August 2017

Algorithms and Theory for Robust PCA and Phase Retrieval

Huishuai Zhang
Syracuse University

ABSTRACT

In this dissertation, we investigate two problems, both of which require the recovery of unknowns from measurements that are potentially corrupted by outliers. The first part focuses on the problem of *robust principal component analysis* (PCA), which aims to recover an unknown low-rank matrix from a corrupted and partially-observed matrix. The robust PCA problem, originally nonconvex itself, has been solved via a convex relaxation based approach *principal component pursuit* (PCP) in the literature. However, previous works assume that the sparse errors uniformly spread over the entire matrix and characterize the condition under which PCP guarantees exact recovery. We generalize these results by allowing non-uniform error corruptions over the low-rank matrix and characterize the conditions on the error corruption probability of each individual entry based on the local coherence of the low-rank matrix, under which correct recovery can be guaranteed by PCP. Our results yield new insights on the graph clustering problem beyond the relevant literature.

The second part of the thesis studies the phase retrieval problem, which requires recovering an unknown vector from only its magnitude measurements. Differently from the first part, we solve this problem directly via optimizing nonconvex objectives. As the nonconvex objective is often constructed in such a way that the true vector is its global optimizer, the difficulty here is to design algorithms to find the global optimizer efficiently and provably. In order to solve this problem, we propose a gradient-like algorithm named reshaped Wirtinger flow (RWF). For random Gaussian measurements, we show that RWF enjoys linear convergence to a global optimizer as long as the number of measurements is on the order of the dimension of the unknown vector. This achieves the best possible sample complexity as well as the state-of-the-art computational efficiency.

Moreover, we study the phase retrieval problem when the measurements are corrupted

by adversarial outliers, which models situations with missing data or sensor failures. In order to resist possible observation outliers in an oblivious manner, we propose a novel median truncation approach to modify the nonconvex approach in both the initialization and the gradient descent steps. We apply the median truncation approach to the Poisson loss and the reshaped quadratic loss respectively, and obtain two algorithms *median-TWF* and *median-RWF*. We show that both algorithms recover the signal from a near-optimal number of independent Gaussian measurements, even when a constant fraction of measurements is corrupted. We further show that both algorithms are stable when measurements are corrupted by both sparse arbitrary outliers and dense bounded noises. We establish our results on the performance guarantee via the development of non-trivial concentration measures of the median-related quantities, which can be of independent interest.

ALGORITHMS AND THEORY FOR ROBUST PCA AND PHASE RETRIEVAL

by

Huishuai Zhang

B.E., University of Science and Technology of China, 2009
M.E., Graduate University of Chinese Academy of Sciences, 2012

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical & Computer Engineering

Syracuse University
August 2017

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. Yingbin Liang, for her unreserved support during my doctoral study. It is a great pleasure to work with her. I sincerely thank her not only for the immense knowledge and techniques she taught me but also for giving me the freedom to pursue eclectic research directions. Her advice and feedback on research were invaluable to me for achieving better results. Every discussion with her inspired me new ideas and helped me to go through the hard times in research. Without her help, I can not imagine how to accomplish the Ph.D journey.

I would like to thank Prof. Biao Chen, Prof. Pinyuen Chen, Prof. M. Cenk Gursoy, Prof. Qinru Qiu, Prof. Pramod K. Vashney for serving on my doctoral defense committee. They carefully read and provided many helpful suggestions to improve the presentation of the thesis.

I would also thank my great collaborators: Prof. Sholomo Shamai, Prof. Lifeng Lai, Prof. Yuejie Chi, Yuanxin Li, Hua Wang, Caiming Xiong, James Bradbury, Richard Socher, and Yi Zhou. Their intelligence and insight made our works even better. Special thanks go to Prof. Lifeng Lai and Prof. Eric Xing for hosting my visits in WPI and CMU.

Thank my lab mates and friends: Runchen, Jiayao, Weiguang, Shaofeng, Anhong, Yi, Yunhao, Zhe for their help in study and life. Their enthusiasm in research and professional attitude to work are great inspiration to me. I also would like to the friends in Syracuse, Pengfei, Xiao, Yu and Shiqi. They made me colorful days

during the intense research life.

I would like to give special thanks to Lei He for the undoubted support. I am very fortunate to know her and to be with her. The relationship brings me countless joy and unparalleled courage. I am indebted to her a lot.

I am greatly thankful to my family, my parents, my grandpa and my sister, who have offered me generous support in every aspect of my life. They are the people who believe in me during hard times and support me to go after my dream without any hesitance.

Thanks to people who ever helped me and encouraged me so far. Special thanks go to my mentors Xiaoyong Zhu, Peng Sun and Jinlin Wang at Institute of Acoustics, who helped me to pursue my research dream.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

This dissertation contains two topics: robust PCA and nonconvex phase retrieval, both of which require recovering unknowns from measurements that are potentially corrupted by outliers. The robust PCA problem is to recover a low-rank matrix from an observed matrix which contains random and sparse error corruptions. We provide a more refined analysis for robust PCA via *local coherence*, which is further shown to have rich implication on the graph clustering problem. The phase retrieval problem is to recover a vector from only its magnitude measurements. We develop new nonconvex methods that are *faster* and (or) *more robust* than existing ones and establish the theoretical performance guarantee for the proposed methods. At last, for the sake of completeness, my earlier Ph.D works investigate the problems on information theoretical security, which are summarized in Section 1.3.

1.1 Robust Principal Component Analysis (PCA)

Classical PCA is a widely-used technique in modern data analysis and dimension reduction. Mathematically, it is to find the first k principal components (eigenvectors corresponding to the k largest eigenvalues) of an observed matrix $\boldsymbol{M}$, and can be formulated as solving

the following optimization problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{L}} \quad & \|\boldsymbol{M}-\boldsymbol{L}\| \\ \text{subject to} \quad & \text{rank}(\boldsymbol{L}) \leq k, \end{aligned}$$

where $\|\cdot\|$ denotes the spectral norm, i.e., the largest singular value. This optimization problem can be solved efficiently via *singular value decomposition* (SVD).

However, the classical PCA is sensitive to grossly corrupted observations: a single corrupted entry in $\boldsymbol{M}$ could make the resulting $\boldsymbol{L}$ arbitrarily far from the ground truth. In order to perform PCA effectively when the data matrix is subject to gross corruptions, the robust PCA problem has been widely studied [5–12]. Mathematically, the robust PCA problem is to decompose an observed matrix $\boldsymbol{M}$ into the sum of a low-rank matrix $\boldsymbol{L}$ and a sparse error matrix $\boldsymbol{S}$. A direct optimization formulation for the robust PCA problem can be written as

$$\begin{aligned} \underset{\boldsymbol{L},\boldsymbol{S}}{\text{minimize}} \quad & \text{rank}(\boldsymbol{L})+\lambda\|\boldsymbol{S}\|_0 \\ \text{subject to} \quad & \boldsymbol{M}=\boldsymbol{L}+\boldsymbol{S}, \end{aligned} \tag{1.1}$$

where $\|\cdot\|_0$ denotes the number of non-zero entries and λ is the hyper-parameter that controls the balance between minimizing the rank of $\boldsymbol{L}$ and the sparsity of $\boldsymbol{S}$. However, the optimization (1.1) is non-convex and intractable in polynomial time.

One way to deal with the above difficulty is to consider the convex relaxation of (1.1). Specifically, the $\text{rank}(\cdot)$ can be relaxed as the nuclear norm $\|\cdot\|_*$, i.e., the sum of singular values, and $\|\cdot\|_0$ can be relaxed as the l_1 norm, i.e., the sum of absolute values of all entries. In [13, 14], the following principal component pursuit (PCP) is proposed to solve the robust PCA problem,

$$\textit{PCP:} \qquad \begin{aligned} \underset{\boldsymbol{L},\boldsymbol{S}}{\text{minimize}} \quad & \|\boldsymbol{L}\|_*+\lambda\|\boldsymbol{S}\|_1 \\ \text{subject to} \quad & \boldsymbol{M}=\boldsymbol{L}+\boldsymbol{S}. \end{aligned} \tag{1.2}$$

Surprisingly, it has been shown in [13, 14] that if the ground truth $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$ satisfy certain

assumptions, then PCP returns the exact $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$ with high probability.

Intuitively, in order to demix $\boldsymbol{M}$ correctly, the two matrices $\boldsymbol{L}_0$ and $\boldsymbol{S}_0$ should be distinguishable from each other in properties. This requires that $\boldsymbol{L}_0$ is not sparse and $\boldsymbol{S}_0$ is not low-rank. One important quantity that determines how similar $\boldsymbol{L}_0$ is to a sparse matrix is the *coherence* of $\boldsymbol{L}_0$, which measures how the column and the row spaces of $\boldsymbol{L}_0$ are aligned with the canonical basis and between themselves. Such a coherence parameter is defined to be the maximum over all of the column and the row spaces of $\boldsymbol{L}$ in [13] , which can be viewed as the *global* parameter for the entire matrix $\boldsymbol{L}$. It has been shown in [13] that exact decomposition can be guaranteed with high probability if the percentage of the corrupted entries is small enough as well as the corrupted entries are uniformly distributed over the entire matrix, and the low-rank matrix is incoherent. Consequently, such a characterization of the result is based on the global (and in fact *the worst case*) coherence parameter.

However, such results do not capture the fact that different parts of the low-rank matrix have different sensitivities to the error corruption. Our contribution is that we associate each entry (i, j) of the low-rank matrix $\boldsymbol{L}$ with a *local* coherence parameter. Such local coherence parameters allow us to study how robust each individual entry of $\boldsymbol{L}$ is to resist the error corruption. Our results also yield rich implications on the graph clustering problem. My Ph.D work with regard to the robust PCA problem so far has lead to one journal preparation and one conference publication [15].

1.1.1 Our Contribution: Robust PCA via Local Coherence

In our study of the robust PCA problem, we assume that the error corruptions are distributed non-uniformly over the entire matrix, and if the data matrix $\boldsymbol{M}$ is partially observed, the probability of observing each entry of $\boldsymbol{M}$ is non-uniform as well. We characterize the conditions that guarantee correct matrix decomposition by PCP in terms of the entry-wise error probability and the entry-wise observation probability with the local coherence parameter.

