 0$, SPARTA incurs total computational cost of $\mathcal{O}(k^2 n \log n \log(1/\epsilon))$, which improves upon the state-of-the-art by at least a factor of k. This computational advantage is paramount in large-scale imaging applications, where the basis factor $n \log n$ is large, typically on the order of millions. In addition, SPARTA can be shown robust to additive noise of bounded support. Extensive simulated tests demonstrate markedly improved exact recovery performance (in the absence of noise), robustness to noise, and runtime speedups relative to the state-of-the-art algorithms.

5.1 Sparse Phase Retrieval

Succinctly stated, the sparse phase retrieval task amounts to reconstructing a sparse $\boldsymbol{x} \in \mathbb{R}^n$ (or $\mathbb{C}^n$) given a system of phaseless quadratic equations taking the form [88]

$$\psi_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|, \quad i = 1, 2, \ldots, m, \qquad \text{subject to} \quad \|\boldsymbol{x}\|_0 \leq k \tag{5.1}$$

where $\{\psi_i\}_{i=1}^m$ are the observed modulus data, and $\{\boldsymbol{a}_i\}_{i=1}^m$ are known sensing (feature) vectors. The sparsity level $k \ll n$ is assumed known *a priori* for theoretical analysis purposes, while numerical implementations with unknown k values will be tested as well. Alternatively, the data can be given in modulus squared (i.e., intensity) form as $\{y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2\}_{i=1}^m$. It has been established that $m = 2k$ generic (e.g., random Gaussian) measurements as in (5.1) are necessary and sufficient for uniquely determining a k-sparse solution in the real case, and $m \geq 4k - 2$ are sufficient in the complex case [3]. In the noisy scenario, stable compressive phase retrieval requires at least as many measurements as the corresponding compressive sensing problem since one is tasked with even less (no phase) information. Hence, stable sparse phase retrieval requires at least $\mathcal{O}(k \log(n/k))$ measurements as in compressive sensing [60]. Indeed, it has been recently demonstrated that $\mathcal{O}(k \log(n/k))$ generic measurements also suffice for stable phase retrieval of a real sparse signal [44].

Given $\{(\boldsymbol{a}_i, \psi_i)\}_{i=1}^m$ and assuming also the existence of a unique k-sparse solution (up to a global sign), our objective is to develop simple yet effective algorithms to provably reconstruct any k-sparse n-dimensional signal $\boldsymbol{x}$ from a small number (far less than n) of phaseless quadratic equations as in (5.1).

Similar to the non-convex model (2.2), recovering a k-sparse solution from a set of quadratic equations can be recast as minimizing the ensuing amplitude-based empirical loss function

$$\underset{\|\boldsymbol{z}\|_0=k}{\text{minimize}} \ \ell(\boldsymbol{z}) := \frac{1}{2m}\sum_{i=1}^{m}\left(\psi_i - |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|\right)^2. \tag{5.2}$$

Clearly, both the objective function and the ℓ_0-norm constraint in (5.2) are nonconvex, which render the optimization problem NP-hard in general [94], and thus computationally intractable. It is worth emphasizing that (thresholded) Wirtinger alternatives dealt with the smooth counterpart of (5.2) based on squared magnitudes $\{y_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|^2\}_{i=1}^m$, which was numerically and experimentally shown to be less effective than the amplitude-based one even when no sparsity is exploited [129, 144]. Although focusing on a formulation similar to (but different than) (5.2), sparse AltMinPhase first estimates the support of the underlying signal, and performs standard phase retrieval of signals with dimension k. More importantly, sparse AltMinPhase relying on alternating minimization with re-sampling entails solving a series of least-squares problems, and performs matrix inversion at every iteration. Numerical tests suggest that a very large number of measurements are required to estimate the support exactly. Once wrong, sparse AltMinPhase confining the phase retrieval task on the estimated support would be impossible to recover the underlying sparse signal. On the other hand, motivated by the iterative hard thresholding (IHT) algorithms for compressive sensing [15, 90], an adaptive hard thresholding procedure that maintains only certain largest entries per iteration during the gradient refinement stage turns out to be effective [18]. Yet both sparse AltMinPhase and ThWF were based on the simple spectral initialization, which was recently shown to be less accurate and robust than the orthogonality-promoting initialization [129].

Broadening the TAF approach and the sparse phase retrieval solver ThWF, the present chapter puts forth a novel iterative solver for (5.2) that proceeds in two stages: S1) a sparse orthogonality-promoting initialization is obtained by solving a PCA-type problem with a few simple power iterations on an estimated support of the underlying sparse signal; and, S2) successive refinements of the initialization are effected by means of a series of truncated gradient

iterations along with a hard thresholding per iteration to set all entries to zero, except for the k ones of largest magnitudes. The two stages are presented in order next.

5.2 Sparse Truncated Amplitude Flow

In this section, the initialization stage and the gradient refinement stage of SPARTA will be described in detail. Assume also without loss of generality that $\|\boldsymbol{x}\|_2 = 1$, which will be justified and generalized shortly.

5.2.1 Sparse orthogonality-promoting initialization

When $\boldsymbol{x}$ is a priori known to be k-sparse with $k \ll n$, one may expect to recover $\boldsymbol{x}$ from a significantly smaller number ($\ll n$) of measurements. The orthogonality-promoting initialization (and spectral based alternatives) requiring m to be on the order of n would fail in the case of phase retrieval for sparse signals given a small number of measurements [91, 22, 32, 129]. By accounting for the sparsity prior information with the ℓ_0 regularization, the same rationale as the orthogonality-promoting initialization in (2.15) would lead to

$$\underset{\|\boldsymbol{z}\|_2=1}{\text{minimize}} \ \boldsymbol{z}^{\mathcal{T}}\boldsymbol{Y}\boldsymbol{z} \quad \text{subject to } \|\boldsymbol{z}\|_0 = k. \tag{5.3}$$

The problem at hand is NP-hard due to the combinatorial constraint. Additionally, it can not be readily converted to a (sparse) PCA problem since the number of data samples available is much smaller than the signal dimension n, thus hardly validating the non-asymptotic result in eq:hard. Although at much higher computational complexity than power iterations, semidefinite relaxation could be applied [39]. Instead of coping with (5.3) directly, we shall take another route and develop our sparse orthogonality-promoting initialization approach to obtain a meaningful sparse initialization from the given limited number of measurements.

Exact support recovery

Along the lines of sparse AltMinPhase and sparse PCA [4], our approach is to first estimate the support of the underlying signal based on a carefully-designed rule; next, we will rely on power iterations to solve (4.1) restricted on the estimated support, thus ensuring a k-sparse estimate

$\tilde{\boldsymbol{z}}^0 \in \mathbb{R}^n$; and, subsequently we will scale $\tilde{\boldsymbol{z}}^0$ by the $\boldsymbol{x}$ norm estimate $\sqrt{\sum_{i=1}^m y_i/m}$ to yield a k-sparse orthogonality-promoting initialization $\boldsymbol{z}^0$.

Starting with the support recovery procedure, assume without loss of generality that $\boldsymbol{x}$ is supported on $\mathcal{S} \subseteq [n] := \{1, \ldots, n\}$ with $|\mathcal{S}| = k \ll n$. Consider the random variables $Z_{i,j} := \psi_i^2 a_{i,j}^2$, $j = 1, \ldots, n$. Recalling that for standardized Gaussian variables, we have $\mathbb{E}[a_{i,j}^4] = 3$, $\mathbb{E}[a_{i,j}^2] = 1$, the rotational invariance property of Gaussian distributions confirms for all $1 \leq j \leq n$ that

$$\begin{aligned} \mathbb{E}[Z_{i,j}] = \mathbb{E}\big[(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x})^2 a_{i,j}^2\big] &= \mathbb{E}\big[a_{i,j}^4 x_j^2 + (\boldsymbol{a}_{i,/j}^{\mathcal{T}}\boldsymbol{x}_{/j})^2 a_{i,j}^2\big] \\ &= 3x_j^2 + \|\boldsymbol{x}_{/j}\|_2^2 \\ &= 2x_j^2 + \|\boldsymbol{x}\|_2^2 \end{aligned} \tag{5.4}$$

where $\boldsymbol{x}_{/j} \in \mathbb{R}^{n-1}$ is obtained by deleting the j-th entry from $\boldsymbol{x} \in \mathbb{R}^n$; and likewise for $\boldsymbol{a}_{i,/j} \in \mathbb{R}^{n-1}$. If $j \in \mathcal{S}$, then $x_j \neq 0$ yielding $\mathbb{E}[Z_{i,j}] = \|\boldsymbol{x}\|_2^2 + 2x_j^2$ in (5.4). If on the other hand $j \notin \mathcal{S}$, it holds that $x_j = 0$, which leads to $\mathbb{E}[Z_{i,j}] = \|\boldsymbol{x}_{/j}\|_2^2 = \|\boldsymbol{x}\|_2^2$. It is now clear that there is a separation of $2x_j^2$ in the expected values of $Z_{i,j}$ for $j \in \mathcal{S}$ and $j \notin \mathcal{S}$. As long as the gap $2x_j^2$ is sufficiently large, the support set $\mathcal{S}$ can be recovered exactly in this way. Specifically, when all $\mathbb{E}[Z_{i,j}]$ values are available, the set of indices corresponding to the k-largest $\mathbb{E}[Z_{i,j}]$ values recover exactly the support of $\boldsymbol{x}$. In practice, $\{\mathbb{E}[Z_{i,j}]\}$ are not available. One has solely access to a number of their independent realizations. Appealing to the strong law of large numbers, the sample average approaches the ensemble one, namely, $\hat{Z}_{i,j} := (1/m)\sum_{i=1}^m Z_{i,j} \to \mathbb{E}[Z_{i,j}]$ as m increases. Hence, the support can be estimated as

$$\hat{\mathcal{S}} := \big\{1 \leq j \leq n \big| \text{indices of top-}k \text{ instances in } \{\hat{Z}_{i,j}\}_{j=1}^n \big\} \tag{5.5}$$

which will be shown to recover $\mathcal{S}$ exactly with high probability provided that $\mathcal{O}(k^2 \log n)$ measurements are taken and the minimum nonzero entry $x_{\min} := \min_{j \in S} |x_j|$ is on the order of $(1/\sqrt{k})\|\boldsymbol{x}\|_2$. The latter is postulated to guarantee such a separation between quantities having their indices belonging or not belonging to the support set. It is worth stressing that $k^2 \log n \ll n$ when $k \ll n$, hence largely reducing the sampling size and also the computational complexity.

Orthogonality-promoting initialization

When the estimated support in (5.5) turns out to be exact, i.e., $\hat{\mathcal{S}} = \mathcal{S}$, one can rewrite $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}| = |\boldsymbol{a}_{i,\hat{\mathcal{S}}}^{\mathcal{T}}\boldsymbol{x}_{\hat{\mathcal{S}}}|$, $i = 1, \ldots, m$, where $\boldsymbol{a}_{i,\hat{\mathcal{S}}} \in \mathbb{R}^k$ includes the j-th entry $a_{i,j}$ of $\boldsymbol{a}_i$ if and only if $j \in \hat{\mathcal{S}}$; and likewise for $\boldsymbol{x}_{\hat{\mathcal{S}}} \in \mathbb{R}^k$. Instead of seeking directly an n-dimensional initialization as in (5.3), one can apply the orthogonality-promoting initialization steps on the dimensionality reduced data $\{(\boldsymbol{a}_{i,\hat{\mathcal{S}}}, \psi_i)\}_{i=1}^m$ to produce a k-dimensional vector

$$\tilde{\boldsymbol{z}}_{\hat{\mathcal{S}}}^0 := \arg\max_{\|\boldsymbol{z}_{\hat{\mathcal{S}}}\|_2=1} \frac{1}{|\bar{\mathcal{I}}^0|}\boldsymbol{z}_{\hat{\mathcal{S}}}^{\mathcal{T}}\Big(\sum_{i\in\bar{\mathcal{I}}^0}\frac{\boldsymbol{a}_{i,\hat{\mathcal{S}}}\boldsymbol{a}_{i,\hat{\mathcal{S}}}^{\mathcal{T}}}{\|\boldsymbol{a}_{i,\hat{\mathcal{S}}}\|_2^2}\Big)\boldsymbol{z}_{\hat{\mathcal{S}}} \tag{5.6}$$

and subsequently reconstruct a k-sparse n-dimensional initialization $\tilde{\boldsymbol{z}}^0$ by zero-padding $\tilde{\boldsymbol{z}}_{\hat{\mathcal{S}}}^0$ at entries with indices not belonging to $\hat{\mathcal{S}}$. Similarly, in the case of $\|\boldsymbol{x}\|_2 \neq 1$, $\tilde{\boldsymbol{z}}^0$ in (5.6) is rescaled by the norm estimate of $\boldsymbol{x}$ to obtain $\boldsymbol{z}^0 = \sqrt{\sum_{i=1}^m y_i/m}\,\tilde{\boldsymbol{z}}^0$. We also note that our proposed algorithm can recover the underlying sparse signal when $\hat{\mathcal{S}} \neq \mathcal{S}$, as long as $\boldsymbol{z}^0$ is sufficiently close to $\boldsymbol{x}$ regardless of support mismatch, which is described further in Lemma 12.

5.2.2 Thresholded truncated gradient iterations

Upon obtaining a sparse orthogonality-promoting initialization $\boldsymbol{z}^0$, our approach to solving (5.2) boils down to iteratively refining $\boldsymbol{z}^0$ by means of a series of k-sparse hard thresholding based truncated gradient iterations, namely,

$$\boldsymbol{z}^{t+1} := \mathcal{H}_k\big(\boldsymbol{z}^t - \mu\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}^t)\big), \quad t = 0, 1, \ldots \tag{5.7}$$

where t is the iteration index, $\mu > 0$ a constant step size, and $\mathcal{H}_k(\boldsymbol{u}) : \mathbb{R}^n \to \mathbb{R}^n$ denotes a k-sparse hard thresholding operation that sets all entries in $\boldsymbol{u}$ to zero except for the k entries of largest magnitudes. If there are multiple such sets comprising the k-largest entries, a set can be chosen either randomly or according to a predefined ordering of the elements. Similar to [129], the truncated (generalized) gradient $\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}^t)$ is

$$\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}^t) := \frac{1}{m}\sum_{i\in\mathcal{I}^{t+1}}\Big(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t - \psi_i\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_i \tag{5.8}$$

Algorithm 6 SPARse Truncated Amplitude flow (SPARTA)

1: **Input:** Data $\{(\boldsymbol{a}_i;\psi_i)\}_{i=1}^m$ and sparsity level k; maximum number of iterations $T=1,000$; step size $\mu=1$, truncation thresholds $|\overline{\mathcal{I}}^0|=\lceil\frac{1}{6}m\rceil$, and $\gamma=1$.

2: **Set** $\hat{\mathcal{S}}$ to include indices corresponding to the k-largest instances in $\big\{\sum_{i=1}^m \psi_i^2|a_{i,j}|^2/m\big\}_{j=1}^n$.

3: **Evaluate** $\overline{\mathcal{I}}^0$ to consist of indices of the top-$|\overline{\mathcal{I}}^0|$ values in $\{\psi_i/\|\boldsymbol{a}_{i,\hat{\mathcal{S}}}\|_2\}_{i=1}^m$ with $\boldsymbol{a}_{i,\hat{\mathcal{S}}}\in\mathbb{R}^k$ removing entries of $\boldsymbol{a}_i\in\mathbb{R}^n$ not belonging to $\hat{S}$; and compute the principal eigenvector $\tilde{\boldsymbol{z}}_{\hat{\mathcal{S}}}^0\in\mathbb{R}^k$ of matrix

$$\boldsymbol{Y}:=\frac{1}{|\overline{\mathcal{I}}^0|}\sum_{i\in\overline{\mathcal{I}}^0}\frac{\boldsymbol{a}_{i,\hat{\mathcal{S}}}\boldsymbol{a}_{i,\hat{\mathcal{S}}}^{\mathcal{T}}}{\|\boldsymbol{a}_{i,\hat{\mathcal{S}}}\|_2^2}$$

based on 100 power iterations.

4: **Initialize** $\boldsymbol{z}^0$ as $\sqrt{\sum_{i=1}^m\psi_i^2/m}\,\tilde{\boldsymbol{z}}^0$, where $\tilde{\boldsymbol{z}}^0\in\mathbb{R}^n$ is obtained by augmenting $\tilde{\boldsymbol{z}}_{\hat{\mathcal{S}}}^0$ in Step 3 with zeros at entries with their indices not in $\hat{\mathcal{S}}$.

5: **Loop: For** $t=0$ **to** $T-1$

$$\boldsymbol{z}^{t+1}=\mathcal{H}_k\Big(\boldsymbol{z}^t-\frac{\mu}{m}\sum_{i\in\mathcal{I}^{t+1}}\Big(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t-\psi_i\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_i\Big)$$

where $\mathcal{I}^{t+1}=\big\{1\le i\le m\big||\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|\ge\psi_i/(1+\gamma)\big\}$, and $\mathcal{H}_k(\boldsymbol{u}):\mathbb{R}^n\to\mathbb{R}^n$ sets all entries of $\boldsymbol{u}$ to zero except for the k-ones of largest magnitudes.

6: **Output:** $\boldsymbol{z}^T$.

where the index set is defined to be

$$\mathcal{I}^{t+1}:=\Big\{1\le i\le m\Big|\frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\ge\frac{1}{1+\gamma}\Big\}\tag{5.9}$$

for some $\gamma>0$ to be determined shortly, where $\{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|=\psi_i\}$ are the given modulus data.

It is clear now that the difficulty of minimizing our nonconvex objective function reduces to that of correctly estimating the signs of $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}$ by $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|$ at each iteration. The truncation rule in (5.9) was shown capable of eliminating most "bad" gradient components involving erroneously estimated signs, i.e., $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|\ne\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}/|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$. This rule improved performance of TAF [129] considerably. Recall that our objective function in (5.2) is also non-smooth at points $\boldsymbol{z}\in\mathbb{R}^n$ obeying $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}=0$. Evidently, the gradient regularization rule in (5.9) keeps only the gradients of component functions (i.e., the summands in (5.2)) that bear large enough $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|$ values; this rule thus maintains $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t$ away from 0 and protects the cost function in (5.2) from

being non-smooth at points satisfying $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z} = 0$. As a consequence, the (truncated) generalized gradient employed in (5.8) reduces to the (truncated) gradient at such points, which also simplifies theoretical convergence analysis.

5.3 Main Results

The proposed sparse phase retrieval solver is summarized in Alg. 6.

Theorem 4 (Exact recovery). *Fix $\boldsymbol{x} \in \mathbb{R}^n$ to be any k-sparse ($k \ll n$) vector of the minimum nonzero entry on the order of $(1/\sqrt{k})\|\boldsymbol{x}\|_2$, namely, $x_{\min}^2 = (C_1/k)\|\boldsymbol{x}\|_2^2$ for some number $C_1 > 0$. Consider the m noiseless measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ from i.i.d. $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le i \le m$. If $m \ge C_0 k^2 \log(mn)$, Step 3 of SPARTA (tabulated in Alg. 6) recovers the support of $\boldsymbol{x}$ exactly with probability at least $1 - 6/m$. Furthermore, there exist numerical constants $\underline{\mu},\ \overline{\mu} > 0$ such that with a fixed step size $\mu \in [\underline{\mu}, \overline{\mu}]$, and a truncation threshold $\gamma = +\infty$, successive estimates of SPARTA obey*

$$\operatorname{dist}(\boldsymbol{z}^t, \boldsymbol{x}) \le \frac{1}{10}(1-\nu)^t \|\boldsymbol{x}\|_2, \quad t = 0, 1, \ldots \tag{5.10}$$

which holds with probability exceeding $1 - c_1 m \mathrm{e}^{-c_0 k} - 7/m$ provided that $m \ge C_2|\overline{\mathcal{I}}^0| \ge C_0 k^2 \log(mn)$. Here, c_0, c_1, C_0, C_2, and $0 < \nu < 1$ are some numerical constants.

Proof of Thm. 4 is deferred to Sec. 5.5 with supporting lemmas presented in the Appendix. We typically take parameters $|\overline{\mathcal{I}}^0| = \lceil \frac{1}{6}m \rceil$, and $\mu = 1$, which will also be validated by our analytical results on the feasible region of the step size. The constant C_0 depends on C_1, ν on μ and C_1, and $\underline{\mu}$ and $\overline{\mu}$ rely on both C_1 and C_0. In the case of phase retrieval of unstructured signals, existing algorithms such as TAF ensures exact recovery when the number of measurements m is about the number of unknowns n, i.e., $m \gtrsim n$. Hence, it would be more meaningful to study the sample complexity bound for phase retrieval of sparse signals when $m \lesssim n$. To this end, the sample complexity bound $m \ge C_0 k^2 \log(mn)$ in Thm. 4 can often be rewritten as $m \ge C_0' k^2 \log n$ for some constant $C_0' > C_0$ and large enough n. Regarding Thm. 4, three observations are in order.

Remark 6. SPARTA recovers exactly any k-sparse signal $\boldsymbol{x}$ of minimum nonzero entries on the order of $(1/\sqrt{k})\|\boldsymbol{x}\|_2$ when there are about $k^2 \log n$ magnitude-only measurements, which

coincides with the number of measurements required by the state-of-the-art algorithms such as CPRL [92], sparse AltMinPhase [91], and ThWF [18].

Remark 7. SPARTA converges at a linear rate to the globally optimal solution $\boldsymbol{x}$ with convergence rate independent of the signal dimension n. In other words, for any given solution accuracy $\epsilon > 0$, after running at most $T = \log(1/\epsilon)$ SPARTA iterations (5.7), the returned estimate $\boldsymbol{z}^T$ is at most $\epsilon\|\boldsymbol{x}\|_2$ away from the global solution $\boldsymbol{x}$.

Remark 8. SPARTA enjoys a low computational complexity of $\mathcal{O}(k^2 n \log n)$, and incurs a total runtime of $\mathcal{O}(k^2 n \log n \log(1/\epsilon))$ to produce an ϵ-accurate solution. The runtime is proportional to the time $\mathcal{O}(k^2 n \log n)$ taken to read the data $\{(\boldsymbol{a}_i, \psi_i)\}_{i=1}^m$. To see this, recall that the support recovery incurs computational complexity $\mathcal{O}(k^2 n \log n + n \log n)$, power iterations incur complexity $\mathcal{O}(k^2 n \log n)$, and thresholded truncated gradient iterations have complexity $\mathcal{O}(k^2 n \log n)$; hence, leading to a total complexity on the order of $k^2 n \log n$. Given the linear convergence rate, SPARTA takes a total runtime of $\mathcal{O}(k^2 n \log n \log(1/\epsilon))$ to achieve any fixed solution accuracy $\epsilon > 0$.

Besides exact recovery guarantees in the case of noiseless measurements, it is worth mentioning that SPARTA exhibits robustness to additive noise, especially when the noise has bounded values. Numerical results using SPARTA for noisy sparse phase retrieval will be presented in the ensuing section.

5.4 Numerical Experiments

Simulated tests examining performance of SPARTA against truncated amplitude flow (TAF) [129] (which does not exploit the sparsity) and thresholded Wirtinger flow (ThWF) [18] are presented in this section. The initialization in each scheme was obtained based upon 100 power iterations, and was subsequently refined by $T = 1,000$ gradient iterations. In all reported experiments, the true k-sparse signal vector $\boldsymbol{x} \in \mathbb{R}^n$ or $\mathbb{C}^n$ was generated first using $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ or $\mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$, followed by setting $(n-k)$ of its n entries to zero uniformly at random. For reproducibility, the Matlab implementation of SPARTA is available at `https://gangwg.github.io/SPARTA/`.

The first experiment evaluates the exact recovery performance of various approaches, where the true signals are real. We fixed the signal dimension to $n = 1,000$, and the sparsity level at $k = 10$, while m/n increases from 0.1 to 3 by 0.1. Curves in Fig. 5.1 clearly demonstrate

markedly improved performance of SPARTA over state-of-the-art alternatives. Even when the exact number of nonzero elements in $\boldsymbol{x}$, namely, k is unknown, setting k in Alg. 6 as an upper limit on the theoretically affordable sparsity level (e.g., $\lceil \sqrt{n} \rceil$ when m is about n according to Thm. 4) works well too (see the magenta curve, denoted SPARTA0). Comparison between TAF and SPARTA shows the advantage of exploiting sparsity in sparse phase retrieval settings.

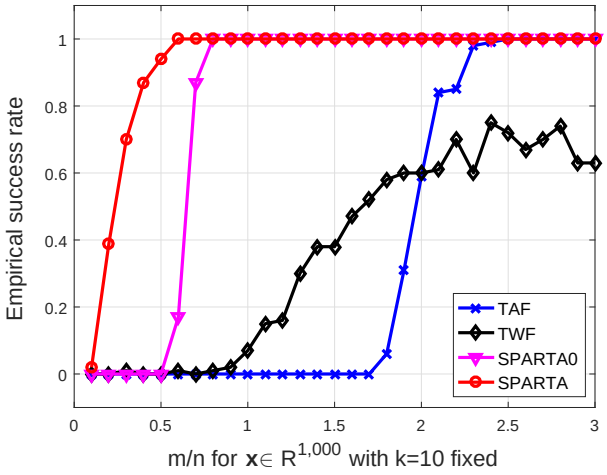

Figure 5.1: Empirical success rate versus m/n.