More specifically, for the full observation model (i.e., all entries are observed), we pro-

vide a more refined analysis of robust PCA than those in [13, 14] that captures how robust each entry of the low-rank matrix combats error corruption. Moreover, we demonstrate that our results have several immediate implications on the graph clustering problem. For the partial observation model, we establish the performance guarantee of matrix recovery from both error corruption and partial observation based on the local coherence by an adaptive PCP. In order to deal with non-uniform error corruption and non-uniform missing observation, our technical proof introduces a new weighted norm and develops/exploits the concentration properties for such a norm.

1.1.2 Related Works

Matrix completion. A closely related but different problem from robust PCA is *matrix completion*, in which a low-rank matrix is partially observed and is to be completed. Such a problem has been previously studied in [16–20], and it was shown that a rank-r n-by-n matrix can be provably recoverable with as few as $\Theta(\max\{\mu_0, \mu_1\}nr\log^2 n)$ observed entries[1], where μ_0 and μ_1 are the coherence parameters determined by the underlying low-rank matrix and the definitions are introduced in Chapter 2. Later on, it was shown in [21] that μ_1 does not affect sample complexity for matrix completion and hence $\Theta(\mu_0 nr\log^2 n)$ observed entries are sufficient for guaranteeing correct matrix completion. It was further shown in [22] that a coherent low-rank matrix (i.e., with large μ_0) can be recovered with $\Theta(nr\log^2 n)$ observations as long as the sampling probability is proportional to the leverage score (i.e., localized μ_0). Our problem can be viewed as its counterpart in robust PCA, where the difference lies in the fact that the local coherence in our problem depends on both localized μ_0 and μ_1.

Moreover, matrix completion with bounded or random noise (i.e., sub-Gaussian noise) has been studied in [23–25], where the error bounds between the recovered matrix and the underlying low-rank matrix were established in terms of the noise magnitude or variance.

[1] $f(n) = \Theta(g(n))$ represents $k_1 \cdot g(n) \leq f(n) \leq k_2 \cdot g(n)$ for some positive k_1, k_2.

The matrix recovery problem with random measurements has also been studied in literature. In [26], matrix recovery is guaranteed from Gaussian measurements (i.e. each entry of a measurement matrix is independent standard Gaussian random variable) via a convex heuristic algorithm, which is in spirit similar to PCP. Furthermore, rank-one measurements were considered in [27], which greatly reduces the complexity. The performance guarantee was established under the *ROP* condition.

Robust PCA. Besides the works [13, 14, 28–30] mentioned in previous sections, [31] further showed that if signs of nonzero entries in the sparse matrix are randomly chosen, then an adjusted convex optimization can produce exact decomposition even when the percentage of corrupted entries goes to one (i.e., error is dense). Moreover, [26] discussed how to decompose two matrices if the observations are linear measurements and [32] considered the low-rank and sparse decomposition under data compression. Compared to previous studies, our work aims at understanding how local coherence affects susceptibility of each matrix entry to error corruption and missing observation, and provides a more refined (i.e. entry-wise) view of robust PCA with random error matrix.

Signal demixing. The problems of decoupling multiple signals which are sparse in corresponding bases have also been studied in convex demixing literature [33–36]. When specialized to the low-rank and sparse demixing problem (i.e., robust PCA), such a formulation assumes that the singular vectors of the low-rank matrix are drawn from orthonormal basis randomly. Sharp phase transition has been established between the low-rank matrix and the sparse matrix that are generally incoherent. Our work differentiates from these studies because we assume that the low-rank matrix is unknown but *deterministic* and moreover we focus on how the coherence pattern (locally) of the low-rank matrix affects the success of PCP.

Nonconvex approach. More recently, provable nonconvex approaches have been developed to solving the matrix decomposition problem, e.g., [37, 38] and the matrix completion problem, e.g., [17, 39–43]. The general idea is to minimize a nonconvex loss function

by first estimating a good initial seed and then refining the estimation via iterative method. The proof of the performance guarantee typically exploits the geometry of the loss function around the neighborhood of the global minimizer. While our focus is along the line of convex relaxation to solve the robust PCA problem, we anticipate that our perspective of analyzing non-uniform error corruption and missing observation is also an important issue to address when applying nonconvex approaches.

1.2 Phase Retrieval via Nonconvex Optimization

The formulation of phase retrieval is motivated by the problem in X-ray crystallography and coherent diffraction imaging [49–51], which requires the recovery of the structure of an object from the recorded far field diffracted intensity when a source beam is incident on the object. Mathematically, the phase retrieval problem is formulated as recovering a vector $\boldsymbol{x} \in \mathbb{R}^n/\mathbb{C}^n$ from measurements y_i given by

$$y_i = |\langle \boldsymbol{a}_i, \mathbf{x} \rangle|^2, \quad \text{for } \ i = 1, \cdots, m, \tag{1.3}$$

where $\boldsymbol{a}_i \in \mathbb{R}^n/\mathbb{C}^n$ are random design vectors (known).

Various algorithms have been proposed to solve this problem since 1970s. The error-reduction methods proposed in [52, 53] work well empirically but lack theoretical guarantees. More recently, by using the lifting trick, the phase retrieval problem is solved via convex relaxations [45, 54–59] which estimate a rank-one positive semidefinite matrix $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^T$ from linear measurements. In particular, when the measurement vectors $\boldsymbol{a}_i$'s are composed of *independent and identically distributed* (i.i.d.) Gaussian entries, Phaselift [45] perfectly recovers all $\boldsymbol{x} \in \mathbb{R}^n$ with high probability as long as the number m of measurements is on the order of signal dimension n. The reader can refer to the review paper [60] to learn more about applications and algorithms of the phase retrieval problem.

While with good theoretical guarantee, these convex methods often suffer from com-

putational complexity especially when the signal dimension is large. Another natural idea is to recover the signal as a solution to some loss minimization problem:

$$\min_{\boldsymbol{z}} \frac{1}{2m} \sum_{i=1}^{m} \ell(\boldsymbol{z}; y_i) \tag{1.4}$$

where $\ell(\boldsymbol{z}, y_i)$ is postulated using the negative likelihood of Gaussian or Poisson noise model. Since the measurements are quadratic in $\boldsymbol{x}$, the objective function is nonconvex.

Recently, Candès et al. introduced *Wirtinger flow* (WF) algorithm [1], which minimizes the following nonconvex loss function

$$\ell_{WF}(\boldsymbol{z}) := \frac{1}{4m} \sum_{i=1}^{m} (|\boldsymbol{a}_i^T \boldsymbol{z}|^2 - y_i)^2. \tag{1.5}$$

The WF algorithm first obtains a good initialization by the spectral method and then iteratively refines the estimation via the gradient descent scheme. It is proved that the WF algorithm guarantees signal recovery with $\mathcal{O}(n \log n)$ Gaussian measurements and attains $\epsilon-$accuracy within $\mathcal{O}(mn^2 \log 1/\epsilon)$ flops.

The WF algorithm is further improved by *truncated Wirtinger flow* (TWF) algorithm [61]. TWF adopts a Poisson loss function of $|\boldsymbol{a}_i^T \boldsymbol{z}|^2$, which is given by

$$\ell_{Poi}(\boldsymbol{z}) := \frac{1}{2m} \sum_{i=1}^{m} \left(|\boldsymbol{a}_i^T \boldsymbol{z}|^2 - y_i \log |\boldsymbol{a}_i^T \boldsymbol{z}|^2 \right). \tag{1.6}$$

When calculating the initial seed and every step of gradient, TWF keeps only well-behaved measurements based on carefully designed truncation thresholds. Such truncation assists to yield linear convergence with a fixed step size and reduces the sample complexity to $\mathcal{O}(n)$ and the convergence time to $\mathcal{O}(mn \log 1/\epsilon)$.

Furthermore, incremental/stochastic methods have been proposed to solve the phase retrieval problem. Specifically, [62, 63] demonstrated the excellent empirical performance of the randomized Kaczmarz method but did not establish the non-asymptotic global conver-

gence guarantee. [64] proposed another stochastic algorithm named *incremental truncated Wirtinger flow* (ITWF) which is developed upon TWF.

Our results improve over previous methods from two directions. We propose an algorithm "reshaped Wirtinger flow (RWF)", which is much faster than other existing nonconvex solvers by employing a new loss function. We further propose a median truncation approach to robustify the process of phase retrieval when observations are corrupted by arbitrarily-valued outliers. In a summary, my Ph.D work on the phase retrieval problem so far has lead to two journal submissions [65, 66] and two conference publications [67, 68].

1.2.1 Our Contribution: Reshaped Wirtinger Flow

We adopt the following loss function for solving the phase retrieval problem:

$$\ell(\boldsymbol{z}) := \frac{1}{2m}\sum_{i=1}^{m}\left(|\boldsymbol{a}_i^T\boldsymbol{z}| - \sqrt{y_i}\right)^2. \tag{1.7}$$

Compared to the loss function (1.5) of WF that adopts the quadratic loss of $|\boldsymbol{a}_i^T\boldsymbol{z}|^2$, the above loss function adopts the quadratic loss of $|\boldsymbol{a}_i^T\boldsymbol{z}|$ and hence has lower-order variables. While both loss functions are nonconvex in $\boldsymbol{z}$, our loss function in (1.7) is also nonsmooth.

To minimize such a nonconvex and nonsmooth loss function (1.7), we develop a gradient descent algorithm, which sets the "gradient" to zero corresponding to nonsmooth samples. We refer to such an algorithm together with an initialization using a new spectral method (different from that employed in TWF or WF) as *reshaped Wirtinger flow* (RWF). We show that this loss function has great advantage in both statistical and computational efficiency, in spite of nonsmoothness. We further develop incremental/stochastic versions of RWF, called incremental RWF (IRWF), and show that IRWF also enjoys the advantageous local curvature of RWF, and achieves excellent statistical and computational performance. Along the way, we establish the convergence analysis of the randomized Kaczmarz method by interpreting it as a variant of IRWF.

Statistically, we show that RWF recovers the true signal with $\mathcal{O}(n)$ Gaussian measurements, which is order-wise optimal. Thus, RWF improves the sample complexity $\mathcal{O}(n \log n)$ of WF, and achieves the same sample complexity as TWF but without truncation in the gradient descent loops. Computationally, RWF converges linearly to the true signal with a constant step size, requiring $\mathcal{O}(mn \log(1/\epsilon))$ flops to reach ϵ-accuracy. Again, without truncation in the gradient descent step, RWF improves the computational cost $\mathcal{O}(mn^2 \log(1/\epsilon)$ of WF and achieves the same computational cost as TWF. Numerically, RWF and its incremental versions require fewer parameters, e.g., truncation thresholds, than TWF in practice. RWF is generally two times faster than TWF and four to six times faster than WF in terms of both the number of iterations and time cost. IRWF also outperforms existing incremental as well as batch algorithms.

1.2.2 Our Contribution: Median Truncation Approach

The WF-type algorithms are evaluated based on their *statistical* and *computational* performances: statistically, we hope the sample complexity m to be as small as possible; computationally, we hope the run time to be as small as possible. As can be seen, existing WF-type algorithms are already near-optimal both statistically and computationally. We now introduces a third consideration, which is the *robustness to outliers*, where we hope that the algorithm continues to perform well even in the presence of outliers that may take arbitrary values. This bears great importance in practice, because outliers arise frequently from the phase imaging applications [71] due to various reasons such as detector failures, recording errors, and missing data.

Specifically, suppose the set of m measurements are given as

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 + \eta_i, \quad i = 1, \cdots, m, \tag{1.8}$$

where $\eta_i \in \mathbb{R}/\mathbb{C}$ for $i = 1, \ldots, m$ are outliers that can take arbitrary values. We assume

that outliers are sparse with no more than sm nonzero entries, i.e., $\|\boldsymbol{\eta}\|_0 \leq sm$, where $\boldsymbol{\eta} = \{\eta_i\}_{i=1}^m \in \mathbb{R}^m/\mathbb{C}^m$. Here, s is a nonzero constant, representing the faction of measurements that is corrupted.

We aim to develop phase retrieval algorithms with both statistical and computational efficiency, and provable robustness to arbitrary-valued outliers. None of the existing algorithms meet all of the three considerations simultaneously. The performance of WF-type algorithms (WF, TWF and RWF) is very sensitive to outliers which introduce anomalous search directions when their values are excessively deviated. While a form of Phaselift [72] is robust to a constant fraction of outliers, it is computationally too expensive.

In the presence of outliers, the signal of interest may no longer be the global optima of (1.4). Therefore, we want to only include the clean samples in the optimization, which is, however, impossible as we do not assume any *a priori* knowledge of the outliers. Our key strategy is to prune the bad samples adaptively and iteratively, using a gradient descent procedure that proceeds as follows:

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{m} \sum_{i \in T_{t+1}} \nabla \ell(\boldsymbol{z}^{(t)}; y_i). \tag{1.9}$$

where $\boldsymbol{z}^{(t)}$ denotes the tth iterate of the algorithm, $\nabla \ell(\boldsymbol{z}^{(t)}; y_i)$ is the gradient of $\ell(\boldsymbol{z}^{(t)}; y_i)$ for $t = 0, 1, \ldots$, and μ is the step size. In each iteration, only a subset T_{t+1} of data-dependent and iteration-varying samples contributes to the search direction. But how to select the set T_{t+1}? Note that the gradient of the loss function typically contains the term $\left|y_i - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2\right|$ (for WF and TWF) or $\left|\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|\right|$ (for RWF), which measures the residual using the current iterate. With y_i being corrupted by arbitrarily large outliers, the gradient can deviate the search direction from the signal arbitrarily.

Inspired by the utility of *median* to combat outliers in robust statistics [73], we prune samples whose gradient components $\nabla \ell(\boldsymbol{z}^{(t)}; y_i)$ are much larger than the *sample median* to control the search direction. Hiding some technical details, this gives the main ingredient

of our *median-truncated gradient descent* update rule[2], i.e., for each iterate $t \geq 0$:

$$T_{t+1} := \{i : |y_i - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2| \lesssim \mathsf{med}(\{|y_i - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2\}_{i=1}^m)\}, \quad \text{for TWF}, \tag{1.10}$$

$$T_{t+1} := \{i : |\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|| \lesssim \mathsf{med}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|\}_{i=1}^m)\}, \quad \text{for RWF}, \tag{1.11}$$

where $\mathsf{med}(\cdot)$ denotes the sample median. The robust property of median lies in the fact that the median cannot be arbitrarily perturbed unless the outliers dominate the inliers [73]. This is in sharp contrast to the sample mean, which can be made arbitrarily large even by a single outlier. Thus, using the sample median in the truncation rule can effectively remove the impact of outliers.

By applying the median truncation strategy to the loss functions used in TWF and RWF, we obtain two new median-truncated gradient descent algorithms, median-TWF and median-RWF, respectively. The median-TWF and median-RWF algorithms do not assume a priori knowledge of the outliers, such as their existence or the number of outliers, and therefore can be used in an oblivious fashion. Importantly, we establish the following performance guarantees.

For the Gaussian measurement model, we show that with high probability, median-TWF and median-RWF recover all signal $\boldsymbol{x}$ up to the global sign at a linear rate of convergence, even with a constant fraction of outliers, as long as the number of measurements m is on the order of $n \log n$. Furthermore, the reconstruction is stable in the presence of additional bounded dense noise. Statistically, the sample complexity of both algorithms is near-optimal up to a logarithmic factor, and to reassure, they continue to work even when outliers are absent. Computationally, both algorithms converge linearly, requiring a mere computational cost of $\mathcal{O}(mn \log 1/\epsilon)$ to reach ϵ-accuracy. More importantly, our algorithms now tolerate a constant fraction of arbitrary outliers, without sacrificing performance otherwise. To the best of our knowledge, this is the first application of the median to

[2]Please see the exact form of the algorithms in Chapter 4.

robustify high-dimensional statistical estimation in the presence of arbitrary outliers with rigorous non-asymptotic performance guarantees.

1.2.3 Related Works

Along the line of developing nonconvex algorithms with global performance guarantee for the phase retrieval problem, [78] developed alternating minimization algorithm, [1, 61, 67, 79] developed/studied first-order gradient-like algorithms, and a recent study [80] characterized geometric structure of the nonconvex objective and designed a second-order trust-region algorithm.

The work "reshaped Wirtinger flow" is most closely related to [1, 61, 67, 69], but develops a new gradient-like algorithm based on new loss function that yields advantageous statistical/computational efficiency. The quadratic loss function of magnitudes was also used in the early literature of phase retrieval [53] with Fourier magnitude measurements. However, no global convergence guarantee was available in [53]. Stochastic algorithms were also developed for the phase retrieval problem. [64] studied the incremental truncated Wirtinger flow (ITWF) and showed that ITWF needs much fewer passes of data than TWF to reach the same accuracy. [62] adapted the Kaczmarz method to solve the phase retrieval problem and demonstrated its fast empirical convergence. We show that IRWF is closely related to Kaczmarz-PR, and empirically runs faster than ITWF thanks to the advantageous curvature of the loss function.

After our work was posted on arXiv, an independent work [70] was subsequently posted, which also adopts the same loss function but develops a slightly different algorithm called TAF (i.e., truncated amplitude flow). One major difference is that RWF does not require truncation in the gradient loops while TAF still employs truncation. Hence, RWF has fewer parameters to tune, and is easier to implement than TAF in practice. Furthermore, RWF demonstrates the performance advantage of adopting a lower-order loss function even without truncation, which cannot be observed from TAF. The two algorithms

also employ different initialization strategies. Moreover, we analyze stochastic algorithms based on the new loss function while [70] does not.

The work of median truncated approach for robust phase retrieval is closely related to the TWF algorithm [61] which is a truncated gradient descent algorithm for phase retrieval. However, the truncation rule in TWF is based on the sample mean, which is very sensitive to outliers. In [71, 72, 98], the problem of phase retrieval under outliers is investigated, but the proposed algorithms either lack performance guarantees or are computationally too expensive.

The adoption of median in machine learning is not unfamiliar, for example, K-median clustering [74] and resilient data aggregation for sensor networks [75]. Our work here further extends the applications of median to robustifying high-dimensional estimation problems with theoretical guarantees. Another popular approach in robust estimation is to use the trimmed mean [73], which has found success in robustifying sparse regression [76], subspace clustering [77], etc. However, using the trimmed mean requires knowledge of the number of outliers, whereas median does not require such information.

More generally, various problems have been studied by minimizing nonconvex loss functions. For example, a partial list of these studies include matrix completion [17, 39–43, 81, 82], low-rank matrix recovery [83–88], robust PCA [37], robust tensor decomposition [89], dictionary learning [90, 91], community detection [92], phase synchronization [93], blind deconvolution [94, 95], etc.

1.3 Secret Key Generation over a Source Network

At he earlier stage of my Ph.D, I studied the secret key generation problems using the information theoretic approaches. I briefly introduce my work on this topic here for the completeness of the thesis.

The problem of secret key generation via public discussion under the source model was

initiated by [99, 100]. In the basic source-type model, two legitimate terminals observe correlated source sequences and wish to establish a common secret key by communicating with each other over a public channel which can be accessed by eavesdroppers. The secret key is required to be kept secure from eavesdroppers. The main observation is that, due to the correlation between two source sequences, terminal $\mathcal{Y}$ can recover terminal $\mathcal{X}$'s source sequence by letting terminal $\mathcal{X}$ send limited amount of information using distributed source coding technique [101]. Then both terminal $\mathcal{X}$ and terminal $\mathcal{Y}$ can generate a shared secret key based on terminal $\mathcal{X}$'s source sequence. The key capacity is given by

$$C(K) = H(X) - H(X|Y) = I(X;Y), \tag{1.12}$$

which can be interpreted as the information rate in terminal $\mathcal{X}$'s source sequence subtracting the rate of information released over the public channel. The close connection between the distributed source coding and secret key generation also holds for more general source-type models [102]. In particular, [102] studied a general source network with multiple terminals, in which a subset of terminals need to agree on a common secret group key and the remaining terminals act as dedicated helpers. It was shown in [102] that the secret key capacity is equal to the joint entropy of all source observations subtracting the minimum amount of communications needed to enable the subset of terminals to recover all source observations. Consider an example case, in which all terminals in the set $\mathcal{M}$ wish to agree on a secret key. Then in [102], the secret key capacity is given by

$$C(K) = H(X_{\mathcal{M}}) - R_{CO}, \tag{1.13}$$

where $X_{\mathcal{M}} = (X_i : i \in \mathcal{M})$ and R_{CO} is the minimum rate of "communication for omniscience". Here, omniscience means that all legitimate terminals recover source observations of all terminals. We note that the secret key capacity in (1.13) can be intuitively interpreted as the entire source information rate $H(X_{\mathcal{M}})$ minus the minimum transmission

rate R_{CO} revealed to the public in order to achieve omniscience at all terminals.