The second experiment examines how SPARTA recovers real signals of various sparsity levels given a fixed number of measurements. Figure 5.2 depicts the empirical success rate versus the sparsity level k, where k equals the exact number of nonzero entries in $\boldsymbol{x}$. The results suggest that with a total of $m = n$ phaseless quadratic equations, TAF representing the state-of-the-art for phase retrieval of unstructured signals fails, as shown by the blue curve. Although ThWF works in some cases, SPARTA significantly outperforms ThWF, and it ensures exact recovery of sparse signals with up to about $25 < \sqrt{n} \approx 32$ nonzero entries (due to existence of polylog factors in the sample complexity), hence justifying our analytical results.

The next experiment validates the robustness of SPARTA against additive noise present in the data. Postulating the noisy Gaussian data model $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}| + \eta_i$ [91], we generated i.i.d. Gaussian noise according to $\eta_i \sim \mathcal{N}(0,\, 0.1^2)$, $i = 1, \ldots, m$. From Fig. 5.1, it is clear that to achieve exact recovery, SPARTA requires about $m = 6k^2 = 600$ measurements, TAF about $3n = 3,000$ measurements, and ThWF much more than $3,000$. In this case, parameters were taken as $n = 1,000$, $m = 3,000$, and $k = 10$, with the number of measurements large

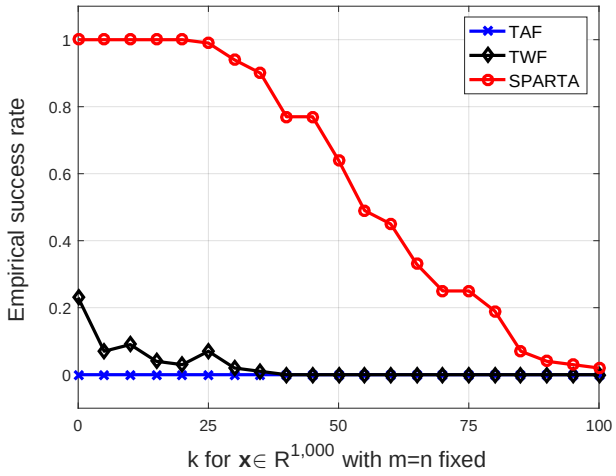

Figure 5.2: Empirical success rate versus sparsity level k.

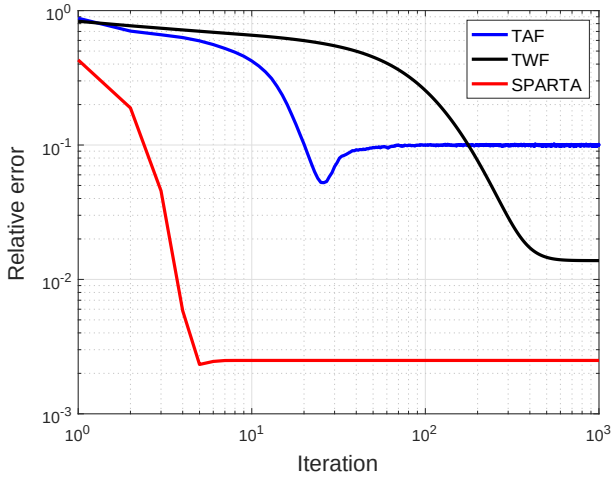

Figure 5.3: Convergence behavior for noisy data with $n = 1,000$, $m = 3,000$, and $k = 10$.

enough to guarantee that ThWF and TAF also work. It is worth mentioning that SPARTA can work with a far smaller number of measurements than $m = 3,000$. As seen from the plots, SPARTA performs only a few gradient iterations to achieve the most accurate solution among the three approaches, while its competing TAF and ThWF require nearly an order more number of iterations to converge to less accurate estimates.

To demonstrate the stability of SPARTA against additive noise, the relative MSE is plotted as

a function of the SNR values in dB. Our experiments are based on the additive Gaussian noise model $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}| + \eta_i$ with a 10-sparse signal $\boldsymbol{x} \in \mathbb{R}^{1,000}$ and the noise $\boldsymbol{\eta} := [\eta_1 \ \cdots \ \eta_m]^{\mathcal{T}} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_m)$, where σ^2 is chosen such that certain $\mathrm{SNR} := 10\log_{10}\sum_{i=1}^{m}|\langle\boldsymbol{a}_i,\boldsymbol{x}\rangle|^2/\sigma^2$ values are achieved. The ratio m/n takes values $\{1,\ 2,\ 3\}$, and the SNR in dB is varied from 5 dB to 55 dB. Averaging over 100 Monte Carlo realizations, Fig. 5.4 demonstrates that the relative MSE for all m/n values scales inversely proportional to SNR, hence corroborating the stability of SPARTA in the presence of additive noise.

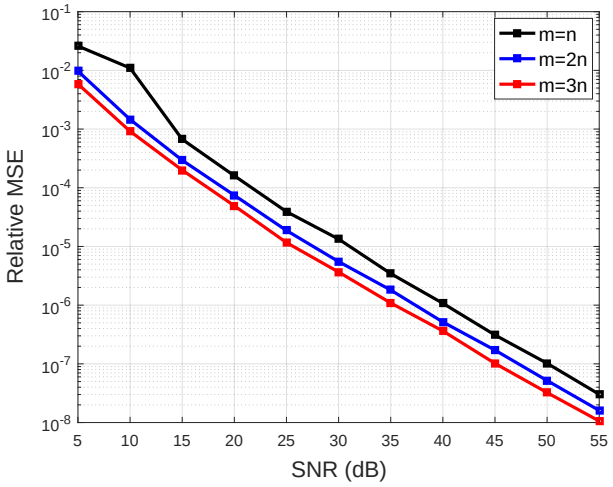

Figure 5.4: Relative MSE versus SNR for SPARTA with the AWGN model.

The last experiment tested the efficacy of SPARTA in the complex-valued setting, where the underlying 10-sparse signal $\boldsymbol{x} \in \mathbb{C}^{20,000}$ was generated using $\boldsymbol{x} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_{20,000}) := \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{20,000}/2) + j\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{20,000}/2)$, and $\boldsymbol{a}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_{20,000})$ for $1 \le i \le 1,000$. The relative MSE versus iteration count was plotted in Fig. 5.5, which validates the scalability and effectiveness of SPARTA in recovering complex signals. In terms of runtime, SPARTA recovers exactly a $20,000$-dimensional complex-valued signal from $1,000$ magnitude-only measurements in a few seconds.

Regarding computation times, SPARTA converges much faster (both in time and in the number of iterations required to achieve certain solution accuracy) than ThWF and TAF in all reported experiments.

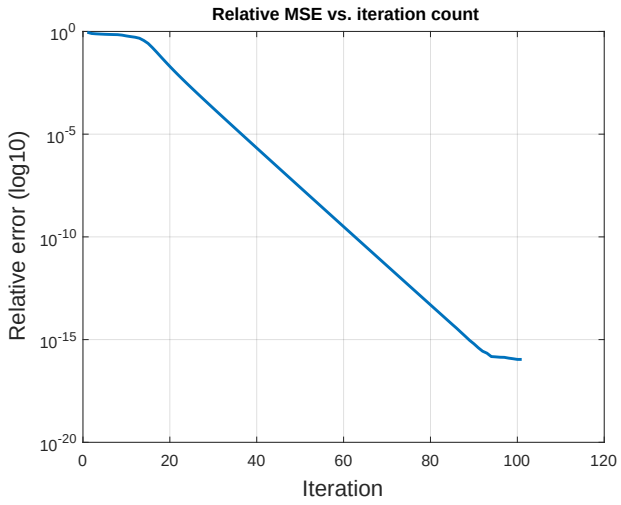

Figure 5.5: Relative MSE versus iteration count for SPARTA in the complex setting.

5.5 Proofs

The proof of Thm. 4 will be provided in this section. To that end, we will first evaluate the performance of our sparse orthogonality-promoting initialization. The following result demonstrates that if the number of measurements is sufficiently large (on the order of k^2 within polylog factors), Step 3 of Alg. 6 reconstructs the support of $\boldsymbol{x}$ exactly with high probability.

Lemma 10. *Consider any k-sparse signal $\boldsymbol{x} \in \mathbb{R}^n$ with support $\mathcal{S}$ and minimum nonzero entries $x_{\min} := \min_{j\in\mathcal{S}} |x_j|$ on the order of $(1/\sqrt{k})\|\boldsymbol{x}\|_2$. If the sensing vectors $\{\boldsymbol{a}_i\}_{i=1}^m$ are i.i.d standard Gaussian, i.e., $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, Step 3 in Alg. 6 recovers $\mathcal{S}$ exactly with probability at least $1 - 6/m$ provided $m \geq C_0 k^2 \log(mn)$ for some absolute constant $C_0 > 0$.*

Upon obtaining the support of the underlying sparse signal, SPARTA subsequently employs the orthogonality-promoting initialization on the reduced-dimension data $\{(\psi_i, \boldsymbol{a}_{i,\hat{S}})\}$. Based on results in [129, Prop. 1], the estimate

$$\boldsymbol{z}_{\hat{S}}^0 := \sqrt{\sum_{i=1}^m \psi_i^2 / m \tilde{\boldsymbol{z}}_{\hat{S}}^0}$$

obtained from Step 3 in Alg. 6 satisfies $\operatorname{dist}(\boldsymbol{z}_{\hat{S}}^0, \boldsymbol{x}_{\hat{S}}) \leq (1/10)\|\boldsymbol{x}_{\hat{S}}\|_2$ with high probability provided that m/k is sufficiently large and k large enough as well. Putting together this result,

Lemma 10, and Step 4 in Alg. 6 leads to the following lemma, which formally summarizes the theoretical performance of our proposed sparse orthogonality-promoting initialization.

Lemma 11. *Let* $\boldsymbol{z}^0 = \sqrt{\sum_{i=1}^m \psi_i^2/m}\,\tilde{\boldsymbol{z}}^0$ *be given by Step* 4*, and* $\tilde{\boldsymbol{z}}^0$ *obtained through the sparse orthogonality-promoting initialization Step* 3 *in Alg. 6. With probability at least* $1-(m+6)\exp(-k/2)-7/m$*, the following holds*

$$\operatorname{dist}(\boldsymbol{z}_0,\, \boldsymbol{x}) \le (1/10)\|\boldsymbol{x}\|_2 \tag{5.11}$$

provided that $m \ge C_0' k$ *for some absolute constant* $C_0' > 0$*.*

The proof can be directly adapted from [129, Prop. 1], and hence it is omitted.

Lemma 12. *Take a constant learning parameter* $\mu \in (\underline{\mu},\, \overline{\mu})$*. There exists an event of probability at least* $1 - c_1 m^{-c_0 k}$*, such that on this event, starting from an initial estimate* $\boldsymbol{z}^0$ *satisfying* $\operatorname{dist}(\boldsymbol{z}^0,\, \boldsymbol{x}) \le (1/10)\|\boldsymbol{x}\|_2$*, successive estimates by Step 5 with* $\gamma = +\infty$ *in Alg. 6 obey*

$$\operatorname{dist}(\boldsymbol{z}^t,\, \boldsymbol{x}) \le (1/10)(1-\nu)^t \|\boldsymbol{x}\|_2, \quad t = 0,\, 1,\, \ldots \tag{5.12}$$

if $m \ge C_0''(3k)\log(n/(3k))$*. Here,* $\underline{\mu}$*,* $\overline{\mu}_0$*,* c_0*,* c_1*,* $C_0'' > 0$ *are certain universal constants.*

It is worth noting that Step 5 of Alg. 6 guarantees linear convergence to the globally optimal solution $\boldsymbol{x}$ as long as the initial guess $\boldsymbol{z}^0$ lands within a small neighborhood of $\boldsymbol{x}$, regardless of whether $\boldsymbol{z}^0$ estimates exactly the support of $\boldsymbol{x}$ or not.

Proof of Lemma 12. To start, let us establish a bit of notation, which will be used only in this section. Define for all $t \ge 0$:

$$\boldsymbol{d}^{t+1} := \boldsymbol{z}^t - \frac{\mu}{m}\sum_{i=1}^m \Big(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t - \psi_i \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}\Big)\boldsymbol{a}_i$$

which represents the estimate prior to the hard thresholding operation in (5.7). With $\mathcal{S}$ and $\hat{\mathcal{S}}^t$ denoting the support set of $\boldsymbol{x}$ and $\boldsymbol{z}^t$, respectively, the reconstruction error $\boldsymbol{x} - \boldsymbol{z}^{t+1}$ is therefore supported on the set $\Theta^{t+1} := \mathcal{S} \cup \hat{\mathcal{S}}^{t+1}$; and likewise, $\boldsymbol{x} - \boldsymbol{z}^t$ is supported on $\Theta^t := \mathcal{S} \cup \hat{\mathcal{S}}^t$. In addition, define the difference between sets Θ^t and Θ^{t+1} as $\Theta^t \setminus \Theta^{t+1}$, which consists of all elements of Θ^t that are not elements of Θ^{t+1}. It is then clear that $|\mathcal{S}| = |\hat{\mathcal{S}}^t| = k$, $|\Theta^t| \le 2k$,

and $|\Theta^t \setminus \Theta^{t+1}| \leq 2k$ as well as $|\Theta^t \cup \Theta^{t+1}| \leq 3k$ for all $t \geq 0$. When using these sets as subscript, for instance, $\boldsymbol{d}_{\Theta^t}$, we mean vectors formed by deleting all but those elements from the vector other than those in the set.

The proof of Lemma 12 will be mainly based on results in [129], and [15], [90]. The former helps establishing the so-termed local regularity condition that will be key to proving linear convergence of iterative optimization algorithms to the globally optimal solutions of nonconvex optimization problems [22], while the latter two offer a standard approach to dealing with the nonlinear hard thresholding operator involved in our proposed SPARTA algorithm. Specifically, based on the triangle inequality of the vector 2-norm, one arrives at

$$\begin{aligned}\big\|\boldsymbol{x}_{\Theta^{t+1}} - \boldsymbol{z}^{t+1}_{\Theta^{t+1}}\big\|_2 &= \big\|\boldsymbol{x}_{\Theta^{t+1}} - \boldsymbol{d}^{t+1}_{\Theta^{t+1}} + \boldsymbol{d}^{t+1}_{\Theta^{t+1}} - \boldsymbol{z}^{t+1}_{\Theta^{t+1}}\big\|_2 \\ &\leq \big\|\boldsymbol{x}_{\Theta^{t+1}} - \boldsymbol{d}^{t+1}_{\Theta^{t+1}}\big\|_2 + \big\|\boldsymbol{z}^{t+1}_{\Theta^{t+1}} - \boldsymbol{d}^{t+1}_{\Theta^{t+1}}\big\|_2 \end{aligned} \tag{5.13}$$

where in the last inequality the first term denotes the distance of $\boldsymbol{x}_{\Theta^{t+1}}$ to the estimate $\boldsymbol{d}^{t+1}_{\Theta^{t+1}}$ before hard thresholding, and the second denotes the distance between $\boldsymbol{d}^{t+1}_{\Theta^{t+1}}$ and its best k-approximation $\boldsymbol{z}^{t+1}_{\Theta^{t+1}}$ because $\boldsymbol{z}^{t+1}_{\Theta^{t+1}}$ has cardinality equal to k. The optimality of $\boldsymbol{z}^{t+1}_{\Theta^{t+1}}$ implies

$$\|\boldsymbol{z}^{t+1}_{\Theta^{t+1}} - \boldsymbol{d}^{t+1}_{\Theta^{t+1}}\|_2 \leq \|\boldsymbol{x}_{\Theta^{t+1}} - \boldsymbol{d}^{t+1}_{\Theta^{t+1}}\|_2.$$

Plugging the latter inequality back into (5.13) yields

$$\big\|\boldsymbol{x}_{\Theta^{t+1}} - \boldsymbol{z}^{t+1}_{\Theta^{t+1}}\big\|_2 \leq 2\big\|\boldsymbol{x}_{\Theta^{t+1}} - \boldsymbol{d}^{t+1}_{\Theta^{t+1}}\big\|_2. \tag{5.14}$$

Define the estimation error $\boldsymbol{h}^t := \boldsymbol{x} - \boldsymbol{z}^t$. Rewriting and substituting

$$\boldsymbol{d}^{t+1} = \boldsymbol{z}^t - \frac{\mu}{m}\sum_{i=1}^m \big(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t - \boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}\big)\boldsymbol{a}_i + \frac{\mu}{m}\sum_{i=1}^m \Big(\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|} - \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\Big)|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\boldsymbol{a}_i \tag{5.15}$$

into (5.14) leads to

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{h}^{t+1}_{\Theta^{t+1}}\|_2 &\leq \Big\|\boldsymbol{h}^t_{\Theta^{t+1}} - \frac{\mu}{m}\sum_{i=1}^m \boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}^t\boldsymbol{a}_{i,\Theta^{t+1}} - \frac{\mu}{m}\sum_{i=1}^m \Big(\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|} - \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\Big)|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\boldsymbol{a}_{i,\Theta^{t+1}}\Big\|_2 \\ &= \Big\|\boldsymbol{h}^t_{\Theta^{t+1}} - \frac{\mu}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}^t_{\Theta^{t+1}} - \frac{\mu}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^t\setminus\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}^t_{\Theta^t\setminus\Theta^{t+1}}\end{aligned}$$

$$
\begin{aligned}
&- \frac{\mu}{m}\sum_{i=1}^{m}\Big(\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}-\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\Big)|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\boldsymbol{a}_{i,\Theta^{t+1}}\Big\|_2 \\
\leq\ &\Big\|\boldsymbol{h}^t_{\Theta^{t+1}}-\frac{\mu}{m}\sum_{i=1}^{m}\boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}^t_{\Theta^{t+1}}\Big\|_2 \\
&+\Big\|\frac{\mu}{m}\sum_{i=1}^{m}\boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^t\setminus\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}^t_{\Theta^t\setminus\Theta^{t+1}}\Big\|_2 \\
&+\Big\|\frac{\mu}{m}\sum_{i=1}^{m}\Big(\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}-\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\Big)|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\boldsymbol{a}_{i,\Theta^{t+1}}\Big\|_2
\end{aligned}
\tag{5.16}
$$

where the equality follows from re-expressing

$$
\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}^t=\boldsymbol{a}_{i,\Theta^t}^{\mathcal{T}}\boldsymbol{h}^t_{\Theta^t}=\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}^t_{\Theta^{t+1}}+\boldsymbol{a}_{i,\Theta^t\setminus\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}^t_{\Theta^t\setminus\Theta^{t+1}}
$$

since $\boldsymbol{h}^t=\boldsymbol{x}-\boldsymbol{z}^t$ is supported on Θ^t. The last inequality is readily obtained with triangle inequality of the ℓ_2-norm.

The task now remains to establish upper bounds for the three terms appearing on the right hand side of (5.16), which will be the subject for the rest of this section. Toward this end, let us recall the concept of the so-called restricted isometry property (RIP) condition in compressive sampling [24]. For each integer $s=1,\,2,\,\ldots,k$, define the isometry constant $0<\delta_s<1$ of a matrix $\boldsymbol{\Phi}\in\mathbb{R}^{m\times n}$ as the smallest quantity such that the following holds for all k-sparse vectors $\boldsymbol{v}\in\mathbb{R}^n$ [24, 90]:

$$
(1-\delta_k)\|\boldsymbol{v}\|_2^2\leq\|\boldsymbol{\Phi}\boldsymbol{v}\|_2^2\leq(1+\delta_k)\|\boldsymbol{v}\|_2^2. \tag{5.17}
$$

For Gaussian matrix $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ whose entries are i.i.d. standard normal variables, then $\frac{1}{\sqrt{m}}\boldsymbol{A}$ satisfies the RIP with constant $\delta_{3k}\leq\epsilon$ with probability at least $1-\mathrm{e}^{-c_0'm}$, provided that $m\geq C_1'\epsilon^{-2}(3k)\log(n/(3k))$ for certain universal constants $c_0',\,C_1'>0$ [24], [90, Eq. (1.2)]. Furthermore, if $\mathcal{K}\subsetneqq\{1,\,2,\,\ldots,\,n\}$ is a set of $3k$ indices or fewer, the following properties of $\boldsymbol{A}$ hold true [90, Prop. 3.1]:

P1) $\|\boldsymbol{A}_{\mathcal{K}}^{\mathcal{T}}\boldsymbol{u}\|_2\leq\sqrt{(1+\delta_{3k})m}\|\boldsymbol{u}\|_2$, for all $\boldsymbol{u}\in\mathbb{R}^m$;

P2) $(1-\delta_{3k})m\|\boldsymbol{v}\|_2\leq\|\boldsymbol{A}_{\mathcal{K}}^{\mathcal{T}}\boldsymbol{A}_{\mathcal{K}}\boldsymbol{v}\|_2\leq(1+\delta_{3k})m\|\boldsymbol{v}\|_2$, for all at most $3k$-sparse vectors $\boldsymbol{v}\in\mathbb{R}^n$;

P3) $\|\boldsymbol{A}_{\mathcal{B}}^{\mathcal{T}}\boldsymbol{A}_{\mathcal{D}}\|_2\leq\delta_{3k}$, where $\mathcal{B}$ and $\mathcal{D}$ are disjoint sets of combined cardinality not exceeding

$3k$;

P4) $\|\boldsymbol{A}_{\mathcal{B}\cup\mathcal{D}}^{\mathcal{T}}\boldsymbol{A}_{\mathcal{B}\cup\mathcal{D}} - \boldsymbol{I}\|_2 \leq \delta_{3k}$.

Having elaborated on the properties of RIP matrices, we are ready to derive bounds for the three terms on the right hand side of (5.16). Regarding the first term, it is easy to check that

$$\begin{aligned}\Big\|\boldsymbol{h}_{\Theta^{t+1}}^t - \frac{\mu}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}_{\Theta^{t+1}}^t\Big\|_2 &= \Big\|\Big(\boldsymbol{I} - \frac{\mu}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\Big)\boldsymbol{h}_{\Theta^{t+1}}^t\Big\|_2 \\ &\leq \Big\|\boldsymbol{I} - \frac{\mu}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\Big\|_2\big\|\boldsymbol{h}_{\Theta^{t+1}}^t\big\|_2 \\ &\leq \max\big\{1-\mu\underline{\lambda},\, \mu\overline{\lambda}-1\big\}\big\|\boldsymbol{h}_{\Theta^{t+1}}^t\big\|_2 \qquad (5.18)\end{aligned}$$

where $\overline{\lambda}$, $\underline{\lambda} > 0$ are the largest and smallest eigenvalue of $(1/m)\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}$, respectively. Specifically, the two inequalities in (5.18) are obtained based on the definition of the induced 2-norm (i.e., the spectral norm) of matrices.

Next, we estimate the eigenvalues $\overline{\lambda}$ and $\underline{\lambda}$. Using P2, it clearly holds that

$$\overline{\lambda} = \lambda_{\max}\Big(\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\Big) \leq 1 + \delta_{2k} \qquad (5.19)$$

due to $|\Theta^{t+1}| \leq 2k$. For the same reason, it further holds that

$$\underline{\lambda} = \lambda_{\min}\Big(\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\Big) \geq 1 - \delta_{2k}. \qquad (5.20)$$

Taking the results in (5.19) and (5.20) into (5.18) yields

$$\begin{aligned}&\Big\|\boldsymbol{h}_{\Theta^{t+1}}^t - \frac{\mu}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}_{\Theta^{t+1}}^t\Big\|_2 \\ &\leq \max\big\{1-\mu(1-\delta_{2k}),\, \mu(1+\delta_{2k})-1\big\}\big\|\boldsymbol{h}_{\Theta^{t+1}}^t\big\|_2. \qquad (5.21)\end{aligned}$$

For the second term in (5.16), since $|\Theta^{t+1}\cup\Theta^t| \leq 3k$, the next holds with high probability

$$\Big\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^t\setminus\Theta^{t+1}}^{\mathcal{T}}\boldsymbol{h}_{\Theta^t\setminus\Theta^{t+1}}^t\Big\|_2 \leq \Big\|\frac{1}{m}\sum_{i=1}^m \boldsymbol{a}_{i,\Theta^{t+1}}\boldsymbol{a}_{i,\Theta^t\setminus\Theta^{t+1}}^{\mathcal{T}}\Big\|_2\big\|\boldsymbol{h}_{\Theta^t\setminus\Theta^{t+1}}^t\big\|_2$$

$$\leq \Big\| \boldsymbol{I} - \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{a}_{i,\Theta^{t+1}\cup\Theta^t} \boldsymbol{a}_{i,\Theta^{t+1}\cup\Theta^t}^{\mathcal{T}} \Big\|_2 \big\| \boldsymbol{h}^t_{\Theta^t \setminus \Theta^{t+1}} \big\|_2$$

$$\leq \delta_{3k} \big\| \boldsymbol{h}^t_{\Theta^t \setminus \Theta^{t+1}} \big\|_2 \tag{5.22}$$

in which the first inequality arises again from the definition of the matrix 2-norm. Deriving the second inequality involves the approximate orthogonality result in P3, while the last result can be obtained by appealing to P4.