Such problems of single-key generation were also studied for other scenarios in [103–107]. However, for some practical scenarios, multiple keys need to be simultaneously generated. For instance, suppose that a number of terminals have different security clearance levels, and each terminal is allowed to access the confidential documents up to its own clearance level. In such a case, terminals with the same clearance level should share the same key, and should be kept ignorant of higher level keys.

Capacity Region for Secret-key Private-key Model. One fundamental multi-key generation model that captures the above hierarchical scenario is initialized by Ye and Narayan [108], which has been referred to as the secret-key private-key model in literature. In the model of [108], three terminals with correlated source observations wish to agree on a common secret key required to be secure from an eavesdropper, while two designated terminals aim to generate another private key required to be secure from both the eavesdropper and the third terminal. [108] provided outer and inner bounds on the key capacity region and showed that the inner and outer bounds match for one special case, but left the problem whether the bounds match for the other two cases open. In our work, we developed a novel random binning and joint decoding scheme which achieves the outer-bound for the other cases and fully characterized the key capacity region. These results have been published in [109, 110].

Multiple Key Generation over a Cellular Network. Another type of multi-key models are referred to as *cellular models*, e.g., [111, 112], in which a central terminal (base station) wishes to agree on independent keys respectively with a number of (mobile) terminals. These models are well-motivated in cellular networks, in which mobile terminals need to share independent secret keys with the base station in order to achieve secure communication. In our work, we studied the models of generating two keys simultaneously over four terminals with a helper and characterized the key capacity region under various secrecy constraints. Moreover, we also characterized the key capacity region for generat-

ing arbitrary number of secret keys simultaneously, which significantly extends the result reported in the previous work [112]. These results have been published in [113–115].

Secret Key Capacity with Vocal Constraint. As we introduce at the beginning of this section, the secret key capacity of generating one key over m terminals is established in [102] by applying the "omniscience" scheme. Namely, each terminal reveals information to the public until the omniscience attains, i.e., every terminal knows all terminals' observations. However, such an omniscience scheme is not always necessary to achieve the secret key capacity. In our work, we characterized a sufficient condition when a specific terminal does not have to talk to achieve the secret key capacity and further gave a sufficient condition when a specific terminal must talk, which disproves the conjecture made in [116] when $m > 3$. Our results published in [118] on various situations, jointly with one more situation studied in [117] fully solved this problem.

1.4 Organization of the Dissertation

The rest of the dissertation is organized as follows. In Chapter 2, we describe the algorithm and theory for robust PCA via local coherence. In Chapter 3, we present the reshaped Wirtinger flow algorithm for solving the phase retrieval problem. In Chapter 4 we present our results for robust phase retrieval. In Chapter 5, we describe directions for future work.

CHAPTER 2
ANALYSIS OF ROBUST PCA VIA LOCAL COHERENCE

In this chapter, we study the robust PCA problem via local coherence. We first formulate the problem mathematically in Section 2.1. We then present our main results on the full observation model and the partial observation model in Section 2.2 and Section 2.3. We describe the numerical experiments in Section 2.4 and summarize this chapter in Section 2.5. At last, we provide technical proofs in Section 2.6.

We provide some notations that are used throughout this chapter. A matrix X is associated with five norms: $\|X\|_F$ denotes the Frobenius norm, $\|X\|_*$ denotes the nuclear norm (i.e., the sum of singular values), $\|X\|$ denotes the spectral norm (i.e., the largest singular value), and $\|X\|_1$ and $\|X\|_\infty$ respectively represent the l_1 and l_∞ norms of the long vector stacked by X. The inner product between two matrices is defined as $\langle X, Y\rangle := trace(X^*Y)$, where X^* represents the *conjugate transpose* of X. For a linear operator $\mathcal{A}$ that acts on the space of matrices, $\|\mathcal{A}\|$ denotes the operator norm given by $\|\mathcal{A}\| = \sup_{\{\|X\|_F=1\}} \|\mathcal{A}X\|_F$.

2.1 Problem Statement

As we introduce in Chapter 1.1, the robust PCA problem aims to decompose an observation matrix, say M, into the sum of a low-rank matrix, say L, and a sparse matrix (i.e., the error matrix), say S. A popular algorithm to solve this problem is PCP, which is formulated as the following optimization problem:

$$\textit{PCP:} \qquad \begin{aligned} \underset{L,S}{\text{minimize}} \quad & \|L\|_* + \lambda \|S\|_1 \\ \text{subject to} \quad & M = L + S, \end{aligned} \tag{2.1}$$

where $\|\cdot\|_*$ denotes the nuclear norm, i.e., the sum of singular values, and $\|\cdot\|_1$ denotes the l_1 norm i.e., the sum of absolute values of all entries.

The success of PCP is highly related with the coherence of the low-rank matrix L and the sparsity of the matrix S. We next introduce the definition of coherence mathematically. Namely, suppose that L is a rank-r matrix with SVD $L = U\Sigma V^*$, where Σ is a $r \times r$ diagonal matrix with singular values as its diagonal entries, U is a $n \times r$ matrix with columns as the left singular vectors of L, V is a $n \times r$ matrix with columns as the right singular vectors of L, and V^* denotes the transpose of V. The *coherence* of L is measured by $\mu = \max\{\mu_0, \mu_1\}$, where μ_0 and μ_1 are defined as

$$\|U^* \boldsymbol{e}_i\| \leq \sqrt{\frac{\mu_0 r}{n}}, \qquad \|V^* \boldsymbol{e}_j\| \leq \sqrt{\frac{\mu_0 r}{n}}, \quad \text{for all } i, j = 1, \cdots, n \tag{2.2}$$

$$\|UV^*\|_\infty \leq \sqrt{\frac{\mu_1 r}{n^2}}. \tag{2.3}$$

Previous studies of robust PCA, e.g., [13, 29–31], often assume that every entry of the low-rank matrix is corrupted with the same probability and show that correct decomposition via PCP requires the error density to be below a certain threshold determined by the coherence parameter μ of the low rank matrix L.

In this chapter, on the contrary, we investigate the robust PCA problem under the as-

sumption that the errors (non-zero entries) are non-uniformly distributed over the entire sparse matrix and entries are observed non-uniformly. Such model generalization naturally requires to connect the robustness of each entry to the local coherence. We thus define the following *local* coherence parameters μ_{ij},

$$\mu_{0ij} := \frac{n}{2r}\left(\|U^* e_i\|^2 + \|V^* e_j\|^2\right), \quad \mu_{1ij} := \frac{n^2([UV^*]_{ij})^2}{r} \tag{2.4}$$

$$\mu_{ij} := \max\{\mu_{0ij}, \mu_{1ij}\}. \tag{2.5}$$

It is clear that $\mu_{0ij} \leq \mu_0$ and $\mu_{1ij} \leq \mu_1$ for all $i, j = 1, \cdots, n$. We note that although $\max_{i,j} \mu_{ij} > 1$, some μ_{ij} might take values as small as zero.

We note that the notion of local coherence was first introduced in [21] for studying the matrix completion problem, in which local coherence determines the local sampling density in order to guarantee correct matrix completion. Here, local coherence plays a similar role, and determines the maximal allowable error density at each entry to guarantee correct matrix decomposition. The difference lies in that local coherence here depends on both localized μ_0 and μ_1 rather than only on localized μ_0 in matrix completion. This difference is unavoidable due to the further challenge arising in robust PCA, in which locations of error corrupted entries are unknown [13, 30].

Mathematically, the model can be described as follows.

- Assume that L is an $n \times n$ rank r matrix, where $r < n$. Assue that L is unknown , to be recovered[1].
- Suppose that L is partially observed on a set $O \subset [n] \times [n]$, where O is random defined by Bernoulli sampling such that

$$\mathbb{P}((i, j) \in O) = \alpha_{ij}. \tag{2.6}$$

- Each observed entry $L_{ij}, (i, j) \in O$ is corrupted by an arbitrary noise S_{ij} with prob-

[1]Here, we focus on square matrices for simplicity. Our results can be extended to rectangular matrices.

ability ρ_{ij} independently. We denote the support set of S by Ω, and hence $\Omega \subseteq O$. Conditioning on $(i,j) \in O$, assume that $(i,j) \in \Omega$ for $i,j \in [n]$ are independent events with

$$\mathbb{P}((i,j) \in \Omega | (i,j) \in O) = \rho_{ij}, \tag{2.7}$$

which implies that

$$\mathbb{P}((i,j) \in \Omega) = \alpha_{ij}\rho_{ij}. \tag{2.8}$$

- Define a set $\Gamma := O \backslash \Omega$, then

$$\mathbb{P}((i,j) \in \Gamma) = \alpha_{ij}(1-\rho_{ij}). \tag{2.9}$$

- We observe $M = \mathcal{P}_O(L) + S$, where $\mathcal{P}_O(\cdot)$ denotes the projection on set O. The aim is to recover the low-rank matrix L based on the observation matrix M.