Consider now the last term in (5.16). For convenience, define

$$\boldsymbol{A}^{\mathcal{T}}_{\Theta^{t+1}} := [\boldsymbol{a}_{1,\Theta^{t+1}} \ \cdots \ \boldsymbol{a}_{m,\Theta^{t+1}}]$$

with $|\Theta^{t+1}| \leq 2k$, and also $\boldsymbol{v}^t := [v_1^t \ \cdots \ v_m^t]^{\mathcal{T}}$ with $v_i^t := (\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t|} - \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|}) |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$ for $i = 1, \ldots, m$. Upon rearranging terms, the induced matrix 2-norm definition implies that

$$\Big\| \frac{1}{m} \sum_{i=1}^{m} \Big(\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t|} - \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \Big) |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}| \boldsymbol{a}_{i,\Theta^{t+1}} \Big\|_2 = \frac{1}{m} \big\| \boldsymbol{A}^{\mathcal{T}}_{\Theta^{t+1}} \boldsymbol{v}^t \big\|_2$$

$$\leq \Big\| \frac{1}{\sqrt{m}} \boldsymbol{A}^{\mathcal{T}}_{\Theta^{t+1}} \Big\|_2 \Big\| \frac{1}{\sqrt{m}} \boldsymbol{v}^t \Big\|_2. \tag{5.23}$$

Property P1 confirms that the largest singular value of $\boldsymbol{A}^{\mathcal{T}}_{\Theta^{t+1}} \in \mathbb{R}^{m \times 2k}$ satisfies $s_{\max}(\boldsymbol{A}^{\mathcal{T}}_{\Theta^{t+1}}) \leq (1 + \delta_{2k})\sqrt{m}$ with high probability. Therefore, the following holds with high probability

$$\Big\| \frac{1}{m} \sum_{i=1}^{m} \Big(\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t|} - \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \Big) |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}| \boldsymbol{a}_{i,\Theta^{t+1}} \Big\|_2 \leq (1 + \delta_{2k}) \frac{1}{\sqrt{m}} \big\| \boldsymbol{v}^t \big\|_2. \tag{5.24}$$

For convenience, define the event

$$\mathcal{K}_i := \left\{ \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}|} \neq \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \right\}. \tag{5.25}$$

Then, it follows that

$$\frac{1}{m} \big\| \boldsymbol{v}^t \big\|_2^2 = \frac{1}{m} \sum_{i=1}^{m} \Big(\frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}^t|} - \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|} \Big)^2 |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|^2$$

$$\leq 4 \cdot \frac{1}{m} \sum_{i=1}^{m} |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}| \cdot |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}^t| \cdot \mathbb{1}_{\mathcal{K}_i}$$

$$\leq \frac{40}{9}\sqrt{1+\epsilon_1}\cdot\Big(\epsilon_1+\frac{1}{10}\sqrt{\frac{21}{20}}\Big)\|\boldsymbol{h}^t\|_2^2 \tag{5.26}$$

where the first inequality follows upon substituting $|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}| \leq |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}^t|$ on the event $\mathcal{K}_i$, and using $\big(\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}-\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\big)^2 \leq 4$. The last inequality can be obtained by appealing to Lemma 19 in the Appendix adapted from [112, Lemma 7.17], which holds for all $(2k)$-sparse vectors $\boldsymbol{h}\in\mathbb{R}^n$. This result has also been employed in the recent sparse phase retrieval approach reported in [62]. Here, we set $\epsilon_0 = 1/10$ in (D.3), and $\epsilon_1 > 0$ can take any sufficiently small values.

Plugging the inequality in (5.26) into (5.24) leads to

$$\begin{aligned}
&\Big\|\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}^t|}-\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\Big)|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\boldsymbol{a}_{i,\Theta^{t+1}}\Big\|_2\\
&\leq (1+\delta_{2k})\cdot\sqrt{\frac{40}{9}\sqrt{1+\epsilon_1}\cdot\Big(\epsilon_1+\frac{1}{10}\sqrt{\frac{21}{20}}\Big)}\|\boldsymbol{h}^t\|_2\\
&:= (1+\delta_{2k})\zeta\|\boldsymbol{h}^t\|_2
\end{aligned} \tag{5.27}$$

where the constant is defined as

$$\zeta := \sqrt{\frac{40}{9}}\sqrt{1+\epsilon_1}\cdot\Big(\epsilon_1+\frac{1}{10}\sqrt{\frac{21}{20}}\Big).$$

Substituting the three bounds in (5.21), (5.22), and (5.27) into (5.16), we obtain

$$\begin{aligned}
\|\boldsymbol{h}^{t+1}\|_2 &\leq 2\max\{1-\mu(1-\delta_{2k}),\,\mu(1+\delta_{2k})-1\}\|\boldsymbol{h}^t_{\Theta^{t+1}}\|_2\\
&\quad + 2\mu\delta_{3k}\|\boldsymbol{h}^t_{\Theta^t\setminus\Theta^{t+1}}\|_2 + 2\mu(1+\delta_{2k})\zeta\|\boldsymbol{h}^t\|_2\\
&\leq 2\sqrt{2}\max\big\{\max\{1-\mu(1-\delta_{2k}),\,\mu(1+\delta_{2k})-1\},\\
&\quad \mu\delta_{3k}\big\}\|\boldsymbol{h}^t\|_2 + 2\mu(1+\delta_{2k})\zeta\|\boldsymbol{h}^t\|_2\\
&\leq 2\Big[\sqrt{2}\max\big\{\max\{1-\mu(1-\delta_{2k}),\,\mu(1+\delta_{2k})-1\},\\
&\quad \mu\delta_{3k}\big\}+\mu(1+\delta_{2k})\zeta\Big]\|\boldsymbol{h}^t\|_2\\
&:= \rho\|\boldsymbol{h}^t\|_2
\end{aligned} \tag{5.28}$$

where the second inequality follows from

$$\left\|\boldsymbol{h}^t_{\Theta^{t+1}}\right\|_2 + \left\|\boldsymbol{h}^t_{\Theta^t\setminus\Theta^{t+1}}\right\|_2 \leq \sqrt{2}\left\|\boldsymbol{h}^t\right\|_2$$

over disjoint sets Θ^{t+1} and $\Theta^t \setminus \Theta^{t+1}$. To ensure linear convergence, it suffices to choose a constant step size $\mu > 0$ such that

$$\rho = 2\big[\sqrt{2}\max\left\{\max\{1-\mu(1-\delta_{2k}),\, \mu(1+\delta_{2k})-1\}, \mu\delta_{3k}\right\} + \mu(1+\delta_{2k})\zeta\big] < 1.$$

For sufficiently small $\delta_{3k} > 0$ and $\epsilon_1 > 0$, one has $\nu := 1-\rho \in (0,1)$, which justifies the linear convergence result in (5.10). □

Theorem 4 can be directly deduced by combining Lemmas 10, 11, and 12. In fact, Lemma 10 guarantees exact support recovery so that the orthogonality-promoting initialization can be effectively performed on the equivalent dimension-reduced data samples. Lemma 11, on the other hand, guarantees that the sparse initialization attained based on the dimensional-reduced data lands within a small neighborhood of the globally optimal solution with high probability. Starting from any point within the basin of attraction, Lemma 12 confirms that successive iterates of SPARTA will be dragged toward the globally optimal solution at a linear rate provided that the step size and the truncation threshold are appropriately selected.

Chapter 6

Summary and Future Directions

Leveraging recent advances in non-convex optimization and statistical signal processing, this thesis contributes to phase retrieval performance analysis and methods of high-dimensional (sparse) signals. In this final chapter, we provide a summary of the main results discussed in this thesis, and also point out a few possible directions for future research.

6.1 Thesis Summary

Building on a high-dimensional stochastic geometry property, a novel orthogonality-promoting initialization is developed in Chapter 2 for phase retrieval of unstructured signals, whose intuition and justification deviates from existing spectral alternatives. To obtain this novel initialization, the power method is invoked. The initialization is subsequently refined by means of a series of gradient-type iterations for minimizing the amplitude-based LS cost function. To further improve the exact recovery performance, a simple yet effective gradient regularization technique is put forth, which is shown to be capable of rejecting spurious search directions that may drag the iterates toward bad solutions. In addition, under random Gaussian sampling vectors, the developed amplitude flow algorithms are proved to converge exponentially fast to the global optimum as soon as the number of noiseless measurements becomes larger than some constant times the number of unknowns. This holds true with no assumption on the signal vector to be recovered. Extensive corroborating numerical tests using both computer generated data and real-world images validate the exact recovery performance analysis, and also demonstrate the merits of the novel approaches relative to the prior art.

In the context of contemporary statistical learning or inference through non-convex optimization, different data examples may contribute differently to the search direction. Building on Chapter 2, Chapter 3 advocates a time-adaptive reweighting technique to further boost the performance of the amplitude flow algorithm, which can judiciously exploit all possible useful information from all data samples. Meanwhile, a weighting technique is also combined with the orthogonality-promoting initialization of Chapter 2. Albeit conceptually simple and easy-to-implement, this idea of reweighting often leads to considerably improved exact recovery performance. Theory on exact phase recovery is also established for the reweighted variants, which is followed by extensive corroborating simulations.

In our era of data deluge, gradient-type amplitude flow algorithms in Chapters 2 and 3 may not scale well to phase retrieval of signals with high dimensionality, due in part to the matrix-vector multiplication per iteration. To endow the amplitude flow variants with scalability, stochastic optimization based iterations are pursued in Chapter 4 for tackling the amplitude-based LS formulation, which incurs per-iteration complexity solely on the order of the number of unknowns, while still maintaining optimal-order total complexity. Geometric convergence of the stochastic amplitude flow approach for exact phase recovery is demonstrated as well. Simulated tests are provided to corroborate the effectiveness and scalability of the stochastic amplitude flow variants.

In many real-world imaging applications however, the signals to be recovered are often naturally sparse or admit sparse representations after certain known and deterministic linear transformations. Leveraging this prior information, both the sample and computational complexity of phase retrieval algorithms can be considerably reduced. Toward this end, the developed amplitude flow algorithm for phase retrieval of general signals is extended to that of sparse signals in Chapter 5. We start with a novel technique for estimating the support of the true sparse signal, which is followed by the proposed initialization on the dimensionality-reduced data samples. Subsequently, a sequence of hard-thresholded iterations are implemented to refine the sparse initialization. On the theoretical side, exact recovery of the support as well as the true sparse signal is established provided that enough measurements are available. Our approach is also numerically demonstrated to exhibit superior performance relative to competitive approaches, and also to be robust to additive noise of bounded support.

6.2 Future Research

The results in this thesis open up interesting directions for a number of future research topics including phase retrieval against outliers, matrix recovery, non-coherent channel estimation, power system monitoring, and deep learning. Next, we outline a couple of them that we are currently pursuing.

6.2.1 Convolutional phase retrieval

In real-world applications, the sampling matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is often structured. For example, in diverse wireless communications applications such as channel estimation [122, 139], and non-coherent optical and underwater acoustic communication [115], the measurements available are usually generated through convolving the signal $\boldsymbol{x} \in \mathbb{C}^n$ with a given filter $\boldsymbol{a} \in \mathbb{C}^m$, namely

$$\boldsymbol{\psi} = |\boldsymbol{a} \circledast \boldsymbol{x}| \tag{6.1}$$

where $\circledast$ is the cyclic convolution modulo m. If $\boldsymbol{A} \in \mathbb{C}^{m\times m}$ denotes the circulant matrix generated by $\boldsymbol{a}$, the convolutional phase retrieval problem boils down to solving a set of phaseless linear equations

$$\boldsymbol{\psi} = |\boldsymbol{A}\boldsymbol{x}|. \tag{6.2}$$

Recently, this problem has been investigated using ordinary gradient descent with the plain-vanilla spectral initialization [97]. It is certainly of interest to address the convolutional phase retrieval task based on the amplitude flow approaches developed in this thesis together with additional gradient regularization techniques. We envision application of the novel amplitude flow approaches in wireless communications as well as in massive MIMO [96].

6.2.2 Learning convolutional neural networks

Deep convolutional neural network (CNN) architectures have recently emerged as popular and powerful tools for automatic knowledge extraction from raw data. These learning machines have led to major breakthroughs in a variety of applications including visual object classification, speech recognition, and natural language processing [74]. The critical challenge there is that training neural networks must deal with extremely high-dimensional non-convex optimization problems, and it is not clear how to guarantee optimality of the learned solutions [113].

Consider a simple rectified linear unit (ReLU)-based convolutional neural network with one input layer, one output layer, and a single hidden layer with a single filter. Formally, if the input sample, e.g., an image, is denoted by $\boldsymbol{x} \in \mathbb{R}^n$, the filter by $\boldsymbol{w} \in \mathbb{R}^d$, and the output weight vector by $\boldsymbol{v} \in \mathbb{R}^k$, the neural network input-output relationship in this case is a nonlinear function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f_{\boldsymbol{w},\boldsymbol{v}}(\boldsymbol{x}) := \boldsymbol{v}^{\mathcal{T}} \sigma(\boldsymbol{w} \circledast \boldsymbol{x}) \tag{6.3}$$

where the ReLU activation $\sigma(t) := \max(t,\, 0)$ is understood entry-wise when applied to a vector. Suppose we have access to a training dataset of m feature and label pairs $(\boldsymbol{x}_i, y_i)$. The goal is to infer the best weight vectors $\boldsymbol{w}$ and $\boldsymbol{v}$ such that the mapping $f_{\boldsymbol{w},\boldsymbol{v}}$ best fits the training data. When using the LS fitting loss, we wish to solve a non-convex optimization problem of the form [113, 42]

$$\underset{\boldsymbol{w}\in\mathbb{R}^d,\, \boldsymbol{v}\in\mathbb{R}^k}{\text{minimize}} \;\; \ell(\boldsymbol{w}, \boldsymbol{v}) := \frac{1}{2m} \sum_{i=1}^{m} \left(y_i - f_{\boldsymbol{w},\boldsymbol{v}}(\boldsymbol{x}_i)\right)^2 . \tag{6.4}$$

Evidently, the optimization (6.4) constitutes a natural generalization of the amplitude-based LS phase retrieval problem in (2.2), or more precisely, of that corresponding to the convolutional phase retrieval (6.1). Evidently, the differences are the use of the ReLU activation versus the absolute-value function as well as an additional weight vector $\boldsymbol{v}$ in (6.4). This suggests possible extensions of our developed amplitude flow algorithms and performance analysis to provable training of shallow CNNs.

6.2.3 Exact power system state recovery

The North American electric grid, the largest machine on earth, is recognized as the greatest engineering achievement of the 20th century [142]. Accurately monitoring the grid's operating condition is critical to several control and optimization tasks, including optimal power flow, reliability analysis, voltage regulation, attack detection, and future network expansion planning [50, 67, 68, 66, 132, 9, 148, 145].

Compliant with the well-known AC power flow model [50], the measurements available through the supervision control and data acquisition (SCADA) system are nonlinearly related with the power system state of interest, namely the complex voltage vector $\boldsymbol{v} := [v_1\; v_2\; \cdots\; v_n]^{\mathcal{H}}$. Suppose that a total of m measurements have been acquired for recovering $\mathbf{v}$, what is referred to as the power system state estimation (PSSE) task, and they are collected in the vector $\boldsymbol{z} \in \mathbb{R}^m$.

Mathematically, the i-th measurement in $\boldsymbol{z}$ obeys the model

$$z_i := \boldsymbol{v}^* \boldsymbol{A}_i \boldsymbol{v} + \eta_i, \quad i = 1, 2, \ldots, m \tag{6.5}$$

where the terms $\eta_i \in \mathbb{R}$ capture the modeling inaccuracies, and the measurement matrices $\boldsymbol{A}_i = \boldsymbol{A}_i^{\mathcal{H}} \in \mathbb{C}^{n \times n}$ are known and deterministic depending on the system topology and grid parameters. The critical goal of AC power system state estimation is to recover $\boldsymbol{v}$ based on the available data $\{(z_i; \boldsymbol{A}_i)\}_{i=1}^m$. The task of AC power system state estimation can be formulated as an empirical loss minimization

$$\underset{\boldsymbol{x} \in \mathbb{C}^n}{\text{minimize}} \;\; \ell(\boldsymbol{x}) := \frac{1}{2m} \sum_{i=1}^{m} (z_i - \boldsymbol{x}^* \boldsymbol{A}_i \boldsymbol{x})^2 \,. \tag{6.6}$$

Due to the quadratic terms inside the LS however, the quartic objective functional $\ell(\boldsymbol{x})$ is non-convex, whose general instance is NP-hard [94]. Hence, it is computationally intractable to compute the LS estimate of $\boldsymbol{v}$ in general.

Although several efforts have been devoted to find approximate PSSE solutions [151, 70, 133, 138, 134, 133, 125], they do not come with exact recovery guarantees even in the absence of noise. The measurement matrices $\boldsymbol{A}_i$ however, are known to be highly sparse (of very few nonzero entries), and also low rank (of at most 2). Under certain practical assumptions on the system operating condition as well as on the measurements acquired, we shall target a link between quadratic data in (6.5) and intensity data in (2.1), such that the developed amplitude flow algorithms and performance analysis are amenable to PSSE. Our goal is to devise efficient and scalable PSSE solvers from the vantage point of non-convex optimization that provide exact recovery guarantees under nominal grid operating conditions.

References

[1] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie, "Beyond pairwise clustering," in *Proc. IEEE Comput. Soc. Conf. on Comput. Vis. Pattern. Recognit.*, vol. 2, 2005, pp. 838–845.

[2] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1711–1732, Mar. 2014.

[3] M. Akcakaya and V. Tarokh, "Sparse signal recovery from a mixture of linear and magnitude-only measurements," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1220–1223, Sep. 2015.

[4] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," *Ann. Stat.*, vol. 37, no. 5B, pp. 2877–2921, Oct. 2009.

[5] A. Andoni, D. Hsu, K. Shi, and X. Sun, "Correspondence retrieval," in *Conf. on Learn. Theory*, 2017, pp. 105–126.

[6] S. Bahmani and J. Romberg, "A flexible convex relaxation for phase retrieval," *Electronic J. Stat.*, vol. 11, no. 2, pp. 5254–5281, 2017.

[7] R. Balan, "On signal reconstruction from its spectrogram," in *Annual Conf. on Inf. Sciences and Syst.*, Princeton, NJ, USA, 2010, pp. 1–4.

[8] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Appl. Comput. Harmon. Anal.*, vol. 20, no. 3, pp. 345–356, May 2006.

[9] M. Bazrafshan, N. Gatsis, A. F. Taha, and J. A. Taylor, "Coupling load-following control with OPF," *IEEE Trans. Smart Grid (*to appear*)*, 2018.

[10] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001, vol. 2.

[11] T. Bendory, Y. C. Eldar, and N. Boumal, "Non-convex phase retrieval from STFT measurements," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 467–484, Jan. 2018.

[12] V. Bentkus, "An inequality for tail probabilities of martingales with differences bounded from one side," *J. Theor. Probab.*, vol. 16, no. 1, pp. 161–173, Jan. 2003.

[13] D. Berberidis, G. Wang, G. B. Giannakis, and V. Kekatos, "Online censoring for large-scale regressions," in *IEEE Asilomar Conference on Signals, Systems and Computers*, Nov 2014, pp. 14–18.

[14] D. K. Berberidis, V. Kekatos, G. Wang, and G. B. Giannakis, "Adaptive censoring for large-scale regressions," in *IEEE Intl. Conf. Acoustics, Speech and Signal Process.*, South Brisbane, QLD, Australia, 2015, pp. 5475–5479.

[15] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, Nov. 2009.

[16] R. Burge, M. Fiddy, A. Greenaway, and G. Ross, "The phase problem," in *Proc. R. Soc. Lond. A*, vol. 350, no. 1661, Aug. 1976, pp. 191–212.

[17] T. Cai, J. Fan, and T. Jiang, "Distributions of angles in random packing on spheres," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1837–1864, Jan. 2013.

[18] T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *Ann. Stat.*, vol. 44, no. 5, pp. 2221–2251, 2016.

[19] S. Cambanis, S. Huang, and G. Simons, "On the theory of elliptically contoured distributions," *J. Multivar. Anal.*, vol. 11, no. 3, pp. 368–385, Sep. 1981.

[20] E. J. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Found. Comput. Math.*, vol. 14, no. 5, pp. 1017–1026, 2014.

[21] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," *Appl. Comput. Harmon. Anal.*, vol. 39, no. 2, pp. 277–299, Sep. 2015.

[22] ——, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[23] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Appl. Comput. Harmon. Anal.*, vol. 66, no. 8, pp. 1241–1274, Nov. 2013.

[24] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

[25] A. Chai, M. Moscoso, and G. Papanicolaou, "Array imaging using intensity-only measurements," *Inverse Probl.*, vol. 27, no. 1, p. 015005, Dec. 2011.

[26] R. Chandra, Z. Zhong, J. Hontz, V. McCulloch, C. Studer, and T. Goldstein, "PhasePack: A phase retrieval library," in *Asilomar Conf. on Signals, Syst., and Comput.*, Pacific Grove, CA, Oct. 29 - Nov. 1, 2017.

[27] H. Chang, S. Marchesini, Y. Lou, and T. Zeng, "Variational phase retrieval with globally convergent preconditioned proximal algorithm," *SIAM J. Imaging Sci.*, vol. 11, no. 1, pp. 56–93, Jan. 2018.

[28] S.-H. Chang, P. C. Cosman, and L. B. Milstein, "Chernoff-type bounds for the Gaussian error function," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 2939–2944, Jul. 2011.

[29] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *IEEE Intl. Conf. Acoustics, Speech and Signal Process.*, Las Vegas, NV, USA, 2008, pp. 3869–3872.

[30] J. Chen, G. Wang, and J. Sun, "Power scheduling for Kalman filtering over lossy wireless sensor networks," *IET Control Theory Appl.*, vol. 11, no. 4, pp. 531–540, Feb. 2017.

[31] Y. Chen and E. J. Candès, "The projected power method: An efficient algorithm for joint alignment from pairwise differences," *arXiv:1609.05820*, 2016.

[32] ——, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Comm. Pure Appl. Math.*, vol. 70, no. 5, pp. 822–883, Dec. 2017.

[33] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *arXiv:1803.07726*, 2018.

[34] Y. Chen, Y. Chi, and A. J. Goldsmith, "Exact and stable covariance estimation from quadratic sampling via convex programming," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 4034–4059, Jul. 2015.

[35] F. H. Clarke, "Generalized gradients and applications," *T. Am. Math. Soc.*, vol. 205, pp. 247–262, 1975.

[36] ——, *Optimization and Nonsmooth Analysis*. SIAM, 1990, vol. 5.

[37] A. Conca, D. Edidin, M. Hering, and C. Vinzant, "An algebraic characterization of injectivity in phase retrieval," *Appl. Comput. Harmon. Anal.*, vol. 38, no. 2, pp. 346–356, Mar. 2015.

[38] J. V. Corbett, "The Pauli problem, state reconstruction and quantum-real numbers," *Rep. Math. Phys.*, vol. 57, pp. 53–68, 2006.

[39] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, "A direct formulation for sparse PCA using semidefinite programming," *SIAM Rev.*, vol. 49, no. 3, pp. 434–448, Jul. 2007.

[40] D. Davis, D. Drusvyatskiy, and C. Paquette, "The nonsmooth landscape of phase retrieval," *arXiv:1711.03247*, 2017.

[41] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu, "Phase retrieval via linear programming: Fundamental limits and algorithmic improvements," *arXiv:1710.05234*, 2017.

[42] S. S. Du, J. D. Lee, Y. Tian, B. Poczos, and A. Singh, "Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima," *arXiv:1712.00779*, 2017.

[43] J. C. Duchi and F. Ruan, "Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval," *arXiv:1705.02356*, 2017.

[44] Y. C. Eldar and S. Mendelson, "Phase retrieval: Stability and recovery guarantees," *Appl. Comput. Harmon. Anal.*, vol. 36, no. 3, pp. 473–494, May 2014.

[45] T. S. Ferguson, "A representation of the symmetric bivariate Cauchy distribution," *Ann. Math. Stat.*, vol. 33, no. 4, pp. 1256–1266, 1962.

[46] C. Fienup and J. Dainty, "Phase retrieval and image reconstruction for astronomy," *Image Recovery: Theory and Application*, pp. 231–275, 1987.

[47] J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," *Opt. Lett.*, vol. 3, no. 1, pp. 27–29, Jul. 1978.

[48] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction," *Optik*, vol. 35, pp. 237–246, Nov. 1972.

[49] R. Ghods, A. S. Lan, T. Goldstein, and C. Studer, "PhaseLin: Linear phase retrieval," *arXiv:1802.00432*, 2018.

[50] G. B. Giannakis, V. Kekatos, N. Gatsis, S.-J. Kim, H. Zhu, and B. Wollenberg, "Monitoring and optimization for power grids: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 107–128, Sep. 2013.

[51] T. Goldstein and S. Studer, "PhaseMax: Convex phase retrieval via basis pursuit," *IEEE Trans. Inf. Theory* (to appear)*;* also *arXiv:1610.07531*, 2018.

[52] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Johns Hopkins University Press, 2012, vol. 3.