We study both the *random sign* and *fixed sign* models for S. For the fixed sign model, we assume that the signs of nonzero entries in S are arbitrary and fixed, whereas for the random sign model, we assume that the signs of nonzero entries in S are independently distributed Bernoulli variables, randomly taking values $+1$ and -1 with probability $1/2$ as follows:

$$[\text{sgn}(S)]_{ij} = \begin{cases} 1 & \text{with prob.} \quad \alpha_{ij}\rho_{ij}/2 \\ 0 & \text{with prob.} \quad 1-\alpha_{ij}\rho_{ij} \\ -1 & \text{with prob.} \quad \alpha_{ij}\rho_{ij}/2. \end{cases} \tag{2.10}$$

Under the assumption of full observation ($\alpha_{ij} = 1$, for all $i,j = 1,\ldots,n$), the problem reduces to the robust PCA problem, i.e., decomposing matrix $M = L + S$ into a low-rank

matrix L and a sparse (error) matrix S.

In this chapter, our goal is to characterize conditions on α_{ij} and ρ_{ij} that guarantee correct recovery of L and S with observation of M via a convex programming algorithm.

2.2 Full Observation Model

In this section, we consider the full observation model, i.e., the robust PCA problem, where M is fully observed, and provide the performance guarantee under which L and S can be recovered correctly via convex programming. Then we discuss several implications of our results on the graph clustering problem.

We first consider the full observation problem under the *random sign model* as introduced in Section 2.1. The following theorem characterizes the condition that guarantees correct recovery by PCP.

Theorem 2.1. *Consider the full observation problem under the random sign model. If*

$$1 - \rho_{ij} \geq \max\left\{ C_0 \sqrt{\frac{\mu_{ij} r}{n}} \log n, \frac{1}{n^3} \right\}$$

for some sufficiently large constant C_0 and for all $i, j \in [n]$, then PCP yields correct matrix decomposition of L and S with $\lambda = \frac{1}{32\sqrt{n \log n}}$, with probability at least $1 - cn^{-10}$ for some constant c.

We note that the term $1/n^3$ is introduced to justify dual certificate conditions in the proof (see Section 2.6). We further note that satisfying the condition in Theorem 2.1 implies $C_0 \sqrt{\mu r / n} \log n \leq 1$, which is an essential bound required in our proof and coincides with the conditions in previous studies [13, 29]. Although we set $\lambda = \frac{1}{32\sqrt{n \log n}}$ for the sake of proof, in practice λ is often determined via cross validation.

The above theorem suggests that the local coherence parameter μ_{ij} is closely related to the robustness of individual entry of L to error corruption in matrix decomposition.

An entry corresponding to smaller μ_{ij} (i.e., lower coherent) tolerates larger error density ρ_{ij}. This is consistent with the result in [21] for matrix completion, in which lower local coherence requires lower local sampling rate. The difference lies in that here both μ_{0ij} and μ_{1ij} play roles in μ_{ij} whereas only μ_{0ij} matters in matrix completion. This is the critical difference between the proofs of robust PCA and matrix completion. The necessity of μ_{1ij} for robust PCA is further demonstrated in Section 2.2.1 via an example.

Theorem 2.1 also provides a more refined view for robust PCA in the dense error regime, in which the error corruption probability approaches one. Such an interesting regime was previously studied in [30, 31]. In [31], it is argued that PCP with adaptive λ yields exact recovery even when the error corruption probability approaches one if errors take random signs and the dimension n is sufficiently large. In [30], it is further shown that PCP with a fixed λ also yields exact recovery and the scaling behavior of the error corruption probability is characterized. The above Theorem 2.1 further provides the scaling behavior of the *local entry-wise* error corruption probability ρ_{ij} as it approaches one, and captures how such scaling behavior depends on local coherence parameters μ_{ij}. Such a result implies that robustness of PCP depends not only on the error density but also on how errors are distributed over the matrix with regard to μ_{ij}.

We next consider the full observation problem under the *fixed sign model* as introduced in Section 2.1. In this case, non-zero entries of the error matrix S can take arbitrary and fixed values, and only locations of non-zero entries are random.

Theorem 2.2. *Consider the full observation problem under the fixed sign model. If*

$$(1-2\rho_{ij}) \geq \max\left\{C_0\sqrt{\frac{\mu_{ij}r}{n}}\log n, \frac{1}{n^3}\right\}$$

for some sufficiently large constant C_0 and for all $i, j \in [n]$, then PCP yields correct recovery with $\lambda = \frac{1}{32\sqrt{n\log n}}$, with probability at least $1 - cn^{-10}$ for some constant c.

Theorem 2.2 follows from Theorem 2.1 by adapting the elimination and derandomiza-

tion arguments [13, Section 2.2] as follows. Let $\boldsymbol{\rho}$ be the matrix with each (i, j)-entry being ρ_{ij}. If PCP yields exact recovery with a certain probability for the random sign model with the parameter $2\boldsymbol{\rho}$, then it also yields exact recovery with at least the same probability for the fixed sign model with locations of non-zero entries sampled using Bernoulli model with the parameter $\boldsymbol{\rho}$. The detailed argument is provided in Section 2.6.2.

We now compare Theorem 2.2 for robust PCA with *non-uniform* error corruption to Theorem 1.1 in [13] for robust PCA with *uniform* error corruption. It is clear that if we set $\rho_{i,j} = \rho$ for all $i, j \in [n]$, then the two models are the same. It can then be easily checked that conditions $\sqrt{\mu r/n} \log n \leq \rho_r$ and $\rho \leq \rho_s$ in Theorem 1.1 of [13] implies the conditions in Theorem 2.2. Thus, Theorem 2.2 provides a more relaxed condition than Theorem 1.1 in [13]. Such benefit of condition relaxation should be attributed to the new golfing scheme introduced in [29,30], and we provides a more refined view of robust PCA by further taking advantage of such a new golfing scheme to analyze *local coherence* conditions.

More importantly, Theorem 2.2 characterizes relationship between local coherence parameters and local error corruption probabilities, which implies that different areas of the low-rank matrix have different levels of ability to resist errors: a less coherent area (i.e., with smaller μ_{ij}) can tolerate more errors. Thus, Theorem 2.2 illustrates the following interesting fact. Whether PCP yields correct recovery depends not only on the total number of errors but also on how errors are distributed. If more errors are distributed to less coherent areas (i.e, with smaller μ_{ij}), then more errors in total can be tolerated. However, if errors are distributed in an opposite manner, then only small number of errors can be tolerated.

2.2.1 Implication on Graph Clustering

In this subsection, we further illustrate our result when the low-rank matrix is a cluster matrix. Although robust PCA and even more sophisticated approaches have been applied to solve clustering problems, e.g., [119–121], our perspective here is to demonstrate how

local coherence affects entry-wise robustness to error corruption, which has not been well addressed in previous studies.

Suppose there are n elements to be clustered. We use a cluster matrix L to represent the clustering relationship of these n elements with $L_{ij} = 1$ if elements i and j are in the same cluster and $L_{ij} = 0$ otherwise. Thus, with appropriate ordering of the elements, L is a block diagonal matrix with all diagonal blocks containing all '1's and off-diagonal blocks containing all '0's. Hence, the rank r of L equals the number of clusters, which is typically much smaller than n. Suppose these entries are corrupted by errors that flip entries from one to zero or from zero to one. This can be thought of as adding a (possibly sparse) error matrix S to L so that the observed matrix is $L + S$. Then PCP can be applied to recover the cluster matrix L.

More specifically, for a cluster matrix L, we use K_l to denote the size of the lth cluster, and further use $K_{(i)}$ to denote the size of the cluster containing element i. We use ρ_{ij} to denote the probability of flipping L_{ij}, i.e., the probability of error corruption of the entry (i, j). Then following Theorem 2.2, we obtain the following conditions that guarantee PCP to recover the true cluster matrix L.

Corollary 2.3. *Consider the problem of recovering the cluster matrix L via PCP. If*

$$1 - 2\rho_{ij} \geq \begin{cases} C_0 \sqrt{\frac{1}{2K_{(i)}} + \frac{1}{2K_{(j)}}} \log n & \text{for } i, j \text{ in different clusters} \\ C_0 \frac{\sqrt{n}}{K_{(i)}} \log n & \text{for } i, j \text{ in the same cluster} \end{cases} \tag{2.11}$$

for some sufficiently large constant C_0 and for all $i, j \in [n]$, then PCP yields correct recovery with $\lambda = \frac{1}{32\sqrt{n \log n}}$, with probability at least $1 - cn^{-10}$ for some constant c.

Proof. The proof of this corollary is immediate by applying Theorem 2 with specific μ_{ij}

given by the clustering problem, where $\mu_{ij} = \max\{\mu_{0ij}, \mu_{1ij}\}$ and

$$\mu_{0ij} = 1 \text{ for all } (i,j), \quad \text{and} \quad \mu_{1ij} = \begin{cases} r, & (i,j) \text{ is in diagonal blocks,} \\ 0, & (i,j) \text{ is in off-diagonal blocks.} \end{cases}$$

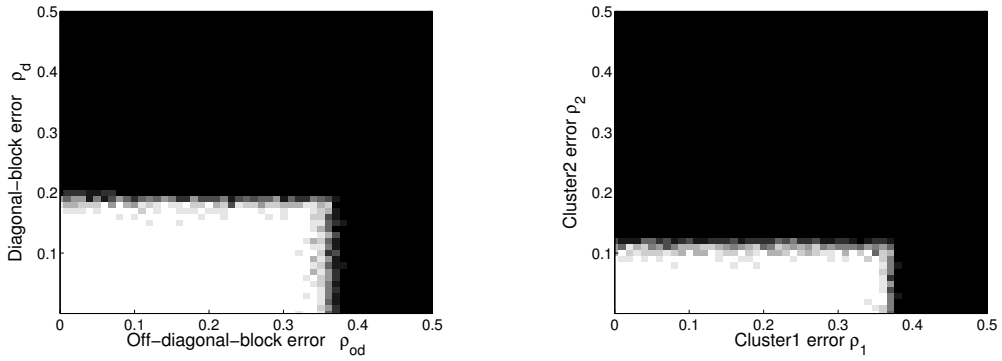
□

In comparison, the result (Theorem 4) in [122] requires $(1 - 2\rho_{ij}) \geq C_0 \frac{\sqrt{n}}{K_{\min}} \log n$ with $K_{\min}$ being the smallest cluster size, which gives a worst-case bound on every ρ_{ij}. However, our result captures the fact that different locations have different sensitivities to error corruption: off-diagonal entries are more robust to error corruptions than block-diagnal entries and large clusters are more robust than small clusters. To further illustrate these facts, we first consider an example with clusters having equal size n/r. We set $n = 600$ and $r = 4$ (i.e., four equal-size clusters). We apply errors to diagonal-block entries and off-diagonal-block entries respectively with the probabilities ρ_d and ρ_{od}. In Figure 2.1(a), we plot recovery accuracy of PCP for each pairs of (ρ_{od}, ρ_d). It is clear from the figure that failure occurs for larger ρ_{od} than ρ_d, which implies that off-diagonal blocks are more robust to errors than diagonal blocks. This can be explained by Corollary 2.3 as follows. For a cluster matrix with equal cluster size n/r, Corollary 2.3 requires

$$1 - 2\rho_{ij} \geq \begin{cases} C_0 \frac{r}{\sqrt{n}} \log n, & (i,j) \text{ is in diagonal blocks} \\ C_0 \sqrt{\frac{r}{n}} \log n, & (i,j) \text{ is in off-diagonal blocks} \end{cases}$$

to guarantee correct recovery. Clearly, off-diagonal entries allow larger ρ_{ij} compared to diagonal entries.

Moreover, this example also demonstrates the necessity of μ_1 in the robust PCA problem. [21] showed that μ_1 is not necessary for matrix completion and argued informally that μ_1 is necessary for robust PCA by connecting the robust PCA problem to hardness of find-



(a) Diagonal-block error vs. off-diagonal-block error

(b) Error vulnerability with respect to cluster sizes 500 vs. 100

Fig. 2.1: Error vulnerability on different parts for cluster matrix. In both cases, for each probability pair, we generate 10 trials of independent random error matrices and count the number of successes of PCP. We declare a trial to be successful if the recovered $\hat{L}$ satisfies $\|\hat{L} - L\|_F / \|L\|_F \leq 10^{-3}$. Color from white to black represents the number of successful trials changes from 10 to 0.

ing a small clique in a large random graph. Here, the above example provides an evidence for such a fact. In the example, μ_{0ij} are the same over the entire matrix, and hence it is μ_{1ij} that differentiates coherence between diagonal blocks and off-diagonal blocks, and thus differentiates their robustness to errors.

We then consider the case with two clusters that have different sizes, i.e., Cluster 1 has size 500 and Cluster 2 has size 100. Hence, $r = 2$. We apply errors to block diagonal entries corresponding to clusters 1 and 2 respectively with the probabilities ρ_1 and ρ_2. In Figure 2.1(b), we plot the recovery accuracy of PCP for each pair of (ρ_1, ρ_2). It is clear that failure occurs for larger ρ_1 than ρ_2, which thus implies that entries corresponding to the larger cluster are more robust to errors than entries corresponding to smaller clusters. This can be explained by Corollary 2.3 because the error corruption probabilities for block-diagnonal entries satisfy $1 - 2\rho_{ij} > C_0 \frac{\sqrt{n}}{K_{(i)}} \log n$ for correct recovery, and hence a larger cluster can resist denser errors. This also coincides with the results on graph clustering in [119].

Furthermore, Corollary 2.3 also implies a natural bound on the minimum cluster size $K_{\min} \geq \sqrt{n} \log n$ as n goes to infinity, which is consistent with the bound on the smallest

clique size in the classical planted clique problem [123] up to a logarithm factor. The planted clique problem is described as follows. For a Erdos-Renyi random graph $G(n, \frac{1}{2})$, randomly pick a subset of $K_{\min}$ nodes and make them fully connected (i.e., these $K_{\min}$ nodes form a clique), where the modified graph is represented by G'. The goal is to find the clique in the graph G'. This problem is believed to be intractable in polynomial time for the regime of $K_{\min} = o(\sqrt{n})$ [124, 125]. Corollary 2.3 is consistent with such a general belief by suggesting that if the size of the planted clique is greater than $\sqrt{n}\log n$, then it is possible to find such a clique via PCP, which performs efficiently in polynomial time.

2.3 Partial Observation Model

In this section, we generalize our main results to the case when the matrix is partially observed, as described in Section 2.1. Such a study can also be viewed as a refined analysis of studies [13, 29, 30], which are based on a global coherence parameter.

In order to establish the performance guarantee of the low-rank matrix recovery from both error corruptions and missing observations, we modify the original PCP as follows:

$$\textit{Adaptive PCP:} \qquad \begin{array}{ll} \underset{L,S}{\text{minimize}} & \|L\|_* + \|\Lambda \circ S\|_1 \\ \text{subject to} & Y = \mathcal{P}_O(L) + S. \end{array} \tag{2.12}$$

We set parameter Λ_{ij} according to the local observation probabilty α_{ij}. Hence, the parameter $\Lambda = [\Lambda_{ij}]$ is an $n \times n$ matrix instead of a scalar in previous results. We note that such setting is for obtaining a tighter bound in the proof rather than for the effectivity of the algorithm. The following theorem characterizes the conditions under which the *adaptive PCP* returns correct recovery.

Theorem 2.4. *Consider the partial observation model under the random sign assumption. Suppose that the partial observation probabilities α_{ij} are known and $\alpha_{ij} > \log^2 n/n$. Then*

if

$$\sqrt{\alpha_{ij}}(1-\rho_{ij}) \geq C_0\sqrt{\frac{\mu_{ij}r}{n}}\log n \quad \textit{and} \quad C_0\sqrt{\frac{\mu r}{n}}\log n \leq 1$$

for some sufficiently large constant C_0, the adaptive PCP with $\Lambda = \left[\frac{1}{32\sqrt{\alpha_{ij}n}\log n}\right]$ recovers the true low-rank matrix L with probability at least $1 - cn^{-10}$ for some constant c.

We note that here the mimimal observation probability is $\frac{\log^2 n}{n}$, which is essentially from the proof of Lemma 4 and consistent with the result of [122]. We note that $\lambda_{\min} \leq \|\Lambda\|_\infty \leq \lambda_{\max}$, where $\lambda_{\min} = \frac{1}{32\sqrt{n}\log n}$ and $\lambda_{\max} = \frac{1}{32\log^2 n}$. This result generalizes Theorem 2.1 to the partial observation model. It coincides with the results in [29, 30] when all ρ_{ij} are the same and all α_{ij} are the same. In the case when the sampling probability α_{ij} is unknown, standard PCP still recovers L by treating unobserved entries as being corrupted by error S. Hence, the following performance guarantee hods.

Proposition 2.5. *The standard PCP with $\lambda = \frac{1}{32\sqrt{n}\log n}$ recovers the true low-rank matrix L with high probability given*

$$\alpha_{ij}(1-\rho_{ij}) \geq C_0\sqrt{\frac{\mu_{ij}r}{n}}\log n \quad \textit{and} \quad C_0\sqrt{\frac{\mu r}{n}}\log n \leq 1,$$

for some sufficiently large constant C_0.

Proof. If we view the missing observation as error corruption, then the entry (i, j) is corrupted with probability $\rho'_{ij} := (1 - \alpha_{ij}) + \alpha_{ij}\rho_{ij}$. Thus $1 - \rho'_{ij}$ satisfies the condition in Theorem 2.1, which concludes the proof. □

2.4 Numerical Experiments

In this section, we provide numerical experiments to demonstrate our theoretical results. In these experiments, we adopt an augmented Lagrange multiplier algorithm in [127] to solve the PCP. We set $\lambda = 1/\sqrt{n\log n}$. A trial of PCP (for a given realization of error locations) is declared to be successful if $\hat{L}$ recovered by PCP satisfies $\|\hat{L} - L\|_F/\|L\|_F \leq 10^{-3}$.

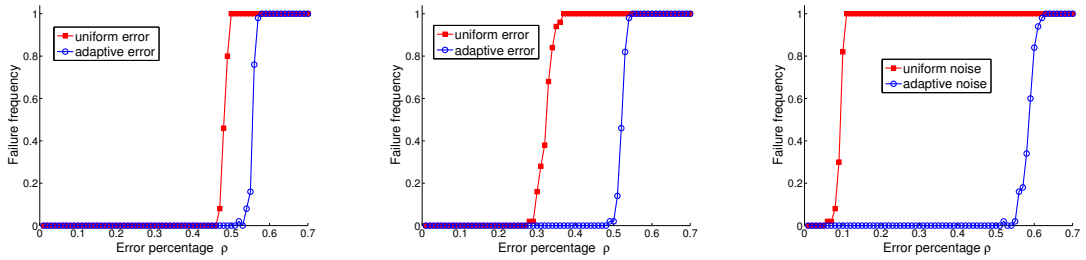
We apply the following three models to construct the low-rank matrix L.

- Bernoulli model: $L = XX^*$ where X is $n \times r$ matrix with entries independently taking values $+1/\sqrt{n}$ and $-1/\sqrt{n}$ equally likely.
- Gaussian model: $L = XX^*$, where X is $n \times r$ matrix with entries independently sampled from Gaussian distribution $\mathcal{N}(0, 1/n)$.
- Cluster model: L is a block diagonal matrix with r equal-size blocks containing all '1's.

In order to demonstrate that the local coherence parameter affects local robustness to error corruptions, we study the following two types of error corruption models.

- Uniform error corruption: $\text{sgn}(S_{ij})$ is generated as (2.10) with $\rho_{ij} = \rho$ for all $i, j \in [n]$, and $S = \text{sgn}(S)$.
- Adaptive error corruption: $\text{sgn}(S_{ij})$ is generated as (2.10) with $\rho_{ij} = \rho \frac{n^2 \sqrt{1/\mu_{ij}}}{\sum_{ij} \sqrt{1/\mu_{ij}}}$ for all $i, j \in [n]$, and $S = \text{sgn}(S)$.