[53] P. Hand and V. Voroninski, "Compressed sensing from phaseless Gaussian measurements via linear programming in the natural parameter space," *arXiv:1611.05985*, 2016.

[54] ——, "An elementary proof of convex phase retrieval in the natural parameter space via the linear program PhaseMax," *arXiv:1611.03935*, 2016.

[55] H. A. Hauptman, "The phase problem of X-ray crystallography," *Rep. Prog. Phys.*, vol. 54, no. 11, p. 1427, Nov 1991.

[56] M. H. Hayes, "The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform," *IEEE Trans. Acoustics, Speech, and Signal Process.*, vol. 30, no. 2, pp. 140–154, Apr. 1982.

[57] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *J. ACM*, vol. 60, no. 6, p. 45, 2013.

[58] E. Hofstetter, "Construction of time-limited functions with specified autocorrelation functions," *IEEE Trans. Inf. Theory*, vol. 10, no. 2, pp. 119–126, Apr. 1964.

[59] K. Huang, Y. C. Eldar, and N. D. Sidiropoulos, "Phase retrieval from 1D Fourier measurements: Convexity, uniqueness, and algorithms," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6105–6117, Dec. 2016.

[60] M. Iwen, A. Viswanathan, and Y. Wang, "Robust sparse phase retrieval made easy," *Appl. Comput. Harmon. Anal.*, vol. 42, no. 1, pp. 135–142, Jan. 2017.

[61] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Phase retrieval: An overview of recent developments," *Opt. Compressive Sens.;* also in *arXiv:1510.07713*, 2015.

[62] G. Jagatap and C. Hedge, "Phase retrieval using structured sparsity: A sample efficient algorithmic framework," *arXiv:1705.06412*, 2017.

[63] G. Jagatap, Z. Chen, C. Hegde, and N. Vaswani, "Sub-diffraction imaging using fourier ptychography and structured sparsity," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Process.*, Calgary, Alberta, CA, April 15-20 2018.

[64] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Adv. Neural Inf. Process. Syst.*, 2013, pp. 315–323.

[65] S. Kaczmarz, "Angenherte auflsung von systemen linearer gleichungen," *Bulletin International de l'Acadmie Polonaise des Sciences et des Lettres. Classe des Sciences Mathmatiques et Naturelles. Srie A, Sciences Mathmatiques*, vol. 37, pp. 355–357, 1937.

[66] V. Kekatos, G. Wang, A. J. Conejo, and G. B. Giannakis, "Stochastic reactive power management in microgrids with renewables," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3386–3395, Nov. 2015.

[67] V. Kekatos, G. Wang, H. Zhu, and G. B. Giannakis, "PSSE redux: Convex relaxation, decentralized, robust, and dynamic approaches," *Advances in Electric Power and Energy; Power Systems Engineering,* M. El-Hawary Editor; also *arXiv:1708.03981*, 2017.

[68] V. Kekatos, L. Zhang, G. B. Giannakis, and R. Baldick, "Voltage regulation algorithms for multiphase power distribution grids," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3913–3923, Sep. 2016.

[69] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[70] S.-J. Kim, G. Wang, and G. B. Giannakis, "Online semidefinite programming for power system state estimation," in *Proc. IEEE Conf. on Acoustics, Speech and Signal Process.*, Florence, Italy, May 2014, pp. 6024–6027.

[71] R. Kolte and S. A. Ozgur, "Phase retrieval via incremental truncated Wirtinger flow," *arXiv:1606.03196*, 2016.

[72] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[73] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Stat.*, vol. 28, no. 5, pp. 1302–1338, 2000.

[74] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[75] H. Y. Lee, G. J. Parka, and H. M. Kim, "A clarification of the Cauchy distribution," *Commun. Stat. Appl. Methods*, vol. 21, no. 2, pp. 183–191, Mar. 2014.

[76] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM J. Appl. Math.*, vol. 45, no. 5, pp. 3019–3033, Sep. 2013.

[77] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," *arXiv:1802.06286*, 2018.

[78] ——, "Nonconvex matrix factorization from rank-one measurements," *arXiv:1802.06286*, 2018.

[79] S. Lu, M. Hong, and Z. Wang, "A nonconvex splitting method for symmetric nonnegative matrix factorization: Convergence analysis and optimality," *IEEE Trans. Signal Process.*, vol. 65, no. 12, pp. 3120–3135, Jun. 2017.

[80] ——, "On the sublinear convergence of randomly perturbed alternating gradient descent to second order stationary solutions," *arXiv:1802.10418*, 2018.

[81] Y. M. Lu and G. Li, "Phase transitions of spectral initialization for high-dimensional nonconvex estimation," *arXiv:1702.06435*, 2017.

[82] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution," *arXiv:1711.10467*, 2017.

[83] J. Ma, J. Xu, and A. Maleki, "Optimization-based AMP for phase retrieval: The impact of initialization and l_2 -regularization," *arXiv:1801.01170*, 2018.

[84] 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, Jul. 1999.

[85] J. Miao, I. Ishikawa, Q. Shen, and T. Earnest, "Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes," *Annu. Rev. Phys. Chem.*, vol. 59, pp. 387–410, May 2008.

[86] R. P. Millane, "Phase retrieval in crystallography and optics," *J. Opt. Soc. Am. A*, vol. 7, no. 3, pp. 394–411, 1990.

[87] M. Mondelli and A. Montanari, "Fundamental limits of weak recovery with applications to phase retrieval," *arXiv:1708.05932*, 2017.

[88] M. L. Moravec, J. K. Romberg, and R. G. Baraniuk, "Compressive phase retrieval," in *Proc. SPIE*, vol. 6701, 2007, pp. 670 120–1.

[89] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Math. Program.*, vol. 39, no. 2, pp. 117–129, 1987.

[90] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.

[91] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4814–4826, Sep. 2015.

[92] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "CPRL–An extension of compressive sensing to the phase retrieval problem," in *Adv. Neural Inf. Process. Syst.*, Stateline, NV, 2012, pp. 1367–1375.

[93] E. Oja, "Simplified neuron model as a principal component analyzer," *J. Math. Biol.*, vol. 15, no. 3, pp. 267–273, Nov. 1982.

[94] P. M. Pardalos and S. A. Vavasis, "Quadratic programming with one negative eigenvalue is NP-hard," *J. Global Optim.*, vol. 1, no. 1, pp. 15–22, 1991.

[95] E. J. R. Pauwels, A. Beck, Y. C. Eldar, and S. Sabach, "On fienup methods for sparse phase retrieval," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 982–991, Feb. 2018.

[96] T. Qiu, X. Fu, N. D. Sidiropoulos, and D. P. Palomar, "MISO channel estimation and tracking from received signal strength feedback," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1691–1704, Apr. 2018.

[97] Q. Qu, Y. Zhang, Y. C. Eldar, and J. Wright, "Convolutional phase retrieval via gradient descent," *arXiv:1712.00716*, 2017.

[98] H. Reichenbach, *Philosophic Foundations of Quantum Mechanics*. Courier Corporation, 1944.

[99] R. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Berlin-Heidelberg: Springer Verlag, 1998.

[100] Y. Saad, *Iterative Methods for Sparse Linear Systems*. SIAM, 2003.

[101] ——, *Numerical Methods for Large Eigenvalue Problems: Revised Edition*. SIAM, 2011.

[102] H. Sahinoglou and S. D. Cabrera, "On phase retrieval of finite-length sequences using the initial time sample," *IEEE Trans. Circuits and Syst.*, vol. 38, no. 8, pp. 954–958, Aug. 1991.

[103] S. Sanghavi, R. Ward, and C. D. White, "The local convexity of solving systems of quadratic equations," *Results Math.*, pp. 1–40, Jun. 2016.

[104] D. Sayre, "The squaring method: A new method for phase determination," *Acta Crystallographica*, vol. 5, no. 1, pp. 60–65, 1952.

[105] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1043–1055, Feb. 2015.

[106] O. Shamir, "Fast stochastic algorithms for SVD and PCA: Convergence properties and convexity," in *Proc. of Intl. Conf. on Mach. Learn.*, New York City, NY, 2016.

[107] Y. Shechtman, A. Beck, and Y. C. Eldar, "GESPAR: Efficient phase retrieval of sparse signals," *IEEE Trans. Signal Process.*, vol. 62, no. 4, pp. 928–938, Feb. 2014.

[108] 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 Process. Mag.*, vol. 32, no. 3, pp. 87–109, May 2015.

[109] N. Z. Shor, "A class of almost-differentiable functions and a minimization method for functions of this class," *Cybern. Syst. Anal.*, vol. 8, no. 4, pp. 599–606, Jul. 1972.

[110] N. Z. Shor, K. C. Kiwiel, and A. Ruszcayǹski, *Minimization Methods for Non-differentiable Functions*. Springer-Verlag New York, Inc., 1985.

[111] M. Soltanolkotabi, "Algorithms and theory for clustering and nonconvex quadratic programming," Ph.D. dissertation, Stanford University, 2014.

[112] ——, "Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization," *arXiv:1702.06175*, 2017.

[113] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *arXiv:1707.04926*, 2017.

[114] M. Stefik, "Inferring DNA structures from segmentation data," *Artificial Intelli.*, vol. 11, no. 1-2, pp. 85–114, Aug. 1978.

[115] M. Stojanovic, J. A. Catipovic, and J. G. Proakis, "Phase-coherent digital communications for underwater acoustic channels," *IEEE J. Ocean. Eng.*, vol. 19, no. 1, pp. 100–111, Jan. 1994.

[116] T. Strohmer and R. Vershynin, "A randomized Kaczmarz algorithm with exponential convergence," *J. Fourier Anal. Appl.*, vol. 15, no. 2, pp. 262–278, 2009.

[117] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *arXiv:1602.06664*, 2016.

[118] Y. S. Tan and R. Vershynin, "Phase retrieval via randomized Kaczmarz: Theoretical guarantees," *arXiv:1706.09993*, 2017.

[119] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv:1011.3027*, 2010.

[120] I. Waldspurger, "Phase retrieval with random Gaussian sensing vectors by alternating projections," *aXiv:1609.03088*, 2016.

[121] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Math. Program.*, vol. 149, no. 1, pp. 47–81, 2015.

[122] P. Walk, H. Becker, and P. Jung, "OFDM channel estimation via phase retrieval," in *IEEE Asilomar Conf. on Signals, Syst. and Comput.*, Pacific Grove, CA, 2015, pp. 1161–1168.

[123] G. Wang, D. Berberidis, V. Kekatos, and G. B. Giannakis, "Online reconstruction from big data via compressive censoring," in *IEEE Global Conf. Signal and Inf. Process.*, Atlanta, GA, 2014, pp. 326–330.

[124] G. Wang and G. B. Giannakis, "Solving random systems of quadratic equations via truncated generalized gradient flow," in *Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016, pp. 568–576.

[125] G. Wang, G. B. Giannakis, and J. Chen, "Robust and scalable power system state estimation via composite optimization," *IEEE Trans. Smart Grid (*submitted*);* see also *arXiv:1708.06013*, 2017.

[126] ——, "Scalable solvers of random quadratic equations via stochastic truncated amplitude flow," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1961–1974, Apr. 2017.

[127] ——, "Solving large-scale systems of random quadratic equations via stochastic truncated amplitude flow," in *25th European Signal Process. Conf.*, Kos Island, Greece, 2017, pp. 1420–1424.

[128] G. Wang, G. B. Giannakis, J. Chen, and M. Akçakaya, "SPARTA: Sparse phase retrieval via truncated amplitude flow," in *IEEE Intl. Conf. Acoustics, Speech and Signal Process.*, New Orleans, LA, USA, 2017.

[129] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 773–794, Feb. 2018.

[130] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen, "Solving most systems of random quadratic equations," in *Adv. in Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 1865–1875.

[131] ——, "Phase retrieval via reweighted amplitude flow," *IEEE Trans. Signal Process.,* 2018 (to appear)*;* also *arXiv:1705.10407*, 2018.

[132] G. Wang, V. Kekatos, A. J. Conejo, and G. B. Giannakis, "Ergodic energy management leveraging resource variability in distribution grids," *IEEE Trans. Power Syst.*, vol. 31, no. 6, pp. 4765–4775, Nov. 2016.

[133] G. Wang, A. S. Zamzam, G. B. Giannakis, and N. D. Sidiropoulos, "Power system state estimation via feasible point pursuit," in *IEEE Global Conf. Signal and Inf. Process.*, Washington, D.C., USA, 2016.

[134] ——, "Power system state estimation via feasible point pursuit: Algorithms and Cramér-Rao bound," *IEEE Trans. Signal Process.*, vol. 66, no. 6, pp. 1649–1658, Mar. 2018.

[135] G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen, "Sparse phase retrieval via truncated amplitude flow," *IEEE Trans. Signal Process.*, vol. 66, no. 2, pp. 479–491, Jan. 2018.

[136] G. Wang, L. Zhang, G. B. Giannakis, and J. Chen, "Sparse phase retrieval via iteratively reweighted amplitude flow," in *26th European Signal Process. Conf.*, Rome, Italy, 2018 (submitted).

[137] G. Wang, J. Chen, and J. Sun, "Stochastic stability of extended filtering for non-linear systems with measurement packet losses," *IET Control Theory & Appl.*, vol. 7, no. 17, pp. 2048–2055, Nov. 2013.

[138] G. Wang, S.-J. Kim, and G. Giannakis, "Moving-horizon dynamic power system state estimation using semidefinite relaxation," in *Proc. IEEE PES General Meeting*, Washington, DC, Jul. 2014, pp. 1–5.

[139] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 33, no. 4, pp. 823–831, Mar. 1985.

[140] K. Wei, "Solving systems of phaseless equations via Kaczmarz methods: A proof of concept study," *Inverse Probl.*, vol. 31, no. 12, p. 125008, 2015.

[141] Z. Wei, W. Chen, C.-W. Qiu, and X. Chen, "Conjugate gradient method for phase retrieval based on the Wirtinger derivative," *JOSA A*, vol. 34, no. 5, pp. 708–712, May 2017.

[142] W. A. Wulf, "Great achievements and grand challenges," *The Bridge*, vol. 30, no. 3/4, pp. 5–10, Fall 2010. [Online]. Available: http://www.greatachievements.org/

[143] P. Xia and G. B. Giannakis, "Design and analysis of transmit-beamforming based on limited-rate feedback," *IEEE Trans. Signal Process.*, vol. 54, no. 5, pp. 1853–1863, May 2006.

[144] L.-H. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller, "Experimental robustness of Fourier ptychography phase retrieval algorithms," *Opt. Express*, vol. 23, no. 26, pp. 33 214–33 240, Dec. 2015.

[145] A. S. Zamzam, N. D. Sidiropoulos, and E. Dall'Anese, "Beyond relaxation and Newton-Raphson: Solving AC OPF for multi-phase systems with renewables," *IEEE Trans. Smart Grid,* to appear 2018.

[146] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow," in *Intl. Conf. on Mach. Learn.*, New York City, NY, USA, 2016, pp. 1022–1031.

[147] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "Reshaped Wirtinger flow and incremental algorithm for solving quadratic system of equations," *arXiv:1605.07719*, 2016.

[148] L. Zhang, V. Kekatos, and G. B. Giannakis, "Scalable electric vehicle charging protocols," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1451–1462, Mar. 2017.

[149] L. Zhang, G. Wang, G. B. Giannakis, and J. Chen, "Compressive phase retrieval via reweighted amplitude flow," *IEEE Trans. Signal Process (*submitted*);* also *arXiv:1712.02426*, 2017.

[150] Y. Zhou, H. Zhang, and Y. Liang, "Geometrical properties and accelerated gradient solvers of non-convex phase retrieval," in *Annual Allerton Conf. on Commun., Control, and Comput.*, Monticello, Illinois, September 27-30 2016, pp. 331–335.

[151] H. Zhu and G. B. Giannakis, "Power system nonlinear state estimation using distributed semidefinite programming," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 6, pp. 1039–1050, Dec. 2014.

Appendix A

Proofs for Chapter 2

A.1 Proof of Lemma 1

By homogeneity of (2.27), it suffices to work with the case where $\|\boldsymbol{x}\| = 1$. It is easy to check that

$$\begin{aligned}\frac{1}{2}\left\|\boldsymbol{x}\boldsymbol{x}^{\mathcal{T}} - \tilde{\boldsymbol{z}}^0(\tilde{\boldsymbol{z}}^0)^{\mathcal{T}}\right\|_F^2 &= \frac{1}{2}\|\boldsymbol{x}\|^4 + \frac{1}{2}\|\tilde{\boldsymbol{z}}^0\|^4 - |\boldsymbol{x}^{\mathcal{T}}\tilde{\boldsymbol{z}}^0|^2 \\ &= 1 - |\boldsymbol{x}^{\mathcal{T}}\tilde{\boldsymbol{z}}^0|^2 \\ &= 1 - \cos^2\theta \end{aligned} \tag{A.1}$$

where $0 \leq \theta \leq \pi/2$ is the angle between the spaces spanned by $\boldsymbol{x}$ and $\tilde{\boldsymbol{z}}^0$. Then one can write

$$\boldsymbol{x} = \cos\theta\,\tilde{\boldsymbol{z}}^0 + \sin\theta(\tilde{\boldsymbol{z}}^0)^{\perp}, \tag{A.2}$$

where $(\tilde{\boldsymbol{z}}^0)^{\perp} \in \mathbb{R}^n$ is a unit vector that is orthogonal to $\tilde{\boldsymbol{z}}^0$ and has a nonnegative inner product with $\boldsymbol{x}$. Likewise,

$$\boldsymbol{x}^{\perp} := -\sin\theta\,\tilde{\boldsymbol{z}}^0 + \cos\theta(\tilde{\boldsymbol{z}}^0)^{\perp}, \tag{A.3}$$

in which $\boldsymbol{x}^{\perp} \in \mathbb{R}^n$ is a unit vector orthogonal to $\boldsymbol{x}$.

Since $\tilde{\boldsymbol{z}}^0$ is the solution to the maximum eigenvalue problem

$$\tilde{\boldsymbol{z}}^0 := \arg\max_{\|\boldsymbol{z}\|=1} \; \boldsymbol{z}^{\mathcal{T}}\bar{\boldsymbol{Y}}_0\boldsymbol{z} \tag{A.4}$$

for $\overline{\boldsymbol{Y}}_0 := \frac{1}{|\overline{\mathcal{I}}^0|}\overline{\boldsymbol{S}}_0^{\mathcal{T}}\overline{\boldsymbol{S}}_0$, it is the leading eigenvector of $\overline{\boldsymbol{Y}}_0$, i.e., $\overline{\boldsymbol{Y}}_0\tilde{\boldsymbol{z}}^0 = \lambda_1\tilde{\boldsymbol{z}}^0$, where $\lambda_1 > 0$ is the largest eigenvalue of $\overline{\boldsymbol{Y}}_0$. Premultiplying (A.2) and (A.3) by $\overline{\boldsymbol{S}}_0$ yields

$$\overline{\boldsymbol{S}}_0\boldsymbol{x} = \cos\theta\,\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0 + \sin\theta\,\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}, \tag{A.5a}$$

$$\overline{\boldsymbol{S}}_0\boldsymbol{x}^{\perp} = -\sin\theta\,\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0 + \cos\theta\,\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}. \tag{A.5b}$$

Pythagoras' relationship now gives

$$\big\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\big\|^2 = \cos^2\theta\big\|\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0\big\|^2 + \sin^2\theta\big\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\big\|^2, \tag{A.6a}$$

$$\big\|\overline{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2 = \sin^2\theta\big\|\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0\big\|^2 + \cos^2\theta\big\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\big\|^2, \tag{A.6b}$$

where the cross-terms vanish because

$$(\tilde{\boldsymbol{z}}^0)^{\mathcal{T}}\overline{\boldsymbol{S}}_0^{\mathcal{T}}\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp} = |\overline{\mathcal{I}}^0|(\tilde{\boldsymbol{z}}^0)^{\mathcal{T}}\overline{\boldsymbol{Y}}_0(\tilde{\boldsymbol{z}}^0)^{\perp} = \lambda_1|\overline{\mathcal{I}}^0|(\tilde{\boldsymbol{z}}^0)^{\mathcal{T}}(\tilde{\boldsymbol{z}}^0)^{\perp} = 0$$

following from the definition of $(\tilde{\boldsymbol{z}}^0)^{\perp}$.

We next construct the following expression:

$$\begin{aligned}
\sin^2\theta\big\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\big\|^2 - \big\|\overline{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2 &= \sin^2\theta\Big(\cos^2\theta\big\|\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0\big\|^2 + \sin^2\theta\big\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\big\|^2\Big) \\
&\quad - \Big(\sin^2\theta\big\|\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0\big\|^2 + \cos^2\theta\big\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\big\|^2\Big) \\
&= \sin^2\theta\Big(\cos^2\theta\big\|\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0\big\|^2 - \big\|\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0\big\|^2 + \sin^2\theta\big\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\big\|^2\Big) \\
&\quad - \cos^2\theta\big\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\big\|^2 \\
&= \sin^4\theta\Big(\big\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\big\|^2 - \big\|\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0\big\|^2\Big) - \cos^2\theta\big\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\big\|^2 \qquad \text{(A.7)} \\
&\le 0.
\end{aligned}$$

Regarding the last inequality, since $\tilde{\boldsymbol{z}}^0$ maximizes the term $\tilde{\boldsymbol{z}}_0^{\mathcal{T}}\overline{\boldsymbol{Y}}_0\tilde{\boldsymbol{z}}_0 = \frac{1}{|\overline{\mathcal{I}}^0|}\tilde{\boldsymbol{z}}_0^{\mathcal{T}}\overline{\boldsymbol{S}}_0^{\mathcal{T}}\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0$ according to (A.4), then in (A.7) the first term $\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\|^2 - \|\overline{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}^0\|^2 \le 0$ holds for any unit vector $(\tilde{\boldsymbol{z}}^0)^{\perp} \in \mathbb{R}^n$. In addition, the second term $-\cos^2\theta\|\overline{\boldsymbol{S}}_0(\tilde{\boldsymbol{z}}^0)^{\perp}\|^2 \le 0$, thus yielding $\sin^2\theta\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\|^2 - \|\overline{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\|^2 \le 0$. For any nonzero $\boldsymbol{x} \in \mathbb{R}^n$, it holds that

$$\sin^2\theta = 1 - \cos^2\theta \le \frac{\big\|\overline{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2}{\big\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\big\|^2}. \tag{A.8}$$

Upon letting $\boldsymbol{u} = \boldsymbol{x}^{\perp}$, the last inequality taken together with (A.1) concludes the proof of (2.28).

A.2 Proof of Lemma 2

Assume $\|\boldsymbol{x}\| = 1$. Let $\boldsymbol{s} \in \mathbb{R}^n$ be sampled uniformly at random on the unit sphere, which has zero mean and covariance matrix $\boldsymbol{I}_n/n$. Let also $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ be a unitary matrix such that $\boldsymbol{U}\boldsymbol{x} = \boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the first canonical vector in $\mathbb{R}^n$. It is then easy to verify that the following holds for any fixed threshold $0 < \tau < 1$ [43]:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{s}\boldsymbol{s}^{\mathcal{T}}|(\boldsymbol{s}^{\mathcal{T}}\boldsymbol{x})^2 > \tau] &= \boldsymbol{U}\mathbb{E}[\boldsymbol{U}^{\mathcal{T}}\boldsymbol{s}\boldsymbol{s}^{\mathcal{T}}\boldsymbol{U}|(\boldsymbol{s}^{\mathcal{T}}\boldsymbol{U}\boldsymbol{U}^{\mathcal{T}}\boldsymbol{x})^2 > \tau]\boldsymbol{U}^{\mathcal{T}} \\
&\overset{(i)}{=} \boldsymbol{U}\mathbb{E}[\tilde{\boldsymbol{s}}\tilde{\boldsymbol{s}}^{\mathcal{T}}|(\tilde{\boldsymbol{s}}^{\mathcal{T}}\boldsymbol{e}_1)^2 > \tau]\boldsymbol{U}^{\mathcal{T}} \\
&= \boldsymbol{U}\mathbb{E}[\tilde{\boldsymbol{s}}\tilde{\boldsymbol{s}}^{\mathcal{T}}|\tilde{s}_1^2 > \tau]\boldsymbol{U}^{\mathcal{T}} \\
&= \boldsymbol{U}\begin{bmatrix} \mathbb{E}[\tilde{s}_1^2|\tilde{s}_1^2 > \tau] & \mathbb{E}[\tilde{s}_1\tilde{\boldsymbol{s}}_{\setminus 1}^{\mathcal{T}}|\tilde{s}_1^2 > \tau] \\ \mathbb{E}[\tilde{s}_1\tilde{\boldsymbol{s}}_{\setminus 1}|\tilde{s}_1^2 > \tau] & \mathbb{E}[\tilde{\boldsymbol{s}}_{\setminus 1}\tilde{\boldsymbol{s}}_{\setminus 1}^{\mathcal{T}}|\tilde{s}_1^2 > \tau] \end{bmatrix}\boldsymbol{U}^{\mathcal{T}} \\
&\overset{(ii)}{=} \boldsymbol{U}\begin{bmatrix} \mathbb{E}[\tilde{s}_1^2|\tilde{s}_1^2 > \tau] & \boldsymbol{0}^{\mathcal{T}} \\ \boldsymbol{0} & \mathbb{E}[\tilde{\boldsymbol{s}}_{\setminus 1}\tilde{\boldsymbol{s}}_{\setminus 1}^{\mathcal{T}}|\tilde{s}_1^2 > \tau] \end{bmatrix}\boldsymbol{U}^{\mathcal{T}} \\
&\overset{(iii)}{=} \mathbb{E}[\tilde{s}_2^2|\tilde{s}_1^2 > \tau]\boldsymbol{I}_n + \big(\mathbb{E}[\tilde{s}_1^2|\tilde{s}_1^2 > \tau] - \mathbb{E}[\tilde{s}_2^2|\tilde{s}_1^2 > \tau]\big)\boldsymbol{x}\boldsymbol{x}^{\mathcal{T}} \\
&\overset{\triangle}{=} C_1\boldsymbol{I}_n + C_2\boldsymbol{x}\boldsymbol{x}^{\mathcal{T}}
\end{aligned} \tag{A.9}$$

with the constants $C_1 := \mathbb{E}[\tilde{s}_2^2|\tilde{s}_1^2 > \tau] < \frac{1-\tau}{n-1}$, $C_2 := \mathbb{E}[\tilde{s}_1^2|\tilde{s}_1^2 > \tau] - C_1 > 0$, and $\boldsymbol{s}_{\setminus 1} \in \mathbb{R}^{n-1}$ denoting the subvector of $\boldsymbol{s} \in \mathbb{R}^n$ after removing the first entry from $\boldsymbol{s}$. Here, the result (i) follows upon defining $\tilde{\boldsymbol{s}} := \boldsymbol{U}^{\mathcal{T}}\boldsymbol{s}$, which obeys the uniformly spherical distribution too using the rotational invariance. The equality (ii) is due to the zero-mean and symmetrical properties of the uniformly spherical distribution. Finally, to derive (iii), we have used the fact $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{e}_1 = \boldsymbol{u}_1$, the first column of $\boldsymbol{U}$, which arises from $\boldsymbol{U}^{\mathcal{T}}\boldsymbol{x} = \boldsymbol{e}_1$ and $\boldsymbol{U}\boldsymbol{U}^{\mathcal{T}} = \boldsymbol{I}_n$.

By the argument above, assume without loss of generality that $\boldsymbol{x} = \boldsymbol{e}_1$. Consider now the truncated vector $\boldsymbol{s}_{\setminus 1}|(\boldsymbol{s}^{\mathcal{T}}\boldsymbol{x})^2 > \tau$, or equivalently, $\boldsymbol{s}_{\setminus 1}|s_1^2 > \tau$. It is then clear that $\boldsymbol{s}_{\setminus 1}|s_1^2 > \tau$ is bounded, and thus subgaussian; furthermore, the next hold

$$\mathbb{E}[\boldsymbol{s}_{\setminus 1}|s_1^2 > \tau] = \boldsymbol{0} \tag{A.10a}$$

$$\mathbb{E}\big[\big(\boldsymbol{s}_{\setminus 1}|s_1^2 > \tau\big)\big(\boldsymbol{s}_{\setminus 1}|s_1^2 > \tau\big)^{\mathcal{T}}\big] = C_1\boldsymbol{I}_{n-1} \tag{A.10b}$$

where (A.10b) is obtained as a submatrix of the first term in (A.9) since the second term $C_2\boldsymbol{e}_1\boldsymbol{e}_1^{\mathcal{T}}$ is removed.

Considering a unit vector $\boldsymbol{x}^{\perp}$ such that $\boldsymbol{x}^{\mathcal{T}}\boldsymbol{x}^{\perp} = \boldsymbol{e}_1^{\mathcal{T}}\boldsymbol{x}^{\perp} = 0$, there exists a unit vector $\boldsymbol{d} \in \mathbb{R}^{n-1}$ such that $\boldsymbol{x}^{\perp} = \big[0\ \boldsymbol{d}^{\mathcal{T}}\big]^{\mathcal{T}}$. Thus, it holds that

$$\big\|\overline{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2 = \Big\|\overline{\boldsymbol{S}}_0\big[0\ \boldsymbol{d}^{\mathcal{T}}\big]^{\mathcal{T}}\Big\|^2 = \big\|\overline{\boldsymbol{S}}_{0,\setminus 1}\boldsymbol{d}\big\|^2 \tag{A.11}$$

where $\overline{\boldsymbol{S}}_{0,\setminus 1} \in \mathbb{R}^{|\overline{\mathcal{I}}^0|\times(n-1)}$ is obtained through deleting the first column in $\overline{\boldsymbol{S}}_0$, which is denoted by $\overline{\boldsymbol{S}}_{0,1}$; that is, $\overline{\boldsymbol{S}}_0 = \big[\overline{\boldsymbol{S}}_{0,1}\ \overline{\boldsymbol{S}}_{0,\setminus 1}\big]$.

The rows of $\overline{\boldsymbol{S}}_{0,\setminus 1}$ may therefore be viewed as independent realizations of the conditional random vector $\boldsymbol{s}_{\setminus 1}^{\mathcal{T}}|s_1^2 > \tau$, with the threshold τ being the $|\overline{\mathcal{I}}^0|$-largest value in $\{y_i/\|\boldsymbol{a}_i\|^2\}_{i=1}^m$. Standard concentration inequalities on the sum of random positive semi-definite matrices composed of independent non-isotropic subgaussian rows [119, Rmk. 5.40] confirm that

$$\left\|\tfrac{1}{|\overline{\mathcal{I}}^0|}\overline{\boldsymbol{S}}_{0,\setminus 1}^{\mathcal{T}}\overline{\boldsymbol{S}}_{0,\setminus 1} - C_1\boldsymbol{I}_{n-1}\right\| \le \sigma C_1 \le \frac{(1-\tau)\sigma}{n-1} \tag{A.12}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_K n}$ as long as $|\overline{\mathcal{I}}^0|/n$ is sufficiently large, where σ is a numerical constant that can take arbitrarily small values, and $c_K > 0$ is a universal constant. Without loss of generality, let us work with $\sigma := 0.005$ in (A.12). Then for any unit vector $\boldsymbol{d} \in \mathbb{R}^{n-1}$, the following inequality holds with probability at least $1 - 2\mathrm{e}^{-c_K n}$:

$$\left|\tfrac{1}{|\overline{\mathcal{I}}^0|}\boldsymbol{d}^{\mathcal{T}}\overline{\boldsymbol{S}}_{0,\setminus 1}^{\mathcal{T}}\overline{\boldsymbol{S}}_{0,\setminus 1}\boldsymbol{d} - C_1\right| \le \frac{0.01}{n} \tag{A.13}$$

for $n \ge 3$. Therefore, one readily concludes that

$$\big\|\overline{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2 = \Big|(\boldsymbol{x}^{\perp})^{\mathcal{T}}\boldsymbol{S}^{\mathcal{T}}\boldsymbol{S}\boldsymbol{x}^{\perp}\Big| \le 1.01|\overline{\mathcal{I}}^0|/n \tag{A.14}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_K n}$, provided that $|\overline{\mathcal{I}}^0|/n$ exceeds some constant. Note that c_K depends on the maximum subgaussian norm of rows of $\boldsymbol{S}$, and we assume without loss of generality $c_K \ge 1/2$. Hence, $\|\overline{\boldsymbol{S}}_0\boldsymbol{u}\|^2$ in (2.28) is upper bounded simply by letting $\boldsymbol{u} = \boldsymbol{x}^{\perp}$ in (A.14).

A.3 Proof of Lemma 3

We next pursue a meaningful lower bound for $\|\bar{\boldsymbol{S}}_0\boldsymbol{x}\|^2$ in (2.30). When $\boldsymbol{x} = \boldsymbol{e}_1$, one has $\|\bar{\boldsymbol{S}}_0\boldsymbol{x}\|^2 = \|\bar{\boldsymbol{S}}_0\boldsymbol{e}_1\|^2 = \sum_{i=1}^{|\bar{\mathcal{I}}^0|} \bar{s}_{i,1}^2$, where $\{\bar{s}_{i,1}\}_{i=1}^{|\bar{\mathcal{I}}^0|}$ are entries of the first column of $\bar{\boldsymbol{S}}_0$. It is further worth mentioning that all squared entries of any spherical random vector obey the *Beta* distribution with parameters $\alpha = \frac{1}{2}$, and $\beta = \frac{n-1}{2}$, i.e., $\bar{s}_{i,j}^2 \sim \mathrm{Beta}\big(\frac{1}{2}, \frac{n-1}{2}\big)$ for all $i,\ j$ [19, Lemma 2]. Although they have closed-form probability density functions (pdfs) that may facilitate deriving a lower bound, we take another route detailed as follows. A simple yet useful inequality is established first.

Lemma 13. *Given m fractions obeying $1 > \frac{p_1}{q_1} \geq \frac{p_2}{q_2} \geq \cdots \geq \frac{p_m}{q_m} > 0$, in which $p_i,\ q_i > 0$, $\forall i \in [m]$, the following holds for all $1 \leq k \leq m$*

$$\sum_{i=1}^{k} \frac{p_i}{q_i} \geq \sum_{i=1}^{k} \frac{p_{[i]}}{q_{[1]}} \tag{A.15}$$

where $p_{[i]}$ denotes the i-th largest one among $\{p_i\}_{i=1}^m$, and hence, $q_{[1]}$ is the maximum in $\{q_i\}_{i=1}^m$.

Proof. For any $k \in [m]$, according to the definition of $q_{[i]}$, it holds that $p_{[1]} \geq p_{[2]} \geq \cdots \geq p_{[k]}$, so $\frac{p_{[1]}}{q_{[1]}} \geq \frac{p_{[2]}}{q_{[1]}} \geq \cdots \geq \frac{p_{[k]}}{q_{[1]}}$. Considering $q_{[1]} \geq q_i$, $\forall i \in [m]$, and letting $j_i \in [m]$ be the index such that $p_{j_i} = p_{[i]}$, then $\frac{p_{j_i}}{q_{j_i}} = \frac{p_{[i]}}{q_{j_i}} \geq \frac{p_{[i]}}{q_{[1]}}$ holds for any $i \in [k]$. Therefore, $\sum_{i=1}^k \frac{p_{j_i}}{q_{j_i}} = \sum_{i=1}^k \frac{p_{[i]}}{q_{j_i}} \geq \sum_{i=1}^k \frac{p_{[i]}}{q_{[1]}}$. Note that $\left\{\frac{p_{[i]}}{q_{j_i}}\right\}_{i=1}^k$ comprise a subset of terms in $\left\{\frac{p_i}{q_i}\right\}_{i=1}^m$. On the other hand, according to our assumption, $\sum_{i=1}^k \frac{p_i}{q_i}$ is the largest among all sums of k summands; hence, $\sum_{i=1}^k \frac{p_i}{q_i} \geq \sum_{i=1}^k \frac{p_{[i]}}{q_{j_i}}$ yields $\sum_{i=1}^k \frac{p_i}{q_i} \geq \sum_{i=1}^k \frac{p_{[i]}}{q_{[1]}}$ concluding the proof. □

Without loss of generality and for simplicity of exposition, let us assume that indices of $\boldsymbol{a}_i$'s have been re-ordered such that

$$\frac{a_{1,1}^2}{\|\boldsymbol{a}_1\|^2} \geq \frac{a_{2,1}^2}{\|\boldsymbol{a}_2\|^2} \geq \cdots \geq \frac{a_{m,1}^2}{\|\boldsymbol{a}_m\|^2}, \tag{A.16}$$

where $a_{i,1}$ denotes the first element of $\boldsymbol{a}_i$. Therefore, writing $\|\bar{\boldsymbol{S}}_0\boldsymbol{e}_1\|^2 = \sum_{i=1}^{|\bar{\mathcal{I}}^0|} a_{i,1}^2/\|\boldsymbol{a}_i\|^2$, the next task amounts to finding the sum of the $|\bar{\mathcal{I}}^0|$ largest out of all m entities in (A.16). Applying

the result (A.15) in Lemma 13 gives

$$\sum_{i=1}^{|\bar{\mathcal{I}}^0|} \frac{a_{i,1}^2}{\|\boldsymbol{a}_i\|^2} \geq \sum_{i=1}^{|\bar{\mathcal{I}}^0|} \frac{a_{[i],1}^2}{\max_{i\in[m]} \|\boldsymbol{a}_i\|^2}, \tag{A.17}$$

in which $a_{[i],1}^2$ stands for the i-th largest entity in $\left\{a_{i,1}^2\right\}_{i=1}^m$.

Observe that for i.i.d. random vectors $\boldsymbol{a}_i \sim \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{I}_n\big)$, the property $\mathbb{P}(\|\boldsymbol{a}_i\|^2 \geq 2.3n) \leq \mathrm{e}^{-n/2}$ holds for large enough n (e.g., $n \geq 20$), which can be understood upon substituting $\xi := n/2$ into the following standard result [73, Lemma 1]

$$\mathbb{P}\left(\|\boldsymbol{a}_i\|^2 - n \geq 2\sqrt{\xi} + 2\xi\right) \leq \mathrm{e}^{-\xi}. \tag{A.18}$$

In addition, one readily concludes that $\mathbb{P}\left(\max_{i\in[m]} \|\boldsymbol{a}_i\| \leq \sqrt{2.3n}\right) \geq 1 - m\mathrm{e}^{-n/2}$. We will henceforth build our subsequent proofs on this event without stating this explicitly each time encountering it. Therefore, (A.17) can be lower bounded by

$$\big\|\bar{\boldsymbol{S}}\boldsymbol{x}\big\|^2 = \sum_{i=1}^{|\bar{\mathcal{I}}^0|} \frac{a_{i,1}^2}{\|\boldsymbol{a}_i\|^2} \geq \sum_{i=1}^{|\bar{\mathcal{I}}^0|} \frac{a_{[i],1}^2}{\max_{i\in[m]} \|\boldsymbol{a}_i\|^2} \geq \frac{1}{2.3n} \sum_{i=1}^{|\bar{\mathcal{I}}^0|} \big|a_{[i],1}\big|^2 \tag{A.19}$$

which holds with probability at least $1 - m\mathrm{e}^{-n/2}$. The task left for bounding $\|\bar{\boldsymbol{S}}\boldsymbol{x}\|^2$ is to derive a meaningful lower bound for $\sum_{i=1}^{|\bar{\mathcal{I}}^0|} a_{[i],1}^2$. Roughly speaking, because the ratio $|\bar{\mathcal{I}}^0|/m$ is small, e.g., $|\bar{\mathcal{I}}^0|/m \leq 1/5$, a trivial result consists of bounding $(1/|\bar{\mathcal{I}}^0|) \sum_{i=1}^{|\bar{\mathcal{I}}^0|} a_{[i],1}^2$ by its sample average $(1/m)\sum_{i=1}^m a_{[i],1}^2$. The latter can be bounded using its ensemble mean, i.e., $\mathbb{E}[a_{i,1}^2] = 1, \forall i \in [\bar{\mathcal{I}}^0]$, to yield $(1/m)\sum_{i=1}^m a_{[i],1}^2 \geq (1-\epsilon)\mathbb{E}[a_{i,1}^2] = 1 - \epsilon$, which holds with high probability for some numerical constant $\epsilon > 0$ [23, Lemma 3.1]. Therefore, one has a candidate lower bound $\sum_{i=1}^{|\bar{\mathcal{I}}^0|} a_{[i],1}^2 \geq (1-\epsilon)|\bar{\mathcal{I}}^0|$. Nonetheless, this lower bound is in general too loose, and it contributes to a relatively large upper bound on the wanted term in (2.28).

To obtain an alternative bound, let us examine first the typical size of the maximum in $\left\{a_{i,1}^2\right\}_{i=1}^m$. Observe obviously that the modulus $|a_{i,1}|$ follows the half-normal distribution having the pdf $p(r) = \sqrt{2/\pi} \cdot \mathrm{e}^{-r^2/2}$, $r > 0$, and it is easy to verify that

$$\mathbb{E}[|a_{i,1}|] = \sqrt{2/\pi}. \tag{A.20}$$

Then integrating the pdf from 0 to $+\infty$ yields the corresponding accumulative distribution function (cdf) expressible in terms of the error function $\mathbb{P}\left(|a_{i,1}| > \xi\right) = 1 - \mathrm{erf}\left(\xi/2\right)$, i.e., $\mathrm{erf}\left(\xi\right) := 2/\sqrt{\pi} \cdot \int_0^{\xi} \mathrm{e}^{-r^2} \mathrm{d}r$. Appealing to a lower bound on the complimentary error function $\mathrm{erfc}\left(\xi\right) := 1 - \mathrm{erf}\left(\xi\right)$ from [28, Thm. 2], one establishes that $\mathbb{P}\left(|a_{i,1}| > \xi\right) = 1 - \mathrm{erf}\left(\xi/2\right) \geq (3/5)\mathrm{e}^{-\xi^2/2}$. Additionally, direct application of probability theory and Taylor expansion confirms that

$$\begin{aligned} \mathbb{P}\big(\max_{i\in[m]} |a_{i,1}| \geq \xi\big) &= 1 - \left[\mathbb{P}\left(|a_{i,1}| \leq \xi\right)\right]^m \\ &\geq 1 - \Big(1 - 0.6\mathrm{e}^{-\xi^2/2}\Big)^m \\ &\geq 1 - \mathrm{e}^{-0.6m\mathrm{e}^{-\xi^2/2}}. \end{aligned} \tag{A.21}$$

Choosing now $\xi := \sqrt{2\log n}$ leads to

$$\mathbb{P}\big(\max_{i\in[m]} |a_{i,1}| \geq \sqrt{2\log n}\big) \geq 1 - \mathrm{e}^{-0.6m/n} \geq 1 - o(1) \tag{A.22}$$

which holds with the proviso that m/n is large enough, and the symbol $o(1)$ represents a small constant probability. Thus, provided that m/n exceeds some large constant, the event $\max_{i\in[m]} a_{i,1}^2 \geq 2\log n$ occurs with high probability. Hence, one may expect a tighter lower bound than $(1-\epsilon_0)|\overline{\mathcal{I}}^0|$, which is on the same order of m under the assumption that $|\overline{\mathcal{I}}^0|/m$ is about a constant.

Although $a_{i,1}^2$ obeys the Chi-square distribution with $k = 1$ degrees of freedom, its cdf is rather complicated and does not admit a nice closed-form expression. A small trick is hence taken in the sequel. Assume without loss of generality that both m and $|\overline{\mathcal{I}}^0|$ are even. Grouping two consecutive $a_{[i],1}^2$'s together, introduce a new variable $\vartheta[i] := a_{[2k-1],1}^2 + a_{[2k],1}^2, \forall k \in [m/2]$, hence yielding a sequence of ordered numbers, i.e., $\vartheta_{[1]} \geq \vartheta_{[2]} \geq \cdots \geq \vartheta_{[m/2]} > 0$. Then, one can equivalently write the wanted sum as

$$\sum_{i=1}^{|\overline{\mathcal{I}}^0|} a_{[i],1}^2 = \sum_{i=1}^{|\overline{\mathcal{I}}^0|/2} \vartheta_{[i]}. \tag{A.23}$$

On the other hand, for i.i.d. standard normal random variables $\{a_{i,1}\}_{i=1}^m$, let us consider grouping randomly two of them and denote the corresponding sum of their squares by $\chi_k :=$

$a_{k_i,1}^2 + a_{k_j,1}^2$, where $k_i \neq k_j \in [m]$, and $k \in [m/2]$. It is self-evident that the χ_k's are identically distributed obeying the Chi-square distribution with $k = 2$ degrees of freedom, having the pdf

$$p\left(r\right) = \frac{1}{2}\mathrm{e}^{-\frac{r}{2}}, \quad r \geq 0, \tag{A.24}$$

and the following complementary cdf (ccdf)

$$\mathbb{P}(\chi_k \geq \xi) := \int_{\xi}^{\infty} \frac{1}{2}\mathrm{e}^{-\frac{r}{2}}\mathrm{d}r = \mathrm{e}^{-\frac{\xi}{2}}, \quad \forall \xi \geq 0. \tag{A.25}$$

Ordering all χ_k's, summing the $|\overline{\mathcal{I}}^0|/2$ largest ones, and comparing the resultant sum with the one in (A.23) confirms that

$$\sum_{i=1}^{|\overline{\mathcal{I}}^0|/2} \chi_{[i]} \leq \sum_{i=1}^{|\overline{\mathcal{I}}^0|/2} \vartheta_{[i]} = \sum_{i=1}^{|\overline{\mathcal{I}}^0|} a_{[i],1}^2, \quad \forall |\overline{\mathcal{I}}^0| \in [m]. \tag{A.26}$$

Upon setting $\mathbb{P}(\chi_k \geq \xi) = |\overline{\mathcal{I}}^0|/m$, one obtains an estimate of $\chi_{|\overline{\mathcal{I}}^0|/2}$, the $(|\overline{\mathcal{I}}^0|/2)$-th largest value in $\{\chi_k\}_{k=1}^{m/2}$ as follows

$$\hat{\chi}_{|\overline{\mathcal{I}}^0|/2} := 2\log\left(m/|\overline{\mathcal{I}}^0|\right). \tag{A.27}$$

Furthermore, applying the Hoeffding-type inequality [119, Prop. 5.10] and leveraging the convexity of the ccdf in (A.25), one readily establishes that

$$\mathbb{P}\left(\hat{\chi}_{|\overline{\mathcal{I}}^0|/2} - \chi_{|\overline{\mathcal{I}}^0|/2} > \xi\right) \leq \mathrm{e}^{-\frac{1}{4}m\xi^2\mathrm{e}^{-\xi}(|\overline{\mathcal{I}}^0|/m)^2}, \quad \forall \xi > 0. \tag{A.28}$$

Taking without loss of generality $\xi := 0.05\hat{\chi}_{|\overline{\mathcal{I}}^0|/2} = 0.1\log\left(m/|\overline{\mathcal{I}}^0|\right)$ gives

$$\mathbb{P}\left(\chi_{|\overline{\mathcal{I}}^0|/2} < 0.95\hat{\chi}_{|\overline{\mathcal{I}}^0|/2}\right) \leq \mathrm{e}^{-c_0 m} \tag{A.29}$$

for some universal constants c_0, $c_\chi > 0$, and sufficiently large n such that $|\overline{\mathcal{I}}^0|/m \gtrsim c_\chi > 0$. The remaining part in this section assumes that this event occurs.

Choosing $\xi := 4\log n$ and substituting this into the ccdf in (A.25) leads to

$$\mathbb{P}\left(\chi \leq 4\log n\right) = 1 - 1/n^2. \tag{A.30}$$

Notice that each summand in $\sum_{i=1}^{|\bar{\mathcal{I}}^0|/2}\chi_{[i]} \geq \sum_{i=1}^{m/2}\chi_i \mathbb{1}_{\tilde{\mathcal{E}}_i}$ is Chi-square distributed, and hence could be unbounded, so we choose to work with the truncation $\sum_{i=1}^{m/2}\chi_i \mathbb{1}_{\tilde{\mathcal{E}}_i}$, where the $\mathbb{1}_{\tilde{\mathcal{E}}_i}$'s are independent copies of $\mathbb{1}_{\tilde{\mathcal{E}}}$, and $\mathbb{1}_{\tilde{\mathcal{E}}}$ denotes the indicator function for the ensuing events

$$\tilde{\mathcal{E}} := \left\{\chi \geq \hat{\chi}_{|\bar{\mathcal{I}}^0|/2}\right\} \cap \{\chi \leq 4\log n\}. \tag{A.31}$$

Apparently, it holds that $\sum_{i=1}^{|\bar{\mathcal{I}}^0|/2}\chi_{[i]} \geq \sum_{i=1}^{m/2}\chi_i \mathbb{1}_{\tilde{\mathcal{E}}_i}$. One further establishes that

$$\begin{aligned}\mathbb{E}\left[\chi_i \mathbb{1}_{\tilde{\mathcal{E}}_i}\right] &:= \int_{\hat{\chi}_{|\bar{\mathcal{I}}^0|/2}}^{4\log n} \frac{1}{2} r \mathrm{e}^{-r/2}\mathrm{d}r \\ &= \left(\hat{\chi}_{|\bar{\mathcal{I}}^0|/2}+2\right)\mathrm{e}^{-\hat{\chi}_{|\bar{\mathcal{I}}^0|/2}/2} - (4\log n+2)\,\mathrm{e}^{-2\log n} \\ &= \frac{2|\bar{\mathcal{I}}^0|}{m}\left[1+\log\left(m/|\bar{\mathcal{I}}^0\right)\right] - \frac{(4\log n+2)}{n^2}.\end{aligned} \tag{A.32}$$

The task of bounding $\sum_{i=1}^{|\bar{\mathcal{I}}^0|} a_{[i],1}^2$ in (A.26) now boils down to bounding $\sum_{i=1}^{m/2}\chi_i \mathbb{1}_{\tilde{\mathcal{E}}_i}$ from its expectation in (A.32). A convenient way to accomplish this is using the Bernstein inequality [119, Prop. 5.16], that deals with bounded random variables. That also justifies introducing the upper-bound truncation on χ in (A.31). Specifically, define

$$\vartheta_i := \chi_i \mathbb{1}_{\tilde{\mathcal{E}}_i} - \mathbb{E}\left[\chi_i \mathbb{1}_{\tilde{\mathcal{E}}_i}\right], \quad 1 \leq i \leq m/2. \tag{A.33}$$

Thus, $\{\vartheta_i\}_{i=1}^{m/2}$ are i.i.d. centered and bounded random variables following from the mean-subtraction and the upper-bound truncation. Further, according to the ccdf (A.25) and the definition of sub-exponential random variables [119, Def. 5.13], the terms $\{\vartheta_i\}_{i=1}^{m/2}$ are sub-exponential. Then, the following

$$\Big|\sum_{i=1}^{m/2}\vartheta_i\Big| \geq \tau \tag{A.34}$$

holds with probability at least $1-2\mathrm{e}^{-c_s\min\left(\tau/K_s, \tau^2/K_s^2\right)}$, in which $c_s>0$ is a universal constant, and $K_s := \max_{i\in[m/2]}\|\vartheta_i\|_{\psi_1}$ represents the maximum subexponential norm of the ϑ_i's.