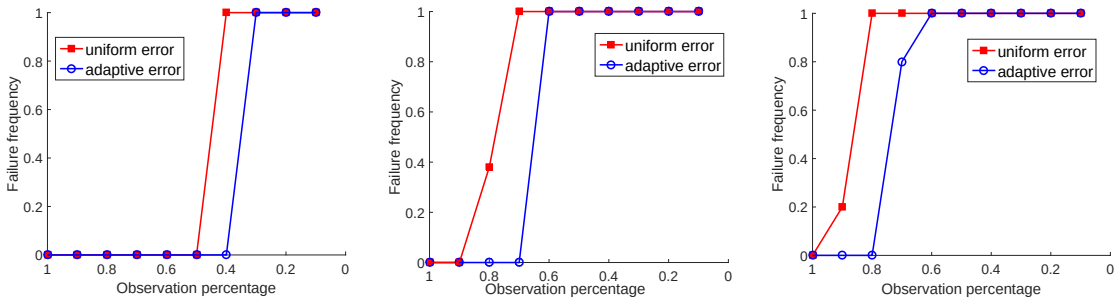
It is clear in both cases, the error matrix has the same average error corruption percentage ρ, but in adaptive error corruption, the local error corruption probability is adaptive to the local coherence.



(a) Bernoulli model (b) Gaussian model (c) Cluster model

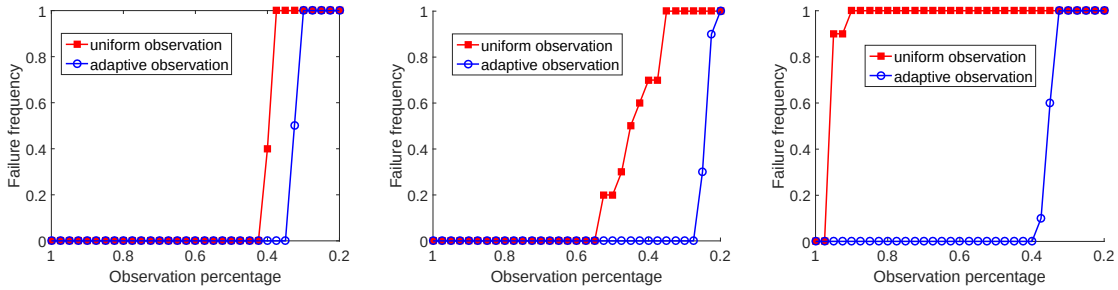
Fig. 2.2: Recovery failure frequency of PCP versus error corruption percentage.

Our first experiment demonstrates that robustness of PCP to error corruption not only depends on the number of errors but also depends on how errors are distributed over the matrix. For all three low-rank matrix models, we set $n = 1200$ and rank $r = 10$. For each



(a) Bernoulli model (b) Gaussian model (c) Cluster model

Fig. 2.3: Recovery failure frequency of PCP versus observation percentage: adaptive error.



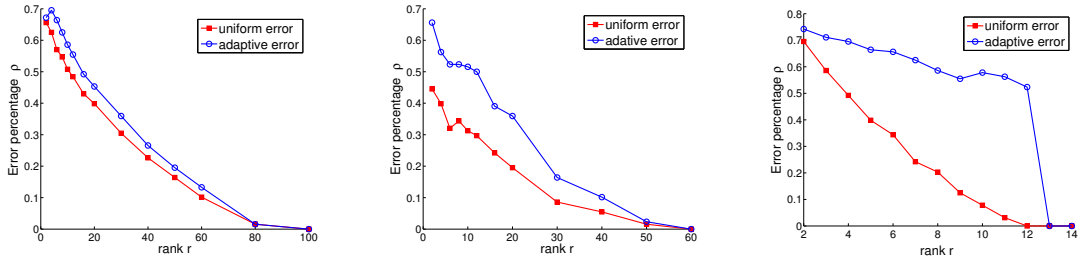
(a) Bernoulli model (b) Gaussian model (c) Cluster model

Fig. 2.4: Recovery failure frequency of PCP versus observation percentage: adaptive observation.

low-rank matrix model, we apply the uniform and adaptive error matrices, and plot the failure frequency of PCP versus the error corruption percentage ρ in Figure 2.2. For each value of ρ, we perform 50 trials of independent error corruption and count the number of failures of PCP. Each plot of Figure 2.2 compares robustness of PCP to uniform error corruption (the red square line) and adaptive error corruption (the blue circle line). We observe that PCP can tolerate more errors in the adaptive case. This is because the adaptive error matrix is distributed based on the local coherence parameter, where error density is higher in areas where matrices can tolerate more errors. Furthermore, comparison among the three plots in Figure 2.2 illustrates that the gap between uniform and adaptive error matrices is the smallest for Bernoulli model and the largest for cluster model. Our theoretic results suggest that the gap is due to the variation of the local coherence parameter across the matrix, which can be measured by the variance of μ_{ij}. Larger variance of μ_{ij} should yield larger gap. Our

numerical calculation of the variances for three models yield $\mathsf{Var}(\mu_{Bernoulli}) = 1.2109$, $\mathsf{Var}(\mu_{Gaussian}) = 2.1678$, and $\mathsf{Var}(\mu_{cluster}) = 7.29$, which confirms our explanation.

We next consider the partial observation setting. For this series of experiments, we assume the uniform observation, i.e., each entry is observed with the same probability. The observation probability changes from 1 to 0.2 with a step size 0.1. For each observed matrix, we apply the uniform and adaptive error matrices, with the same error percentage 0.2. We perform 50 trials for each of the Bernoulli Model, Gaussian Model and Cluster Model, and Figure 2.3 shows the simulation results. It is clear from the figure that PCP becomes more successful with an adaptive error under the partial observation setting. Note that the gaps between uniform error and adaptive error are not as large as those under the fully observation setting. This is because missing observations play the same role as uniform errors, so that the total errors do not adapt to local coherence as well as in the full observation model. We then further consider the setting where the noise is uniformly distributed with the error percentage 0.1, and the observation can be either adapted to the local coherence μ_{ij} or not with the probability of observation ranging from 1 to 0.2. The results are shown in Figure 2.4. Each plot of Figure 2.4 compares robustness of PCP to a uniform error corruption under either uniform missing observations (the red square line) or adaptive missing observations (the blue circle line). We observe that PCP can tolerate more missing observations in the adaptive case. This is because the adaptive missing observations are distributed based on the local coherence parameter.



(a) Bernoulli model (b) Guassian model (c) Cluster model

Fig. 2.5: Allowable error corruption percentage versus rank of L given PCP yields correct recovery.

Lastly, we study the phase transition between the rank of the low-rank matrix and error corruption probability. For the three low-rank matrix models, we set $n = 1200$. In Figure 2.5, we plot the error corruption percentage versus the rank L for both uniform and adaptive error corruption models. Each point on the curve records the maximum allowable error corruption percentage under the corresponding rank such that PCP yields correction recovery. We count a (r, ρ) pair to be successful if nine trials out of ten are successful. We first observe that in each plot of Figure 2.5, PCP is more robust in adaptive error corruption due to the same reason explained above. We further observe that the gap between the uniform and adaptive error corruption changes as the rank changes. In the low-rank regime, the gap is largely determined by the variance of coherence parameter μ_{ij} as we argued before. As the rank increases, the gap is more dominated by the rank and less affected by the local coherence. Eventually for large enough rank, no error can be tolerated no matter how errors are distributed. This is because the correctness of PCP is influenced by the coherence parameter and the rank in a same manner, as indicated in Theorem 2.1.

2.5 Conclusion

We characterize refined conditions under which PCP recovers correctly the low-rank matrix under both error corruptions and missing observations. Our result shows that the correct recovery via PCP depends not only on the total number of corrupted and missing entries but also on locations of these entries, more essentially on the local coherence of the low-rank matrix. Such result is well supported by our numerical experiments. Moreover, our result has rich implication when the low-rank matrix is a cluster matrix, and our result not only coincides with recent studies on clustering problems via low-rank cluster matrix recovery but also provides complementary insights for these problems. Our introduction of the new weighted norm for analysis of performance guarantee can be useful for studying other problems under non-uniform error corruptions or non-uniform random sampling.

2.6 Technical Proofs

2.6.1 Proof of Theorem 2.4 (and Theorem 2.1)

In this subsection, we provide the proof for Theorem 2.4, which yields Theorem 2.1 by specializing $\alpha_{ij} = 1$ and $\Lambda_{ij} = \frac{1}{32\sqrt{n}\log n}$ for all $i, j = 1, \ldots, n$.

The proof of Theorem 2.4 follows the general idea established in [13] and further developed in [29, 30]. The main novel technical development lies in analysis of non-uniform error corruption based on local coherence parameters, for which we introduce a new weighted norm $l_{w(\infty)}$, and establish concentration properties and bounds associated with this norm. As a generalization of matrix infinity norm, $l_{w(\infty)}$ incorporates both μ_{0ij} and μ_{1ij}, and is hence different from the weighted norms $l_{\mu(\infty)}$ and $l_{\mu(\infty,2)}$ in [22]. We next provide the main steps of the proof, with the proofs of technical lemmas regulated to Section 2.6.3.

We first introduce some notations. We define the subspace $T := \{UX^* + YV^* : X, Y \in \mathbb{R}^{n\times r}\}$, where U, V are left and right singular matrix of L. Then T induces a projection operator $\mathcal{P}_T$ given by $\mathcal{P}_T(M) = UU^*M + MVV^* - UU^*MVV^*$. Moreover, the complement subspace $T^\perp$ to T induces an orthogonal projection operator $\mathcal{P}_{T^\perp}$ with $\mathcal{P}_{T^\perp}(M) = (I - UU^*)M(I - VV^*)$. We further define two operators associated with Bernoulli sampling. Let Ω_0 denote a generic subset of $[n] \times [n]$. We define a corresponding projection operator $\mathcal{P}_{\Omega_0}$ as $\mathcal{P}_{\Omega_0}(M) = \sum_{ij} \mathbb{I}_{\{(i,j)\in\Omega_0\}}\langle M, e_ie_j^*\rangle e_ie_j^*$, where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. If Ω_0 is a random set generated by Bernoulli sampling with $\mathbb{P}((i,j) \in \Omega_0) = t_{ij}$ with $0 < t_{ij} \leq 1$ for all $i, j \in [n]$, we further define a linear operator $\mathcal{R}_{\Omega_0}$ as $\mathcal{R}_{\Omega_0}(M) = \sum_{ij} \frac{1}{t_{ij}}\mathbb{I}_{\{(i,j)\in\Omega_0\}}\langle M, e_ie_j^*\rangle e_ie_j^*$. For two variables a and b, $a \vee b = \max\{a, b\}$.