Indeed, K_s can be found as follows [119, Def. 5.13]:

$$K_s := \sup_{p\geq 1} p^{-1}\left(\mathbb{E}\left[|\vartheta_i|^p\right]\right)^{1/p}$$

$$\leq \left(4\log n - 2\log\left(m/|\overline{\mathcal{I}}^0|\right)\right)\left[|\overline{\mathcal{I}}^0|/m - 1/n^2\right]$$
$$\leq \frac{2|\overline{\mathcal{I}}^0|}{m}\log\left(n^2|\overline{\mathcal{I}}^0|/m\right)$$
$$\leq \frac{4|\overline{\mathcal{I}}^0|}{m}\log n. \tag{A.35}$$

Choosing $\tau := 8|\overline{\mathcal{I}}^0|/(c_s m)\cdot\log^2 n$ in (A.34) yields

$$\sum_{i=1}^{m/2}\chi_i \mathbb{1}_{\tilde{\mathcal{E}}_i} \geq |\overline{\mathcal{I}}^0|\left[1+\log\left(m/|\overline{\mathcal{I}}^0|\right)\right] - 8|\overline{\mathcal{I}}^0|/(c_s m)\cdot\log^2 n$$
$$- m\left(2\log n + 1\right)/n^2$$
$$\geq (1-\epsilon_s)|\overline{\mathcal{I}}^0|\left[1+\log\left(m/|\overline{\mathcal{I}}^0|\right)\right] \tag{A.36}$$

for some small constant $\epsilon_s > 0$, which holds with probability at least $1 - m\mathrm{e}^{-n/2} - \mathrm{e}^{-c_0 m} - 1/n^2$ as long as m/n exceeds some numerical constant and n is sufficiently large. Therefore, combining (A.19), (A.26), and (A.36), one concludes that the following holds with high probability

$$\left\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\right\|^2 = \sum_{i=1}^{|\overline{\mathcal{I}}^0|}\frac{a_{i,1}^2}{\|\boldsymbol{a}_i\|^2} \geq (1-\epsilon_s)\frac{|\overline{\mathcal{I}}^0|}{2.3n}\left[1+\log\left(m/|\overline{\mathcal{I}}^0|\right)\right]. \tag{A.37}$$

Taking $\epsilon_s := 0.01$ without loss of generality concludes the proof of Lemma 3.

A.4 Proof of Lemma 5

Let us first prove the argument for a fixed pair $\boldsymbol{h}$ and $\boldsymbol{x}$, such that $\boldsymbol{h}$ and $\boldsymbol{z}$ are independent of $\{\boldsymbol{a}_i\}_{i=1}^m$, and then apply a covering argument. To start, introduce a Lipschitz-continuous counterpart for the discontinuous indicator function [32, A.2]

$$\chi_E(\theta) := \begin{cases} 1, & |\theta| \geq \frac{\sqrt{1.01}}{1+\gamma}, \\ 100(1+\gamma)^2\theta^2 - 100, & \frac{1}{1+\gamma} \leq |\theta| < \frac{\sqrt{1.01}}{1+\gamma}, \\ 0, & |\theta| < \frac{1}{1+\gamma} \end{cases} \tag{A.38}$$

with Lipschitz constant $\mathcal{O}(1)$. Recall $\mathcal{E}_i = \left\{ \left| \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \geq \frac{1}{1+\gamma} \right\}$, so it holds that $0 \leq \chi_E \left(\left| \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right) \leq \mathbb{1}_{\mathcal{E}_i}$ for any $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^n$, thus yielding

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right)^2 \mathbb{1}_{\mathcal{E}_i} &\geq \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right)^2 \chi_E \left(\left| \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{z}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right) \\ &= \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right)^2 \chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right). \end{aligned} \tag{A.39}$$

By homogeneity and rotational invariance of normal distributions, it suffices to prove the case where $\boldsymbol{x} = \boldsymbol{e}_1$ and $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| = \|\boldsymbol{h}\| \leq \rho$. According to (A.39), lower bounding the first term in (2.52) can be achieved by lower bounding $\sum_{i=1}^{m} (\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h})^2 \chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right)$ instead. To that end, let us find the mean of $\left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right)^2 \chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right)$. Note that $\left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right)^2$ and $\chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right)$ are dependent. Introduce an orthonormal matrix $\boldsymbol{U}_{\boldsymbol{h}}$ that contains $\boldsymbol{h}^{\mathcal{T}}/\|\boldsymbol{h}\|$ as its first row, namely,

$$\boldsymbol{U}_{\boldsymbol{h}} := \begin{bmatrix} \boldsymbol{h}^{\mathcal{T}}/\|\boldsymbol{h}\| \\ \widetilde{\boldsymbol{U}}_{\boldsymbol{h}} \end{bmatrix} \tag{A.40}$$

for some orthogonal matrix $\widetilde{\boldsymbol{U}}_{\boldsymbol{h}} \in \mathbb{R}^{(n-1)\times n}$ such that $\boldsymbol{U}_{\boldsymbol{h}}$ is orthonormal. Moreover, define $\tilde{\boldsymbol{h}} := \boldsymbol{U}_{\boldsymbol{h}} \boldsymbol{h}$, and $\tilde{\boldsymbol{a}}_i := \boldsymbol{U}_{\boldsymbol{h}} \boldsymbol{a}_i$; and let $\tilde{a}_{i,1}$ and $\tilde{\boldsymbol{a}}_{i,\setminus 1}$ denote the first entry and the remaining entries in the vector $\tilde{\boldsymbol{a}}_i$; likewise for $\tilde{\boldsymbol{h}}$. Then, for any $\boldsymbol{h}$ such that $\|\boldsymbol{h}\| \leq \rho$, we have

$$\begin{aligned} &\mathbb{E}\left[(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h})^2 \chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right) \right] \\ &= \mathbb{E}\left[(\tilde{a}_{i,1} \tilde{h}_1)^2 \chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right) \right] + \mathbb{E}\left[(\tilde{\boldsymbol{a}}_{i,\setminus 1}^{\mathcal{T}} \tilde{\boldsymbol{h}}_{\setminus 1})^2 \chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}} \right| \right) \right] \\ &= \tilde{h}_1^2 \, \mathbb{E}\left[\tilde{a}_{i,1}^2 \, \chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{a_{i,1}} \right| \right) \right] + \mathbb{E}\left[(\tilde{\boldsymbol{a}}_{i,\setminus 1}^{\mathcal{T}} \tilde{\boldsymbol{h}}_{\setminus 1})^2 \right] \mathbb{E}\left[\chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{a_{i,1}} \right| \right) \right] \\ &= \tilde{h}_1^2 \mathbb{E}\left[\tilde{a}_{i,1}^2 \chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{a_{i,1}} \right| \right) \right] + \|\tilde{\boldsymbol{h}}_{\setminus 1}\|^2 \mathbb{E}\left[\chi_E \left(\left| 1 + \frac{\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h}}{a_{i,1}} \right| \right) \right] \\ &\geq \left(\tilde{h}_1^2 + \|\tilde{\boldsymbol{h}}_{\setminus 1}\|^2 \right) \min \left\{ \mathbb{E}\left[a_{i,1}^2 \chi_E \left(\left| 1 + h_1 + \frac{\boldsymbol{a}_{i,\setminus 1}^{\mathcal{T}} \boldsymbol{h}_{\setminus 1}}{a_{i,1}} \right| \right) \right], \right. \\ &\qquad \left. \mathbb{E}\left[\chi_E \left(\left| 1 + h_1 + \frac{\boldsymbol{a}_{i,\setminus 1}^{\mathcal{T}} \boldsymbol{h}_{\setminus 1}}{a_{i,1}} \right| \right) \right] \right\} \end{aligned}$$

$$\geq \|\boldsymbol{h}\|^2 \min\left\{\mathbb{E}\left[a_{i,1}^2\chi_E\left(\left|1-\rho+\frac{a_{i,2}}{a_{i,1}}\rho\right|\right)\right], \mathbb{E}\left[\chi_E\left(1-\rho+\frac{a_{i,2}}{a_{i,1}}\rho\right)\right]\right\}$$
$$= (1-\zeta_1)\|\boldsymbol{h}\|^2 \tag{A.41}$$

where the second equality follows from the independence between $\tilde{\boldsymbol{a}}_{i,\setminus 1}^{\mathcal{T}}\tilde{\boldsymbol{h}}_{\setminus 1}$ and $\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}$, the second inequality holds for $\rho \leq 1/10$ and $\gamma \geq 1/2$, and the last equality comes from the definition of ζ_1 in (A.33). Notice that $\varrho := (\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right) \leq (\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^2 \stackrel{d}{=} \|\boldsymbol{h}\|^2 a_{i,1}^2$ is a subexponential variable, and thus its subexponential norm $\|\varrho\|_{\psi_1} := \sup_{p\geq 1}\left[\mathbb{E}(|\varrho|^p)\right]^{1/p}$ is finite.

Direct application of the Berstein-type inequality [119, Prop. 5.16] confirms that for any $\epsilon > 0$, the following

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right) \geq \mathbb{E}\left[\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)\right] - \epsilon\|\boldsymbol{h}\|^2$$
$$\geq (1-\zeta_1-\epsilon)\|\boldsymbol{h}\|^2 \tag{A.42}$$

holds with probability at least $1-\mathrm{e}^{-c_5 m\epsilon^2}$ for some numerical constant $c_5 > 0$ provided that $\epsilon \leq \|\varrho\|_{\psi_1}$ by assumption.

To obtain uniform control over all vectors $\boldsymbol{z}$ and $\boldsymbol{x}$ such that $\|\boldsymbol{z}-\boldsymbol{x}\| \leq \rho$, the net covering argument is applied [119, Def. 5.1]. Let $\mathcal{S}_\epsilon$ be an ϵ-net of the unit sphere, $\mathcal{L}_\epsilon$ be an ϵ-net of $[0,\ \rho]$, and define

$$\mathcal{N}_\epsilon := \{(\boldsymbol{z},\ \boldsymbol{h},\ t) : (\boldsymbol{z}_0,\ \boldsymbol{h}_0,\ t_0) \in \mathcal{S}_\epsilon \times \mathcal{S}_\epsilon \times \mathcal{L}_\epsilon\}. \tag{A.43}$$

Since the cardinality $|\mathcal{S}_\epsilon| \leq (1+2/\epsilon)^n$ [119, Lemma 5.2], then

$$|\mathcal{N}_\epsilon| \leq (1+2/\epsilon)^{2n}\rho/\epsilon \leq (1+2/\epsilon)^{2n+1} \tag{A.44}$$

due to the fact that $\rho/\epsilon < 2/\epsilon < 1+2/\epsilon$ for $0 < \rho < 1$.

Consider now any $(\boldsymbol{z},\ \boldsymbol{h},\ t)$ obeying $\|\boldsymbol{h}\| = t \leq \rho$. There exists a pair $(\boldsymbol{z}_0,\ \boldsymbol{h}_0,\ t_0) \in \mathcal{N}_\epsilon$ such that $\|\boldsymbol{z}-\boldsymbol{z}_0\|$, $\|\boldsymbol{h}-\boldsymbol{h}_0\|$, and $|t-t_0|$ are each at most ϵ. Taking the union bound yields

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)$$
$$\geq \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0\right)^2\chi_E\left(\left|1-t_0+\frac{a_{i,2}}{a_{i,1}}t_0\right|\right)$$

$$\geq (1-\zeta_1-\epsilon)\,\|\boldsymbol{h}_0\|^2, \quad \forall\,(\boldsymbol{z}_0,\,\boldsymbol{h}_0,\,t_0)\in\mathcal{N}_\epsilon \tag{A.45}$$

with probability at least $1-(1+2/\epsilon)^{2n+1}\,\mathrm{e}^{-c_5\epsilon^2 m}\geq 1-\mathrm{e}^{-c_0 m}$, which follows by choosing m such that $m\geq\left(c_6\cdot\epsilon^{-2}\log\epsilon^{-1}\right)n$ for some constant $c_6>0$.

Recall that $\chi_E\,(\tau)$ is Lipschitz continuous, thus

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)-\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}^{\mathcal{T}}\boldsymbol{h}_0}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)\right| \\
&\lesssim \frac{1}{m}\sum_{i=1}^{m}\left|\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2-\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0\right)^2\right| \\
&= \frac{1}{m}\sum_{i=1}^{m}\left|\boldsymbol{a}_i^{\mathcal{T}}\left(\boldsymbol{h}\boldsymbol{h}^{\mathcal{T}}-\boldsymbol{h}_0\boldsymbol{h}_0^{\mathcal{T}}\right)\boldsymbol{a}_i\right| \\
&\lesssim c_7\sum_{i=1}^{m}\left|\boldsymbol{h}\boldsymbol{h}^{\mathcal{T}}-\boldsymbol{h}_0\boldsymbol{h}_0^{\mathcal{T}}\right| \\
&\leq 2.5c_7\,\|\boldsymbol{h}-\boldsymbol{h}_0\|\,\|\boldsymbol{h}\| \\
&\leq 2.5c_7\rho\epsilon
\end{aligned} \tag{A.46}$$

for some numerical constant c_7 and provided that $\epsilon<1/2$ and $m\geq\left(c_6\cdot\epsilon^{-2}\log\epsilon^{-1}\right)n$, where the first inequality arises from the Lipschitz property of $\chi_E(\tau)$, the second uses the results in Lemma 1 in [32], and the third from Lemma 2 in [32].

Putting all results together confirms that with probability exceeding $1-2\mathrm{e}^{-c_0 m}$, we have

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)\geq\left[1-\zeta_1-(1+2.5c_7\rho)\,\epsilon\right]\|\boldsymbol{h}\|^2 \tag{A.47}$$

for all vectors $\|\boldsymbol{h}\|\,/\,\|\boldsymbol{x}\|\leq\rho$, concluding the proof.

A.5 Proof of Lemma 6

Similar to the proof in Sec. A.4, it is convenient to work with the following auxiliary function instead of the discontinuous indicator function

$$\chi_D(\theta) := \begin{cases} 1, & |\theta| \geq \frac{2+\gamma}{1+\gamma} \\ -100\left(\frac{1+\gamma}{2+\gamma}\right)^2\theta^2 + 100, & \sqrt{0.99}\cdot\frac{2+\gamma}{1+\gamma} \leq |\theta| < \frac{2+\gamma}{1+\gamma} \\ 0, & |\theta| < \sqrt{0.99}\cdot\frac{2+\gamma}{1+\gamma} \end{cases} \tag{A.48}$$

which is Lipschitz continuous in θ with Lipschitz constant $\mathcal{O}(1)$. For $\mathcal{D}_i = \left\{\left|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right| \geq \frac{2+\gamma}{1+\gamma}\right\}$, it holds that $0 \leq \mathbb{1}_{\mathcal{D}_i} \leq \chi_D\left(\left|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)$ for any $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^n$. Assume without loss of generality that $\boldsymbol{x} = \boldsymbol{e}_1$. Then for $\gamma > 0$ and $\rho \leq 1/10$, it holds that

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\geq\frac{2+\gamma}{1+\gamma}\right\}} &\leq \frac{1}{m}\sum_{i=1}^{m}\chi_D\left(\left|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}\chi_D\left(\left|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{a_{i,1}}\right|\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}\chi_D\left(\left|h_1 + \frac{\boldsymbol{a}_{i,\setminus 1}^{\mathcal{T}}\boldsymbol{h}_{\setminus 1}}{a_{i,1}}\right|\right) \\ &= \frac{1}{m}\sum_{i=1}^{m}\chi_D\left(\left|h_1 + \frac{a_{i,2}}{a_{i,1}}\left\|\boldsymbol{h}_{\setminus 1}\right\|\right|\right) \\ &\overset{\text{(i)}}{\leq} \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\left|h_1+\frac{a_{i,2}}{a_{i,1}}\|\boldsymbol{h}_{\setminus 1}\|\right|\geq\sqrt{0.99}\cdot\frac{2+\gamma}{1+\gamma}\right\}} \end{aligned} \tag{A.49}$$

where the last inequality arises from the definition of χ_D. Note that $a_{i,2}/a_{i,1}$ obeys the standard Cauchy distribution, i.e., $a_{i,2}/a_{i,1} \sim \text{Cauchy}(0,1)$ [45]. Transformation properties of Cauchy distributions assert that $h_1 + \frac{a_{i,2}}{a_{i,1}}\|\boldsymbol{h}_{\setminus 1}\| \sim \text{Cauchy}(h_1, \|\boldsymbol{h}_{\setminus 1}\|)$ [75]. Recall that the cdf of a Cauchy distributed random variable $w \sim \text{Cauchy}\,(\mu_0, \alpha)$ is given by [45]

$$F(w;\mu_0,\alpha) = \frac{1}{\pi}\arctan\left(\frac{w-\mu_0}{\alpha}\right) + \frac{1}{2}. \tag{A.50}$$

It is easy to check that when $\|\boldsymbol{h}_{\setminus 1}\| = 0$, the indicator function $\mathbb{1}_{\mathcal{D}_i} = 0$ due to $|h_1| \leq \rho < \sqrt{0.99}(2+\gamma)/(1+\gamma)$. Consider only $\|\boldsymbol{h}_{\setminus 1}\| \neq 0$ next. Define for notational brevity

$w := a_{i,2}/a_{i,1}$, $\alpha := \|\boldsymbol{h}_{\setminus 1}\|$, as well as $\mu_0 := h_1/\alpha$ and $w_0 := \sqrt{0.99}\frac{2+\gamma}{\alpha(1+\gamma)}$. Then,

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{\{|\mu_0+w|\geq w_0\}}] &= 1 - \big[F(w_0;\mu_0,1) - F(-w_0;\mu_0,1)\big] \\
&= 1 - \frac{1}{\pi}\big[\arctan(w_0-\mu_0) - \arctan(-w_0-\mu_0)\big] \\
&\overset{\text{(i)}}{=} \frac{1}{\pi}\arctan\left(\frac{2w_0}{w_0^2-\mu_0^2-1}\right) \\
&\overset{\text{(ii)}}{\leq} \frac{1}{\pi}\cdot\frac{2w_0}{w_0^2-\mu_0^2-1} \\
&\overset{\text{(iii)}}{\leq} \frac{1}{\pi}\cdot\frac{2\sqrt{0.99}\rho(2+\gamma)/(1+\gamma)}{0.99(2+\gamma)^2/(1+\gamma)^2-\rho^2} \\
&\leq 0.0646
\end{aligned} \tag{A.51}$$

for all $\gamma > 0$ and $\rho \leq 1/10$. In deriving (i), we used the property $\arctan(u) + \arctan(v) = \arctan\left(\frac{u+v}{1-uv}\right) \pmod{\pi}$ for any $uv \neq 1$. Concerning (ii), the inequality $\arctan(x) \leq x$ for $x \geq 0$ is employed. Plugging given parameter values and using $\|\boldsymbol{h}_{\setminus 1}\| \leq \|\boldsymbol{h}\| \leq \rho$ confirms (iii). Next, $\mathbb{1}_{\{|\mu_0+w|\geq w_0\}}$ is bounded; and it is known that all bounded random variables are subexponential. Thus, upon applying the Bernstein-type inequality [119, Cor. 5.17], the next holds with probability at least $1-\mathrm{e}^{-c_5 m\epsilon^2}$ for some numerical constant $c_5 > 0$ and any sufficiently small $\epsilon > 0$:

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\geq\frac{2+\gamma}{1+\gamma}\right\}} &\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\left|h_1+\frac{a_{i,2}}{a_{i,1}}\|\boldsymbol{h}_{\setminus 1}\|\right|\geq\sqrt{0.99}\frac{2+\gamma}{1+\gamma}\right\}} \\
&\leq (1+\epsilon)\mathbb{E}\left[\mathbb{1}_{\left\{\left|h_1+\frac{a_{i,2}}{a_{i,1}}\|\boldsymbol{h}_{\setminus 1}\|\right|\geq\sqrt{0.99}\frac{2+\gamma}{1+\gamma}\right\}}\right] \\
&\leq \frac{1+\epsilon}{\pi}\cdot\frac{2\sqrt{0.99}\rho(2+\gamma)/(1+\gamma)}{0.99(2+\gamma)^2/(1+\gamma)^2-\rho^2}.
\end{aligned} \tag{A.52}$$

On the other hand, it is easy to establish that the following holds true for any fixed $\boldsymbol{h} \in \mathbb{R}^n$:

$$\mathbb{E}\left[(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^4\right] = \mathbb{E}\left[a_{i,1}^4\right]\|\boldsymbol{h}\|^4 = 3\|\boldsymbol{h}\|^4 \tag{A.53}$$

which has also been used in Lemma 1 [32] and Lemma 6.1 [117]. Furthermore, recalling our working assumption $\|\boldsymbol{a}_i\| \leq \sqrt{2.3n}$ and $\|\boldsymbol{h}\| \leq \rho\|\boldsymbol{x}\|$, the random variables $(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h})^4$ are bounded, and thus they are subexponential [119]. Appealing again to the Bernstein-type

inequality for subexponential random variables [119, Prop. 5.16] and provided that $m/n > c_6 \cdot \epsilon^{-2} \log \epsilon^{-1}$ for some numerical constant $c_6 > 0$, we have

$$\frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right)^4 \leq 3(1+\epsilon) \left\| \boldsymbol{h} \right\|^4 \tag{A.54}$$

which holds with probability exceeding $1 - \mathrm{e}^{-c_5 m \epsilon^2}$ for some universal constant $c_5 > 0$ and any sufficiently small $\epsilon > 0$.

Combining results (A.52), (A.54), leveraging the Cauchy-Schwartz inequality, and considering $\mathcal{D}_i \cap \mathcal{K}_i$ only consisting of a spherical cap, the following holds for any $\rho \leq 1/10$ and $\gamma > 0$:

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right)^2 \mathbb{1}_{\mathcal{D}_i \cap \mathcal{K}_i} &\leq \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right)^4} \sqrt{\frac{1}{2} \cdot \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\left\{ \frac{\left| \boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{h} \right|}{\left| \boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x} \right|} \geq \frac{2+\gamma}{1+\gamma} \right\}}} \\ &\leq \sqrt{3(1+\epsilon) \left\| \boldsymbol{h} \right\|^4} \sqrt{\frac{1+\epsilon}{\pi} \cdot \frac{\sqrt{0.99}\rho(2+\gamma)/(1+\gamma)}{0.99(2+\gamma)^2/(1+\gamma)^2 - \rho^2}} \\ &\triangleq (\zeta_2' + \epsilon') \left\| \boldsymbol{h} \right\|^2 \end{aligned} \tag{A.55}$$

where $\zeta_2' := 0.9748\sqrt{\rho\tau/(0.99\tau^2 - \rho^2)}$ with $\tau := (2+\gamma)/(1+\gamma)$, which holds with probability at least $1 - 2\mathrm{e}^{-c_0 m}$. The latter arises upon choosing $c_0 \leq c_5 \epsilon^2$ in $1 - 2\mathrm{e}^{-c_5 m \epsilon^2}$, which can be accomplished by taking m/n sufficiently large.

Appendix B

Proofs for Chapter 3

B.1 Proof of Lemma 8

Let $\{\boldsymbol{b}_i^*\}_{i=1}^{|\mathcal{S}|}$ denote rows of $\boldsymbol{B} \in \mathbb{R}^{|\mathcal{S}|\times n}$, which are obtained by scaling rows of $\boldsymbol{A}_{\mathcal{S}} := \{\boldsymbol{a}_i^*\}_{i\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times n}$ by weights $\{w_i = \psi_i^{\gamma/2}\}_{i\in\mathcal{S}}$ [cf. (3.13)]. Since $\boldsymbol{x} = \boldsymbol{e}_1$, we have $\boldsymbol{\psi} = |\boldsymbol{A}\boldsymbol{e}_1| = |\boldsymbol{A}_1|$, while the index set $\mathcal{S}$ depends solely on the first column of $\boldsymbol{A}$, and is independent of the other columns of $\boldsymbol{A}$. Using this, partition accordingly $\boldsymbol{A}^{\mathcal{S}} := [\boldsymbol{A}_1^{\mathcal{S}}\ \boldsymbol{A}_r^{\mathcal{S}}]$, where $\boldsymbol{A}_1^{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times 1}$ denotes the first column of $\boldsymbol{A}^{\mathcal{S}}$, and $\boldsymbol{A}_r^{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times(n-1)}$ collects the remaining ones. Likewise, partition $\boldsymbol{B} = [\boldsymbol{B}_1\ \boldsymbol{B}_r]$ with $\boldsymbol{B}_1 \in \mathbb{R}^{|\mathcal{S}|\times 1}$ and $\boldsymbol{B}_r \in \mathbb{R}^{|\mathcal{S}|\times(n-1)}$. By this argument, rows of $\boldsymbol{A}^{\mathcal{S}}$ are mutually independent, and Gaussian distributed with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{I}_{n-1}$. Furthermore, the weights $\psi_i^{\gamma/2} = |\boldsymbol{a}_i^*\boldsymbol{e}_1|^{\gamma/2} = |a_{i,1}|^{\gamma/2}, \forall i \in \mathcal{S}$ are also independent of the entries in $\boldsymbol{A}^{\mathcal{S}}$. As a consequence, rows of $\boldsymbol{B}_r$ are mutually independent, and one can explicitly write its i-th row as $\boldsymbol{b}_{r,i} = |\boldsymbol{a}_{[i]}^*\boldsymbol{e}_1|^{\gamma/2}\boldsymbol{a}_{[i],\backslash 1} = |a_{[i],1}|^{\gamma/2}\boldsymbol{a}_{[i],\backslash 1}$, where $\boldsymbol{a}_{[i],\backslash 1} \in \mathbb{R}^{n-1}$ is obtained after removing the first entry of $\boldsymbol{a}_{[i]}$. It is easy to verify that $\mathbb{E}[\boldsymbol{b}_{r,i}] = \boldsymbol{0}$, and $\mathbb{E}[\boldsymbol{b}_{r,i}\boldsymbol{b}_{r,i}^*] = C_\gamma \boldsymbol{I}_{n-1}$, where the constant $C_\gamma := \sqrt{2^\gamma/\pi}\Gamma(\gamma+1/2)\|\boldsymbol{x}\|^\gamma = \sqrt{2^\gamma/\pi}\Gamma(\gamma+1/2)$, and $\Gamma(\cdot)$ is the Gamma function.

Given $\boldsymbol{x}^*\boldsymbol{x}^\perp = \boldsymbol{e}_1^*\boldsymbol{x}^\perp = 0$, one can write $\boldsymbol{x}^\perp = [0\ \boldsymbol{r}^*]^*$ with any unit vector $\boldsymbol{r} \in \mathbb{R}^{n-1}$; hence,

$$\|\boldsymbol{B}\boldsymbol{x}^\perp\|^2 = \|\boldsymbol{B}[0\ \boldsymbol{r}^*]^*\|^2 = \|\boldsymbol{B}_r\boldsymbol{r}\|^2 \tag{B.1}$$

with independent sub-Gaussian rows $\boldsymbol{b}_{r,i} = |a_{j,1}|^{\gamma/2}\boldsymbol{a}_{j,\backslash 1}$ if $0 \le \gamma \le 1$. Standard concentration

results on the sum of random positive semi-definite matrices composed of independent non-isotropic sub-Gaussian rows [119, Rmk. 5.40.1] assert that

$$\left\| \frac{1}{|\mathcal{S}|} \boldsymbol{B}_r^* \boldsymbol{B}_r - C_\gamma \boldsymbol{I}_{n-1} \right\| \leq \delta \tag{B.2}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_5 n}$ provided that $|\mathcal{S}|/n$ is larger than some positive constant. Here, $\delta > 0$ is a numerical constant that can take arbitrarily small values, and $c_5 > 0$ is a constant depending on δ. With no loss of generality, take $\delta := 0.01 C_\gamma$ in (B.2). For any unit vector $\boldsymbol{r} \in \mathbb{R}^{n-1}$, the following holds with probability at least $1 - 2\mathrm{e}^{-c_5 n}$

$$\left\| \frac{1}{|\mathcal{S}|} \boldsymbol{r}^* \boldsymbol{B}_r^* \boldsymbol{B}_r \boldsymbol{r} - C_\gamma \boldsymbol{r}^* \boldsymbol{r} \right\| \leq \delta \boldsymbol{r}^* \boldsymbol{r} = \delta \tag{B.3}$$

or

$$\|\boldsymbol{B}_r \boldsymbol{r}\|^2 = \boldsymbol{r}^* \boldsymbol{B}_r^* \boldsymbol{B}_r \boldsymbol{r} \leq 1.01 C_\gamma |\mathcal{S}|. \tag{B.4}$$

Taking (B.4) back to (B.1) confirms that

$$\|\boldsymbol{B}\boldsymbol{x}^\perp\|^2 \leq 1.01 C_\gamma |\mathcal{S}| \tag{B.5}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_5 n}$ if $|\mathcal{S}|/n$ exceeds some constant. Note that c_5 depends on the maximum sub-Gaussian norm of the rows $\boldsymbol{b}_i$ in $\boldsymbol{B}_r$, and we assume without loss of generality $c_5 \geq 1/2$. Therefore, one confirms that the numerator $\|\boldsymbol{B}\boldsymbol{u}\|^2$ in (3.14) is upper bounded after replacing $\boldsymbol{x}^\perp$ with $\boldsymbol{u}$ in (B.5).

B.2 Proof of Lemma 9

This section is devoted to obtaining a meaningful lower bound for the denominator $\|\boldsymbol{B}\boldsymbol{x}\|^2$ in (3.17). Note first that

$$\|\boldsymbol{B}\boldsymbol{x}\|^2 = \sum_{i=1}^{|\mathcal{S}|} \|\boldsymbol{b}_i^* \boldsymbol{x}\|^2 = \sum_{i=1}^{|\mathcal{S}|} \psi_{[i]}^\gamma |\boldsymbol{a}_{[i]}^* \boldsymbol{x}|^2 = \sum_{i=1}^{|\mathcal{S}|} |\boldsymbol{a}_{[i]}^* \boldsymbol{x}|^{2+\gamma}.$$

Taking without loss of generality $\boldsymbol{x} = \boldsymbol{e}_1$, the term on the right side of the last equality reduces to

$$\|\boldsymbol{B}\boldsymbol{x}\|^2 = \sum_{i=1}^{|\mathcal{S}|} |a_{[i],1}|^{2+\gamma}. \tag{B.6}$$

Since $a_{[i],1}$ follows the standardized normal distribution, the probability density function (pdf) of random variables $|a_{[i],1}|^{2+\gamma}$ can be given in closed form as

$$p(t) = \sqrt{\frac{2}{\pi}} \cdot \frac{1}{2+\gamma} t^{-\frac{1+\gamma}{2+\gamma}} \mathrm{e}^{-\frac{1}{2}t^{\frac{2}{2+\gamma}}}, \quad t > 0 \tag{B.7}$$

which is rather complicated and whose cumulative density function (cdf) does not come in closed form in general. Therefore, instead of dealing with the pdf in (B.7) directly, we shall take a different route by deriving a lower bound that is a bit looser yet suffices for our purpose.

Since $|a_{[|\mathcal{S}|],1}| \leq \cdots \leq |a_{[2],1}| \leq |a_{[1],1}|$, then it holds for all $1 \leq i \leq |\mathcal{S}|$ that $|a_{[i],1}|^{2+\gamma} \geq |a_{[|\mathcal{S}|],1}|^{\gamma} a_{[i],1}^2$, which yields

$$\|\boldsymbol{B}\boldsymbol{x}\|^2 = \sum_{i=1}^{|\mathcal{S}|} |a_{[i],1}|^{2+\gamma} \geq |a_{[|\mathcal{S}|],1}|^{\gamma} \sum_{i=1}^{|\mathcal{S}|} a_{[i],1}^2. \tag{B.8}$$

We will next demonstrate next that deriving a lower bound for $\|\boldsymbol{B}\boldsymbol{x}\|^2$ suffices to derive a lower bound for the summation on the right hand side (B.8). The latter can be achieved by appealing to a result in [129, Lemma 3], which for completeness is included in the following.

Lemma 14. *For an arbitrary unit-norm vector $\boldsymbol{x} \in \mathbb{R}^n$, let $\psi_i = |\boldsymbol{a}_i^* \boldsymbol{x}|$, $1 \leq i \leq m$ be m noiseless measurements. Then with probability at least $1 - \mathrm{e}^{-c_2 m}$, the following holds*

$$\sum_{i=1}^{|\mathcal{S}|} a_{[i],1}^2 \geq 0.99|\mathcal{S}|\big[1 + \log(m/|\mathcal{S}|)\big] \tag{B.9}$$

provided that $m \geq c_0|\mathcal{S}| \geq c_1 n$ for some numerical constants c_0, c_1, $c_2 > 0$.

Combining the results in Lemma 14 and (B.8), one further deduces that

$$\|\boldsymbol{B}\boldsymbol{x}\|^2 \geq |a_{[|\mathcal{S}|],1}|^{\gamma} \sum_{i=1}^{|\mathcal{S}|} a_{[i],1}^2 \geq |a_{[|\mathcal{S}|],1}|^{\gamma} \cdot 0.99|\mathcal{S}|\big[1 + \log(m/|\mathcal{S}|)\big]. \tag{B.10}$$

The task remains to estimate the size of $|a_{[|\mathcal{S}|],1}|$, which we recall is the $|\mathcal{S}|$-th largest among the m independent realizations $\{\psi_i = |a_{i,1}|\}_{i=1}^m$. Taking $\gamma = -1$ in (B.7) gives the pdf of the half-normal distribution

$$p(t) = \sqrt{\frac{2}{\pi}} \mathrm{e}^{-\frac{1}{2}t^2}, \quad t > 0 \tag{B.11}$$

whose corresponding cdf is

$$F(\tau) = \mathrm{erf}(\tau/\sqrt{2}). \tag{B.12}$$

Setting $F(\tau_{|\mathcal{S}|}) := 1 - |\mathcal{S}|/m$ or using the complementary cdf $|\mathcal{S}|/m := \mathrm{erfc}(\tau/\sqrt{2})$ based on the complementary error function gives rise to an estimate of the size of the $|\mathcal{S}|$-th largest (or equivalently, the $(m - |\mathcal{S}|)$-th smallest) entry in the m realizations, namely

$$\tau_{|\mathcal{S}|} = \sqrt{2}\,\mathrm{erfc}^{-1}(|\mathcal{S}|/m) \tag{B.13}$$

where $\mathrm{erfc}^{-1}(\cdot)$ represents the inverse complementary error function. In the sequel, we show that the deviation of the $|\mathcal{S}|$-th largest realization $\psi_{|\mathcal{S}|}$ from its expected value $\tau_{|\mathcal{S}|}$ in (B.13) is bounded with high probability.

For random variable $\psi = |a|$ with a obeying the standard Gaussian distribution, consider the event $\psi \leq \tau_{|\mathcal{S}|} - \delta$ for a fixed constant $\delta > 0$. Define the indicator random variable $\chi := \mathbb{1}_{\{\psi \leq \tau_{|\mathcal{S}|} - \delta\}}$, whose expectation can be obtained by substituting $\tau = \tau_{|\mathcal{S}|} - \delta$ into the pdf in (B.12) as

$$\mathbb{E}[\chi_i] = \mathrm{erf}(\tau_{|\mathcal{S}|} - \delta/\sqrt{2}). \tag{B.14}$$

Considering now the m independent copies $\{\chi_i = \mathbb{1}_{\{\psi_i \leq \tau_{|\mathcal{S}|} - \delta\}}\}_{i=1}^m$ of χ, the following holds

$$\begin{aligned}
\mathbb{P}(\psi_{|\mathcal{S}|} \leq \tau_{|\mathcal{S}|} - \delta) &= \mathbb{P}\Big(\sum_{i=1}^m \chi_i \leq m - |\mathcal{S}|\Big) \\
&= \mathbb{P}\Big(\frac{1}{m}\sum_{i=1}^m (\chi_i - \mathbb{E}[\chi_i]) \leq 1 - \frac{|\mathcal{S}|}{m} - \mathbb{E}[\chi_i]\Big)
\end{aligned}$$

Clearly, since random variables χ_i are bounded, they are subgaussian [119]. For notational brevity, let $t := 1 - |\mathcal{S}|/m - \mathbb{E}[\chi_i] = 1 - |\mathcal{S}|/m - \mathrm{erf}(\tau_{|\mathcal{S}|} - \delta/\sqrt{2})$. Appealing to a large deviation

inequality for sums of independent sub-Gaussian random variables, one establishes that

$$\mathbb{P}(\psi_{|\mathcal{S}|} \leq \tau_{|\mathcal{S}|} - \delta) = \mathbb{P}\Big(\frac{1}{m}\sum_{i=1}^{m}(\chi_i - \mathbb{E}[\chi_i]) \leq 1 - \frac{|\mathcal{S}|}{m} - \mathbb{E}[\chi_i]\Big) \leq \mathrm{e}^{-c_5 m t^2} \tag{B.15}$$

where $c_5 > 0$ is some absolute constant. On the other hand, using the definition of the error function and properties of integration gives rise to

$$t = 1 - |\mathcal{S}|/m - \mathrm{erf}(\tau_{|\mathcal{S}|} - \delta/\sqrt{2}) = \frac{2}{\sqrt{\pi}}\int_{(\tau_{|\mathcal{S}|}-\delta)/\sqrt{2}}^{\tau_{|\mathcal{S}|}/\sqrt{2}} \mathrm{e}^{-s^2}\mathrm{d}s \geq \sqrt{\frac{2}{\pi}}\delta\mathrm{e}^{-\frac{\tau_{|\mathcal{S}|}^2}{2}} \geq \sqrt{\frac{2}{\pi}}\delta. \tag{B.16}$$

Taking the results in (B.15) and (B.16) together, one concludes that fixing any constant $\delta > 0$, the following holds with probability at least $1 - \mathrm{e}^{-c_2 m}$:

$$\psi_{|\mathcal{S}|} \geq \tau_{|\mathcal{S}|} - \delta \geq \sqrt{2}\,\mathrm{erfc}^{-1}(|\mathcal{S}|/m) - \delta$$

where $c_2 := 2/\pi \cdot c_5\delta^2$. Furthermore, choosing without loss of generality $\delta := 0.01\tau_{|\mathcal{S}|}$ above leads to $\psi_{|\mathcal{S}|} \geq 1.4\,\mathrm{erfc}^{-1}(|\mathcal{S}|/m)$.

Substituting the last inequality into (B.10), and under our working assumption $|\mathcal{S}|/m \leq 0.25$, one readily obtains that

$$\begin{aligned}\|\boldsymbol{B}\boldsymbol{x}\|^2 &\geq [1.4\,\mathrm{erfc}^{-1}(|\mathcal{S}|/m)]^{\gamma} \cdot 0.99|\mathcal{S}|\big[1 + \log(m/|\mathcal{S}|)\big] \\ &\geq 0.99 \cdot 1.14^{\gamma}|\mathcal{S}|\big[1 + \log(m/|\mathcal{S}|)\big]\end{aligned}$$

which holds with probability exceeding $1 - \mathrm{e}^{-c_2 m}$ for some constant $c_2 > 0$, thus concluding the proof of Lemma 9.

B.3 Proof of Proposition 9

To proceed, let us introduce the following events for all $i = 1, 2, \ldots, m$:

$$\mathcal{D}_i := \big\{(\boldsymbol{a}_i^*\boldsymbol{x})(\boldsymbol{a}_i^*\boldsymbol{z}) < 0\big\} \tag{B.17}$$

$$\mathcal{E}_i := \left\{\frac{|\boldsymbol{a}_i^*\boldsymbol{z}|}{|\boldsymbol{a}_i^*\boldsymbol{x}|} \geq \frac{1}{1+\eta}\right\} \tag{B.18}$$

for some fixed constant $\eta > 0$, in which the former corresponds to the gradients involving wrongly estimated signs, namely $\frac{\boldsymbol{a}_i^*\boldsymbol{z}}{|\boldsymbol{a}_i^*\boldsymbol{z}|} \neq \frac{\boldsymbol{a}_i^*\boldsymbol{x}}{|\boldsymbol{a}_i^*\boldsymbol{x}|}$, and the second will be useful for deriving error bounds. Based on the definition of $\mathcal{D}_i$ and with $\mathbb{1}_{\mathcal{D}_i}$ denoting the indicator function of the event $\mathcal{D}_i$, we have

$$\begin{aligned}
\langle \ell_{\mathrm{rw}}(\boldsymbol{z}), \boldsymbol{h}\rangle &= \frac{1}{m}\sum_{i=1}^{m} w_i \left(\boldsymbol{a}_i^*\boldsymbol{z} - |\boldsymbol{a}_i^*\boldsymbol{x}|\frac{\boldsymbol{a}_i^*\boldsymbol{z}}{|\boldsymbol{a}_i^*\boldsymbol{z}|}\right)(\boldsymbol{a}_i^*\boldsymbol{h}) \\
&= \frac{1}{m}\sum_{i=1}^{m} w_i \left(\boldsymbol{a}_i^*\boldsymbol{h} + \boldsymbol{a}_i^*\boldsymbol{x} - |\boldsymbol{a}_i^*\boldsymbol{x}|\frac{\boldsymbol{a}_i^*\boldsymbol{z}}{|\boldsymbol{a}_i^*\boldsymbol{z}|}\right)(\boldsymbol{a}_i^*\boldsymbol{h}) \\
&= \frac{1}{m}\sum_{i=1}^{m} w_i (\boldsymbol{a}_i^*\boldsymbol{h})^2 + \frac{1}{m}\sum_{i=1}^{m} 2w_i \big(\boldsymbol{a}_i^*\boldsymbol{x}\big)(\boldsymbol{a}_i^*\boldsymbol{h})\mathbb{1}_{\mathcal{D}_i} \\
&\geq \frac{1}{m}\sum_{i=1}^{m} w_i (\boldsymbol{a}_i^*\boldsymbol{h})^2 - \frac{1}{m}\sum_{i=1}^{m} 2w_i \big|\boldsymbol{a}_i^*\boldsymbol{x}\big|\big|\boldsymbol{a}_i^*\boldsymbol{h}\big|\mathbb{1}_{\mathcal{D}_i}. \qquad \text{(B.19)}
\end{aligned}$$

In the following, we will derive a lower bound for the term on the right hand side of (B.19). Specifically, a lower bound for the first term $(1/m)\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2$ and an upper bound for the second term $(1/m)\sum_{i=1}^{m} 2w_i|\boldsymbol{a}_i^*\boldsymbol{x}|\big|\boldsymbol{a}_i^*\boldsymbol{h}|\mathbb{1}_{\mathcal{D}_i}$ will be obtained, based on Lemmas 15 and 16, with their proofs postponed to Appendix B.4 and Appendix B.5, respectively.

Lemma 15. *Fix fixed $\eta,\ \beta > 0$, and any sufficiently small constant $\epsilon > 0$, the following holds with probability at least $1 - 2\mathrm{e}^{-c_5\epsilon^2 m}$*

$$\frac{1}{m}\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2 \geq \frac{1-\zeta_1-\epsilon}{1+\beta(1+\eta)}\|\boldsymbol{h}\|^2 \qquad \text{(B.20)}$$

with $w_i = 1/[1+\beta/(|\boldsymbol{a}_i^*\boldsymbol{z}|/|\boldsymbol{a}_i^*\boldsymbol{x}|)]$ for all $1 \leq i \leq m$, provided that $m/n > (c_6 \cdot \epsilon^{-2}\log\epsilon^{-1})$ for certain numerical constants $c_5,\ c_6 > 0$.

Now we turn to the second term in (B.19). For ease of exposition, let us first introduce the following events

$$\mathcal{B}_i := \big\{|\boldsymbol{a}_i^*\boldsymbol{x}| < |\boldsymbol{a}_i^*\boldsymbol{h}| \leq (k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|\big\} \qquad \text{(B.21)}$$

$$\mathcal{O}_i := \big\{(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}| < |\boldsymbol{a}_i^*\boldsymbol{h}|\big\} \qquad \text{(B.22)}$$

for all $1 \le i \le m$ and some fixed constant $k > 0$. The second term can be bounded as follows

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m} 2w_i\big|\boldsymbol{a}_i^*\boldsymbol{x}\big|\big|\boldsymbol{a}_i^*\boldsymbol{h}\big|\mathbb{1}_{\mathcal{D}_i} &\le \frac{1}{m}\sum_{i=1}^{m} w_i\big[(\boldsymbol{a}_i^*\boldsymbol{x})^2 + (\boldsymbol{a}_i^*\boldsymbol{h})^2\big]\,\mathbb{1}_{\{(\boldsymbol{a}_i^*\boldsymbol{z})(\boldsymbol{a}_i^*\boldsymbol{x})<0\}} \\
&= \frac{1}{m}\sum_{i=1}^{m} w_i\big[(\boldsymbol{a}_i^*\boldsymbol{x})^2 + (\boldsymbol{a}_i^*\boldsymbol{h})^2\big]\,\mathbb{1}_{\{(\boldsymbol{a}_i^*\boldsymbol{h})(\boldsymbol{a}_i^*\boldsymbol{x})+(\boldsymbol{a}_i^*\boldsymbol{x})^2<0\}} \\
&\le \frac{1}{m}\sum_{i=1}^{m} w_i\big[(\boldsymbol{a}_i^*\boldsymbol{x})^2 + (\boldsymbol{a}_i^*\boldsymbol{h})^2\big]\,\mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}} \\
&\le \frac{2}{m}\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2\mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}} \\
&= \frac{2}{m}\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2\mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\le(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|\}} \\
&\quad + \frac{2}{m}\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2\mathbb{1}_{\{(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}} \\
&= \frac{2}{m}\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2\mathbb{1}_{\mathcal{B}_i} + \frac{2}{m}\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2\mathbb{1}_{\mathcal{O}_i} \qquad \text{(B.23)}
\end{aligned}$$

where the first equality is derived by substituting $\boldsymbol{z} = \boldsymbol{h} + \boldsymbol{x}$ according to the definition of $\boldsymbol{h}$, the second event suffices for $(\boldsymbol{a}_i^*\boldsymbol{h})(\boldsymbol{a}_i^*\boldsymbol{x}) + (\boldsymbol{a}_i^*\boldsymbol{x})^2 < 0$, and the second equality follows from writing the indicator function $\mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}}$ as the summation of two indicator functions of two events $\mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\le(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|\}}$ and $\mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{h}|>(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|\}}$.

The task so far remains to derive upper bounds for the two terms on the right hand side of (B.23), which leads to Lemma 16.

Lemma 16. *Fixing a fixed $k > 0$, define ζ_2 to be the maximum of $\mathbb{E}[w_i]$ in* (B.32) *for $\varrho = 0.01$ and $\nu = 0.1$, which depends only on k. For any $\epsilon > 0$, if $m/n > c_6\epsilon^{-2}\log\epsilon^{-1}$, the following hold simultaneously with probability at least $1 - c_3\mathrm{e}^{-c_2\epsilon^2 m}$*

$$\frac{1}{m}\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2\mathbb{1}_{\mathcal{O}_i} \le (\zeta_2 + \epsilon)\|\boldsymbol{h}\|^2 \qquad \text{(B.24)}$$

and

$$\frac{1}{m}\sum_{i=1}^{m} w_i(\boldsymbol{a}_i^*\boldsymbol{h})^2\mathbb{1}_{\mathcal{B}_i} \le \frac{0.1271 - \zeta_2 + \epsilon}{1 + \beta/k}\|\boldsymbol{h}\|^2 \qquad \text{(B.25)}$$

for all $\boldsymbol{h} \in \mathbb{R}^n$ *obeying* $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| \leq 1/10$*, where* $c_1,\, c_2,\, c_3 > 0$ *are some universal constants.*

Substituting (B.20), (B.23), and (B.24)-(B.25) established in Lemmas 15 and 16 back into (B.19), we conclude that

$$\begin{aligned}\langle \ell_{\mathrm{rw}}(\boldsymbol{z}), \boldsymbol{h}\rangle &\geq \frac{1}{m}\sum_{i=1}^{m} w_i (\boldsymbol{a}_i^*\boldsymbol{h})^2 \mathbb{1}_{\mathcal{E}_i} - \frac{1}{m}\sum_{i=1}^{m} 2w_i \big|\boldsymbol{a}_i^*\boldsymbol{x}\big|\big|\boldsymbol{a}_i^*\boldsymbol{h}\big| \mathbb{1}_{\mathcal{D}_i} \\ &= \zeta_e \|\boldsymbol{h}\|^2 \end{aligned} \tag{B.26}$$

which will be rendered positive, provided that $\beta > 0$ is small enough, and that parameters $\eta,\, k > 0$ are suitably chosen.

B.4 Proof of Lemma 15

Plugging in the weighting parameters $w_i = \frac{1}{1+\beta/(|\boldsymbol{a}_i^*\boldsymbol{z}|/|\boldsymbol{a}_i^*\boldsymbol{x}|)}$ and based on the definition of $\mathcal{E}_i$, the first term in (B.19) can be lower bounded as

$$\frac{1}{m}\sum_{i=1}^{m} w_i (\boldsymbol{a}_i^*\boldsymbol{h})^2 \geq \frac{1}{m}\sum_{i=1}^{m} \frac{(\boldsymbol{a}_i^*\boldsymbol{h})^2 \mathbb{1}_{\mathcal{E}_i}}{1+\beta/(|\boldsymbol{a}_i^*\boldsymbol{z}|/|\boldsymbol{a}_i^*\boldsymbol{x}|)} \tag{B.27}$$

$$\begin{aligned} &\geq \frac{1}{m}\sum_{i=1}^{m} \frac{1}{1+\beta(1+\eta)} (\boldsymbol{a}_i^*\boldsymbol{h})^2 \mathbb{1}_{\left\{\frac{|\boldsymbol{a}_i^*\boldsymbol{z}|}{|\boldsymbol{a}_i^*\boldsymbol{x}|} \geq \frac{1}{1+\eta}\right\}} \\ &= \frac{1}{1+\beta(1+\eta)} \cdot \frac{1}{m}\sum_{i=1}^{m} (\boldsymbol{a}_i^*\boldsymbol{h})^2 \mathbb{1}_{\mathcal{E}_i} \end{aligned} \tag{B.28}$$

where the first inequality arises from dropping some nonnegative terms from the left hand side, and the second one after replacing the ratio $|\boldsymbol{a}_i^*\boldsymbol{z}|/|\boldsymbol{a}_i^*\boldsymbol{x}|$ in the weights by its lower bound $1/(1+\eta)$ because the weights are monotonically increasing functions of $|\boldsymbol{a}_i^*\boldsymbol{z}|/|\boldsymbol{a}_i^*\boldsymbol{x}|$. Using [129, Lemma 5], the last term in (B.28) can be further bounded by

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} w_i (\boldsymbol{a}_i^*\boldsymbol{h})^2 &\geq \frac{1}{1+\beta(1+\eta)} \cdot \frac{1}{m}\sum_{i=1}^{m} (\boldsymbol{a}_i^*\boldsymbol{h})^2 \mathbb{1}_{\mathcal{E}_i} \\ &\geq \frac{1-\zeta_1-\epsilon}{1+\beta(1+\eta)} \|\boldsymbol{h}\|^2 \end{aligned} \tag{B.29}$$

for any fixed sufficiently small constant $\epsilon > 0$, which holds with probability at least $1 - 2\mathrm{e}^{-c_5\epsilon^2 m}$, if $m > (c_6 \cdot \epsilon^{-2} \log \epsilon^{-1})n$.

B.5 Proof of Lemma 16

The proof is adapted from [147, Lemma 9]. We first prove the bound (B.24) for any fixed $\boldsymbol{h}$ obeying $\|\boldsymbol{h}\| \leq \|\boldsymbol{x}\|/10$, and subsequently develop a uniform bound at the end of this section. The bound (B.25) can be derived directly after subtracting the bound in (B.24) with k from that bound with $k = 0$, followed by an application of the Bernstein-type sub-exponential tail bound [119]. We only discuss the first bound (B.24). Because of the discontinuity hence non-Lipschitz of the indicator functions, let us approximate them by a sequence of auxiliary Lipschitz functions. Specifically, with some constant $\varrho > 0$, define for all $1 \leq i \leq m$ the ensuing continuous functions

$$\chi_i(s) := \begin{cases} s, & s > (1+k)^2(\boldsymbol{a}_i^*\boldsymbol{x})^2 \\ \frac{1}{\varrho}[s-(k+1)^2(\boldsymbol{a}_i^*\boldsymbol{x})^2] & \\ \quad +(k+1)^2(\boldsymbol{a}_i^*\boldsymbol{x})^2, & (1-\varrho)(k+1)^2(\boldsymbol{a}_i^*\boldsymbol{x})^2 \leq s \leq (k+1)^2(\boldsymbol{a}_i^*\boldsymbol{x})^2 \\ 0, & \text{otherwise.} \end{cases} \tag{B.30}$$

Clearly, all χ_i(s)'s are random Lipschitz functions with constant $1/\varrho$. Furthermore, it is easy to verify that

$$|\boldsymbol{a}_i^*\boldsymbol{h}|^2 \mathbb{1}_{\{(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}} \leq \chi_i(|\boldsymbol{a}_i^*\boldsymbol{h}|^2) \leq |\boldsymbol{a}_i^*\boldsymbol{h}|^2 \mathbb{1}_{\{\sqrt{1-\varrho}(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}}. \tag{B.31}$$

Since the second term involves the addition event $\mathcal{G}_i$ in (B.18), define

$$w_i := \frac{|\boldsymbol{a}_i^*\boldsymbol{h}|^2}{\|\boldsymbol{h}\|^2} \mathbb{1}_{\{\sqrt{1-\varrho}(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}}$$

for $1 \leq i \leq m$, and $\nu := \frac{\|\boldsymbol{h}\|}{\|\boldsymbol{x}\|}$ for convenience. If $f(\tau_1, \tau_2)$ denotes the density of two joint Gaussian random variables with correlation coefficient $\rho = \frac{\boldsymbol{h}^*\boldsymbol{x}}{\|\boldsymbol{h}\|\|\boldsymbol{x}\|} \in (-1, 1)$, then the expectation of w_i can be obtained using the conditional expectation

$$\mathbb{E}[w_i] = \int_{-\infty}^{\infty} \mathbb{E}[w_i | \boldsymbol{a}_i^*\boldsymbol{x} = \tau_1\|\boldsymbol{x}\|, \boldsymbol{a}_i^*\boldsymbol{h} = \tau_1\|\boldsymbol{h}\|] f(\tau_1,\tau_2) \mathrm{d}\tau_1 \mathrm{d}\tau_2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tau_2^2 \mathbb{1}_{\{\sqrt{1-\varrho}(k+1)|\tau_1|<|\tau_2|\nu\}} f(\tau_1,\tau_2)\mathrm{d}\tau_1\mathrm{d}\tau_2$$

$$= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} \tau_2^2 \exp(-\tau_2^2/2)\bigg[\mathrm{erf}\Big(\frac{(\nu/[\sqrt{1-\varrho}(k+1)]-\rho)\tau_2}{\sqrt{2(1-\rho^2)}}\Big) + \mathrm{erf}\Big(\frac{(\nu/[\sqrt{1-\varrho}(k+1)]+\rho)\tau_2}{\sqrt{2(1-\rho^2)}}\Big)\bigg]\mathrm{d}\tau_2 \tag{B.32}$$

$$:= \zeta_2. \tag{B.33}$$

It is not difficult to see that $\mathbb{E}[w_i] = 0$ for $\rho = \pm 1$, and $\mathbb{E}[w_i]$ is continuous over $\rho \in (-1, 1)$ due to the integration property of continuous functions over a continuous interval. Although the last term in (B.32) can not be expressed in closed form, it can be evaluated numerically. Note first that for fixed parameters $\varrho > 0$ and $\nu \leq 0.1$, the integration in (B.32) is monotonically decreasing in $k \geq 0$, and achieves the maximum at $k = 0$. For parameter values $k = 5$, $\nu = 0.1$ and $\varrho = 0.01$, Fig. B.1 plots $\mathbb{E}[w_i]$ as a function of ρ, whose maximum $\zeta_2 = 0.0213.$ is achieved at $\rho = 0$. Further, from the integration in (B.32) for fixed $k \geq 0$, $\mathbb{E}[w_i]$ is a monotonically increasing function of both ν and ϱ, and it is therefore safe to conclude that for all $0 < \nu \leq 0.1$, and $\varrho = 0.01$, we have

$$\mathbb{E}[w_i] \leq \zeta_2 = 0.0213. \tag{B.34}$$

Hence, we can infer that $\mathbb{E}[\chi_i(|\boldsymbol{a}_i^*\boldsymbol{h}|^2)] \leq 0.0213\|\boldsymbol{h}\|^2$ for $\nu < 0.1$, $\varrho = 0.01$, and $k = 5$. Since $[\chi_i(|\boldsymbol{a}_i^*\boldsymbol{h}|^2$'s are sub-exponential with sub-exponential norm of the order $\mathcal{O}(\|\boldsymbol{h}\|^2)$, Bernstein-type sub-exponential tail bound [119] confirms that

$$\mathbb{p}\Big(\frac{1}{m}\sum_{i=1}^{m}\frac{\chi_i(|\boldsymbol{a}_i^*\boldsymbol{h}|^2)}{\|\boldsymbol{h}\|^2} > (\zeta_2+\epsilon)\Big) < \mathrm{e}^{-c_7 m\epsilon^2} \tag{B.35}$$

for some numerical constant $\epsilon > 0$, provided that $\|\boldsymbol{h}\| \leq \|\boldsymbol{x}\|/10$. Finally, due to the fact that $w_i \leq 1$ for all $1 \leq i \leq m$, the following holds

$$\frac{1}{m}\sum_{i=1}^{m} w_i\chi_i(|\boldsymbol{a}_i^*\boldsymbol{h}|^2) < (\zeta_2+\epsilon)\|\boldsymbol{h}\|^2 \tag{B.36}$$

with probability at least $1 - \mathrm{e}^{-c_7 m\epsilon^2}$.

We have proved the bound in (B.24) for a fixed vector $\boldsymbol{h}$, and the uniform bound for all vectors $\boldsymbol{h}$ obeying $\|\boldsymbol{h}\| \leq \|\boldsymbol{x}\|/10$ can be obtained by similar arguments in the proof [147,

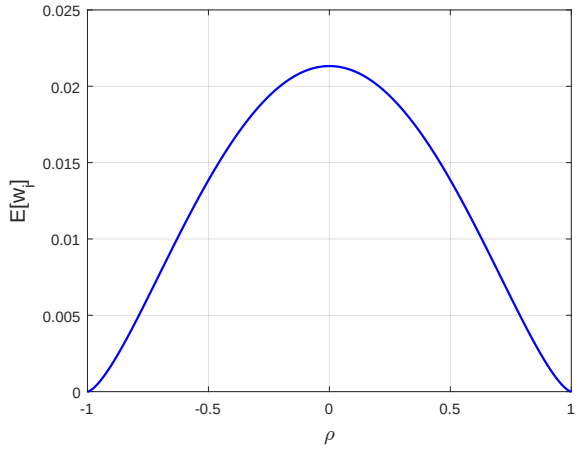

Figure B.1: The expectation $\mathbb{E}[w_i]$ as a function of ρ over $[-1, 1]$.

Lemma 9] with only minor changes in the constants.

Regarding the second bound (B.25), it is easy to see that

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} |\boldsymbol{a}_i^*\boldsymbol{h}|^2 \mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|\}} \\
&= \frac{1}{m}\sum_{i=1}^{m} \left[|\boldsymbol{a}_i^*\boldsymbol{h}|^2 \mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}} - |\boldsymbol{a}_i^*\boldsymbol{h}|^2 \mathbb{1}_{\{(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\}}\right] \\
&\leq (0.1271 - \zeta_2 + \epsilon)\|\boldsymbol{h}\|^2
\end{aligned} \tag{B.37}$$

where the last inequality follows from subtracting the bound in (B.24) of k from that corresponding to $k = 0$. To account for the weights $w_i = 1/[1 + \beta/(|\boldsymbol{a}_i^*\boldsymbol{z}|/|\boldsymbol{a}_i^*\boldsymbol{x}|)]$, first notice that $\boldsymbol{a}_i^*\boldsymbol{h} = \boldsymbol{a}_i^*\boldsymbol{z} - \boldsymbol{a}_i^*\boldsymbol{x}$, and that our second bound works with $(\boldsymbol{a}_i^*\boldsymbol{z})(\boldsymbol{a}_i^*\boldsymbol{x}) < 0$ in (B.19), hence $\frac{|\boldsymbol{a}_i^*\boldsymbol{z}|}{|\boldsymbol{a}_i^*\boldsymbol{x}|} \leq \frac{|\boldsymbol{a}_i^*\boldsymbol{h}|}{|\boldsymbol{a}_i^*\boldsymbol{x}|} - 1$. Recall that the second bound (B.25) assumes the event $\{|\boldsymbol{a}_i^*\boldsymbol{x}| < |\boldsymbol{a}_i^*\boldsymbol{h}| \leq (k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|\}$, implying $\frac{|\boldsymbol{a}_i^*\boldsymbol{z}|}{|\boldsymbol{a}_i^*\boldsymbol{x}|} \leq \frac{|\boldsymbol{a}_i^*\boldsymbol{h}|}{|\boldsymbol{a}_i^*\boldsymbol{x}|} - 1 \leq k$. Further, because w_i is monotonically increasing in $\frac{|\boldsymbol{a}_i^*\boldsymbol{z}|}{|\boldsymbol{a}_i^*\boldsymbol{x}|}$, then $w_i \leq \frac{1}{1+\beta/k}$. Taking this result back to (B.37) yields

$$\frac{1}{m}\sum_{i=1}^{m} w_i |\boldsymbol{a}_i^*\boldsymbol{h}|^2 \mathbb{1}_{\{|\boldsymbol{a}_i^*\boldsymbol{x}|<|\boldsymbol{a}_i^*\boldsymbol{h}|\leq(k+1)|\boldsymbol{a}_i^*\boldsymbol{x}|\}} \leq \frac{0.1271-\zeta_2+\epsilon}{1+\beta/k}\|\boldsymbol{h}\|^2 \tag{B.38}$$

which proves the second bound in (B.25).

Appendix C

Proofs for Chapter 5

C.1 Proof of Lemma 10

As elaborated in Sec. 5.2.1, there is a clear separation in the expected values $\mathbb{E}[Z_{i,j}] = \mathbb{E}[\psi_i^2 a_{i,j}^2]$ for $j \in \mathcal{S}$ and $j \notin \mathcal{S}$; that is,

$$\begin{aligned}\mathbb{E}[Z_{i,j}] = \mathbb{E}\big[(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x})^2 a_{i,j}^2\big] &= \mathbb{E}\big[a_{i,j}^4 x_j^2 + (\boldsymbol{a}_{i,/j}^{\mathcal{T}}\boldsymbol{x}_{/j})^2 a_{i,j}^2\big] \\ &= \begin{cases} \|\boldsymbol{x}\|_2^2, & j \notin \mathcal{S}, \\ \|\boldsymbol{x}\|_2^2 + 2x_j^2, & j \in \mathcal{S}. \end{cases}\end{aligned} \tag{C.1}$$

Consider the case of $j \in \mathcal{S}$ first. Based on $\mathbb{E}[a_{i,j}^{2p}] = (2p-1)!!$ with p being a positive integer and the symbol !! denoting the double factorial, $Z_{i,j}$ has second-order moment

$$\begin{aligned}\mathbb{E}[Z_{i,j}^2] &= \mathbb{E}\big[(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x})^4 a_{i,j}^4\big] \\ &= \mathbb{E}\big[a_{i,j}^8 x_j^4 + a_{i,j}^4 a_{i,\ell\neq j}^4 \|\boldsymbol{x}_{/j}\|_2^4 + 6a_{i,j}^6 x_j^2 a_{i,\ell\neq j}^2 \|\boldsymbol{x}_{/j}\|_2^2\big] \\ &= 105x_j^4 + 9\|\boldsymbol{x}_{/j}\|_2^4 + 90x_j^2\|\boldsymbol{x}_{/j}\|_2^2 \\ &= 9\|\boldsymbol{x}\|_2^4 + 24x_j^4 + 72x_j^2\|\boldsymbol{x}\|_2^2\end{aligned} \tag{C.2}$$

where $\ell \in \{1, 2, \ldots, n\}$ is some index from different than j. Letting $\tilde{Z}_j := \|\boldsymbol{x}\|_2^2 + 2x_j^2 - Z_{i,j}$ for all $j \in \mathcal{S}$, it holds that

$$\tilde{Z}_j \leq \|\boldsymbol{x}\|_2^2 + 2x_j^2 \leq 3\|\boldsymbol{x}\|_2^2.$$

Furthermore, one has $\mathbb{E}[\tilde{Z}_j] = 0$, and

$$\begin{aligned}\mathbb{E}[\tilde{Z}_j^2] &= \|\boldsymbol{x}\|_2^4 + 4x_j^4 + 4x_j^2\|\boldsymbol{x}\|_2^2 + \mathbb{E}[Z_{i,j}^2] - \big(2\|\boldsymbol{x}\|_2^2 + 4x_j^2\big)\mathbb{E}[Z_{i,j}] \\ &= 8\|\boldsymbol{x}\|_2^4 + 68x_j^2\|\boldsymbol{x}\|_2^2 + 20x_j^4 \\ &\le 96\|\boldsymbol{x}\|_2^4.\end{aligned}$$

Appealing to Lemma 17, one establishes for all $j \in \mathcal{S}$ that

$$\Pr\Big(\frac{1}{m}\sum_{i=1}^m \psi_i^2 a_{i,j}^2 - (\|\boldsymbol{x}\|_2^2 + 2x_j^2) \le -\epsilon\Big) \le \exp\Big(-\frac{m\epsilon^2}{192\|\boldsymbol{x}\|_2^4}\Big).$$

Taking $\epsilon = x_{\min}^2 := \min_{j\in\mathcal{S}} x_j^2 \le x_j^2$ leads to

$$\Pr\Big(\frac{1}{m}\sum_{i=1}^m \psi_i^2 a_{i,j}^2 \le \|\boldsymbol{x}\|_2^2 + x_{\min}^2\Big) \le \exp\Big(-\frac{mx_{\min}^4}{192\|\boldsymbol{x}\|_2^4}\Big).$$

Recalling our assumption that $x_{\min}^2$ is on the order of $(1/k)\|\boldsymbol{x}\|_2^2$, i.e.,

$$x_{\min}^2 = (C_1/k)\|\boldsymbol{x}\|_2^2$$

for some constant $C_1 > 0$, the following holds with probability at least $1 - 1/m$ for all $j \in \mathcal{S}$:

$$\min_{j\in\mathcal{S}} \frac{1}{m}\sum_{i=1}^m \psi_i^2 a_{i,j}^2 \ge \|\boldsymbol{x}\|_2^2 + x_{\min}^2 = \Big(1 + \frac{C_1}{k}\Big)\|\boldsymbol{x}\|_2^2 \tag{C.3}$$

provided that $m \ge C_0 k^2 \log(mn)$ for some absolute constant $C_0 > 0$.

Now let us turn to the case of $j \notin \mathcal{S}$, in which $\sum_{i=1}^m Z_{i,j} = \sum_{i=1}^m \psi_i^2 a_{i,j}^2$ is a weighted sum of χ_1^2 random variables. According to Lemma 18, it holds that

$$\Pr\Big(\sum_{i=1}^m \psi_i^2(a_{i,j}^2 - 1) > 2\sqrt{\epsilon}\Big(\sum_{i=1}^m \psi_i^4\Big)^{\frac{1}{2}} + 2\epsilon \max_i \psi_i^2\Big) \le \exp(-\epsilon). \tag{C.4}$$

In addition, for any constants ϵ', $\epsilon'' > 0$, Chebyshev's inequality together with the union bound

confirms that

$$\Pr\Big(\sum_{i=1}^{m}\psi_i^4 > \big(3m+\sqrt{96m}\epsilon'\big)\|\boldsymbol{x}\|_2^4\Big) \le 1/(\epsilon')^2 \tag{C.5a}$$

$$\Pr\Big(\max_{i=1}^{m}\psi_i^2 > \epsilon''\|\boldsymbol{x}\|_2^2\Big) \le 2m\exp(-\epsilon''/2). \tag{C.5b}$$

Take $\epsilon := \log(mn)$ in (C.4), $\epsilon' := \sqrt{m}$ and $\epsilon'' := 4\log(mn)$ in (C.5). Then, with probability at least $1-4/m$, the next holds for all $j \notin \mathcal{S}$ and $m > C'$

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\psi_i^2(a_{i,j}^2-1) &\le \frac{2}{m}\sqrt{\log(mn)}\sqrt{3m+\sqrt{96m}\sqrt{m}}\|\boldsymbol{x}\|_2^2 \\ &\quad + \frac{8}{m}\big(\log(mn)\big)^2\|\boldsymbol{x}\|_2^2 \\ &\le 8\sqrt{\frac{\log(mn)}{m}}\|\boldsymbol{x}\|_2^2 \end{aligned} \tag{C.6}$$

for some absolute constant $C' > 0$ depending on n.

On the other hand, the rotational invariance property of Gaussian distributions confirms that [23]

$$\psi_i^2 = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|^2 = |\boldsymbol{a}_{i,\mathcal{S}}^{\mathcal{T}}\boldsymbol{x}_{\mathcal{S}}|^2 \overset{d}{=} a_{i,j}^2\|\boldsymbol{x}\|_2^2$$

in which the symbol $\overset{d}{=}$ means that terms involved on both sides of the equality enjoy the same distribution. Since the χ^2 variables $a_{i,j}^2$ are sub-exponential, an application of Bernstein's inequality produces the tail bound

$$\Pr\Big(\frac{1}{m}\sum_{i=1}^{m}a_{i,j}^2 - 1 \ge \epsilon\Big) \le \exp(-m\epsilon^2/8) \tag{C.7}$$

for any $\epsilon \in (0,1)$, which can also be easily verified with a direct tail probability calculation from the tail probability of standard Gaussian distribution. Choosing $\epsilon := \sqrt{16\log(m)/m}$ with $m > C'$ gives rise to

$$\frac{1}{m}\sum_{i=1}^{m}\psi_{i,j}^2 \le \left(1+4\sqrt{\frac{\log m}{m}}\right)\|\boldsymbol{x}\|_2^2 \tag{C.8}$$

which holds true with probability at least $1-1/m$ for all $j \in [m]$. Putting results in (C.6) and

(C.8) together leads to

$$\max_{j\notin\mathcal{S}\subseteq[m]} \frac{1}{m}\sum_{i=1}^{m}\psi_i^2 a_{i,j}^2 \leq \left(1+12\sqrt{\frac{\log(mn)}{m}}\right)\|\boldsymbol{x}\|_2^2 \tag{C.9}$$

which holds with probability exceeding $1-5/m$ for large enough m.

The last inequality taken collectively with (C.3) suggests that there exists an event E_0 on which with probability at least $1-6/m$, the following holds

$$\begin{aligned}\min_{j\in\mathcal{S}} \frac{1}{m}\sum_{i=1}^{m}\psi_i^2 a_{i,j}^2 &\geq \left(1+\frac{C_1}{k}\right)\|\boldsymbol{x}\|_2^2 \\ &> \left(1+12\sqrt{\frac{\log(mn)}{m}}\right)\|\boldsymbol{x}\|_2^2 \\ &\geq \max_{j\notin\mathcal{S}} \frac{1}{m}\sum_{i=1}^{m}\psi_i^2 a_{i,j}^2 \end{aligned} \tag{C.10}$$

provided that $m \geq C_0 k^2 \log(mn)$ such that $C_0 \geq 144/C_1^2$ with $x_{\min}^2 = (C_1/k)\|\boldsymbol{x}\|_2^2$.

Appendix D

Supporting Lemmas

Lemma 17 ([12])**.** *For i.i.d. zero-mean random variables $X_1,\, X_2,\, \ldots,\, X_m$, if there exists some nonrandom constant $b > 0$ such that $X_i \leq b$ for $1 \leq i \leq m$, and $\mathbb{E}[X_i^2] = v^2$, then the following holds*

$$\Pr(X_1 + \cdots + X_m \geq y) \leq \min\Big(\exp\big(-\frac{y^2}{2\sigma^2}\big), c_0 - c_0\Phi\big(\frac{y}{\sigma}\big)\Big) \tag{D.1}$$

for $\sigma^2 := m\max(b^2,\, v^2)$, and the cumulative distribution function of the standard normal distribution $\Phi(\cdot)$, where one can take $c_0 = 25$.

Lemma 18 ([73])**.** *Let $X_1,\, X_2,\, \ldots,\, X_m$ be i.i.d. Gaussian random variables with zero mean and variance 1, and $b_1,\, b_2,\, \ldots, b_m$ be nonnegative. The following inequality holds for any $\epsilon > 0$*

$$\Pr\Big(\sum_{i=1}^m b_i(X_i^2 - 1) \geq 2\Big(\sum_{i=1}^m b_i^2\Big)^{\frac{1}{2}} \sqrt{\epsilon} + 2\big(\max_{i=1}^m b_i\big)\epsilon\Big) \leq \exp(-\epsilon). \tag{D.2}$$

Lemma 19. *[112, Lemma 7.17] For any k-sparse $\boldsymbol{x} \in \mathbb{R}^n$ supported on $\mathcal{S}$, assume noise-free measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ generated from i.i.d. Gaussian sampling vectors $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $i = 1,\, 2,\, \ldots,\, m$. Fixing any $\epsilon_1 > 0$, and for all $(2k)$-sparse $\boldsymbol{h} \in \mathbb{R}^n$, the following holds with probability at least $1 - 3\mathrm{e}^{-c_5 m}$*

$$\frac{1}{m}\sum_{i=1}^m \Big(\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|} - \frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\Big) |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}) \leq 2\frac{\sqrt{1+\epsilon_1}}{1-\rho_0}\Big(\epsilon_1 + \sqrt{\frac{21}{20}}\rho_0\Big)\|\boldsymbol{h}\|_2^2 \tag{D.3}$$

for all $\boldsymbol{z} \in \mathbb{R}^n$ obeying $\|\boldsymbol{z} - \boldsymbol{x}\|_2 \leq \rho_0\|\boldsymbol{x}\|_2$, provided that $m > c_6(2s)\log(n/(2s))$ for some fixed numerical constants $c_5,\, c_6 > 0$. Here, $\rho_0 = 1/10$.