We introduce a new weighted norm. Suppose that μ_{ij}'s are local coherence parameters of L as defined in (2.5). Let $\hat{w}_{ij} = \sqrt{\frac{\mu_{ij}r}{n^2}}$ and $w_{ij} = \max\{\hat{w}_{ij}, \epsilon\}$, where ϵ is the smallest

nonzero $\hat{w}_{ij}$. Here ϵ is introduced to avoid singularity. Then for any matrix Z, define

$$\|Z\|_{w(\infty)} = \max_{i,j} \frac{|Z_{ij}|}{w_{ij}}. \tag{2.13}$$

It is easy to verify that $\|\cdot\|_{w(\infty)}$ is a well defined norm. We establish several concentration inequalities for this weighted infinity norm, which facilitate the proof of dual certificate construction.

We further note that "with high probability" means "with probability at least $1-cn^{-10}$", where the constant c may be different in different contexts.

Our proof includes two main steps: establishing that existence of a certain dual certificate is sufficient to guarantee correct recovery and constructing such a dual certificate. We first establish a number of supporting lemmas, and then proceed the two main steps.

We next provide a number of concentration properties for non-uniform sampling operators. These properties are in parallel to those under uniform sampling operators used in [13, 29, 30]. More specifically, Lemma 2.6 is proven in [22], which readily implies Lemma 2.8. We develop other lemmas based on the local coherence, the proofs of which are provided in Section 2.6.3.

Lemma 2.6. *[22, Lemma 9] Suppose* $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ *for all* $i,j \in [n]$. *If* $q_{ij} \geq \min\{C_0(\mu_{0ij} r \log n)/n, 1\}$ *for some sufficiently large constant* C_0 *and for all* $i,j \in [n]$, *then with high probability*

$$\|\mathcal{P}_T - \mathcal{P}_T \mathcal{R}_{\Omega_0} \mathcal{P}_T\| \leq \frac{1}{2}. \tag{2.14}$$

Lemma 2.7. *Suppose* $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ *for all* $i,j \in [n]$. *If* $\|\mathcal{P}_T - \mathcal{P}_T \mathcal{R}_{\Omega_0} \mathcal{P}_T\| \leq \frac{1}{2}$ *and* $q_{ij} \geq p_0$ *for all* $i,j \in [n]$, *then*

(a) $\|\mathcal{P}_T \mathcal{R}_{\Omega_0}\| \leq \sqrt{\frac{3}{2p_0}}$;

(b) $\mathcal{P}_{\Omega_0} \mathcal{P}_T$ *is injective on* T.

Lemma 2.8. *Suppose* $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ *for all* $i, j \in [n]$. *For a fixed matrix* $Z \in T$, *if* $q_{ij} \geq \min\{C_0(\mu_{ij} r \log n)/n, 1\}$ *for some sufficiently large constant* C_0 *and for all* $i, j \in [n]$, *then with high probability*

$$\|Z - \mathcal{P}_T \mathcal{R}_{\Omega_0}(Z)\|_F \leq \frac{1}{2}\|Z\|_F. \tag{2.15}$$

Lemma 2.9. *Suppose* $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ *for all* $i, j \in [n]$. *For a fixed matrix* $Z \in T$, *if* $q_{ij} \geq \min\left\{C_0(\sqrt{\mu_{ij} r} \vee \mu_{ij} r)\frac{\log n}{n}, 1\right\}$ *for some sufficiently large constant* C_0 *and for all* $i, j \in [n]$, *then with high probability*

$$\|(\mathcal{R}_{\Omega_0} - I)Z\| \leq \frac{C}{C_0}\|Z\|_{w(\infty)} \tag{2.16}$$

for some constant C.

Lemma 2.10. *Suppose* $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ *for all* $i, j \in [n]$. *Suppose* $\beta > 0$ *is a scaling factor. For a fixed matrix* $Z \in T$, *if* $q_{ij} \geq \min\left\{C_0 \beta^{-2}(\mu_{ij} r)\frac{\log n}{n}, 1\right\}$ *for some sufficiently large* C_0 *and for all* $i, j \in [n]$, *then with high probability*

$$\|(\mathcal{P}_T \mathcal{R}_{\Omega_0} - \mathcal{P}_T)Z\|_{w(\infty)} \leq \frac{1}{2}\beta\|Z\|_{w(\infty)}. \tag{2.17}$$

Dual Certificate Condition

We adopt the idea of [29] and introduce an equivalent model that is easy to deal with.

1. Define two independent random subsets of $[n] \times [n]$: Γ with $\mathbb{P}((i,j) \in \Gamma) = \alpha_{ij}(1 - \rho_{ij})$ and Ω' with $\mathbb{P}((i,j) \in \Omega') = \frac{\alpha_{ij}\rho_{ij}}{1-\alpha_{ij}+\alpha_{ij}\rho_{ij}}$. Let $O = \Gamma \cup \Omega'$, then $\mathbb{P}((i,j) \in O) = \alpha_{ij}$.

2. Define $\Omega := \Omega' \backslash \Gamma = \{(i,j) : (i,j) \in \Omega' \text{ and } (i,j) \notin \Gamma\}$. Then $\mathbb{P}\{(i,j) \in \Omega\} = \alpha_{ij}\rho_{ij}$.

3. Define a matrix W with each entry W_{ij} being $+1$ or -1 with equal probability and independently across (i,j). That is $\mathbb{P}\{W_{ij} = +1\} = \mathbb{P}\{W_{ij} = -1\} = 1/2$ for all

$(i,j) \in [n] \times [n]$.

4. Let S be a matrix supported on Ω. The signs of S coincide with W on Ω. This means $[\text{sgn}(S)]_{ij}$ are independent random variables with the following distribution

$$[\text{sgn}(S)]_{ij} = \begin{cases} 1 & \text{with prob.} \quad \frac{\alpha_{ij}\rho_{ij}}{2}, \\ 0 & \text{with prob.} \quad 1 - \alpha_{ij}\rho_{ij}, \\ -1 & \text{with prob.} \quad \frac{\alpha_{ij}\rho_{ij}}{2}. \end{cases} \tag{2.18}$$

We establish the following proposition to capture the central role of dual certificate in the performance guarantee of adaptive PCP.

Proposition 2.11. *Suppose* $\|\mathcal{P}_T\mathcal{R}_\Gamma\| \leq \sqrt{\frac{3}{2p_0}}$ *and* $\mathcal{P}_\Gamma\mathcal{P}_T$ *is injective on* T. *The adaptive PCP program produces a unique solution if there exists a dual certificate* Y *obeying*

$$\mathcal{P}_{\Gamma^c}Y = 0, \tag{2.19}$$

$$\|\mathcal{P}_\Gamma Y \circ \frac{1}{\Lambda}\|_\infty \leq \frac{1}{4}, \tag{2.20}$$

$$\|\mathcal{P}_{T^\perp}(\Lambda \circ sgn(S) + Y)\| \leq \frac{1}{4}, \tag{2.21}$$

$$\|\mathcal{P}_T(Y + \Lambda \circ sgn(S) - UV^*)\|_F \leq \frac{\lambda_{\min}}{n^2} \tag{2.22}$$

where Λ *is a matrix with each entry* $\Lambda_{ij} = \frac{1}{\sqrt{\alpha_{ij}n}\log n}$, $\lambda_{\min} = \frac{1}{\sqrt{n}\log n}$, *and "*$\circ$*" represents entrywise product.*

Proof. See Section 2.6.3. □

Dual Certificate Construction

Proposition 2.11 suggests that it suffices to prove Theorem 2.1 if we find a dual certificate Y that satisfies the dual certificate conditions (2.19)-(2.22). Thus, the second step is to construct Y via the golfing scheme. Although we adapt the steps in [29] to construct the

dual certificate Y, our analysis requires new technical development based on local incoherence parameters.

Note that $\mathbb{P}((i,j) \in \Gamma) = \alpha_{ij}(1-\rho_{ij}) := p_{ij}$. Let $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \cdots \cup \Gamma_l$, and $\{\Gamma_k\}$'s are independent random sets given by

$$\begin{aligned}
\mathbb{P}((i,j) \in \Gamma_1) &= \frac{p_{ij}}{6}, \\
\mathbb{P}((i,j) \in \Gamma_2) &= \frac{p_{ij}}{6}, \\
\mathbb{P}((i,j) \in \Gamma_k) &= q_{ij}, \quad \text{for} \quad k = 3, \cdots, l.
\end{aligned}$$

Thus if $1 - p_{ij} = (1 - \frac{p_{ij}}{6})^2 (1 - q_{ij})^{l-2}$, the sampling strategies are equivalent. Because of the overlap between $\{\Gamma_k\}$, it is clear that $q_{ij} \geq \frac{2p_{ij}}{3(l-2)}$. We set $l = \lfloor 5 \log n + 1 \rfloor$.

We construct a dual certificate Y in the following iterative way:

$$Z_0 = \mathcal{P}_T(UV^* - \Lambda \circ \text{sgn}(S)) \tag{2.23}$$

$$Z_k = (\mathcal{P}_T - \mathcal{P}_T \mathcal{R}_{\Gamma_k} \mathcal{P}_T) Z_{k-1}, \quad for \quad k = 1, \cdots, l \tag{2.24}$$

$$Y = \sum_{k=1}^{l} \mathcal{R}_{\Gamma_k} Z_{k-1}. \tag{2.25}$$

We next show that such constructed Y satisfies the dual certificate conditions by bounding various norms of Z_0 and showing each iteration (2.24) reduces these norms at least by half. Our proof applies the technical lemmas in Section 2.6.1.

Note that we require in the theorem $\sqrt{\alpha_{ij}}(1-\rho_{ij}) \geq C_0 \sqrt{\frac{\mu_{ij} r}{n}} \log n$ and $\alpha_{ij} \geq \frac{\log^2 n}{n}$, which imply

$$p_{ij} = \alpha_{ij}(1-\rho_{ij}) \geq C_0 \sqrt{\frac{\alpha_{ij}\mu_{ij} r}{n}} \log n \geq C_0 \left(\frac{\mu_{ij} r}{n} \vee \frac{\sqrt{\mu_{ij} r}}{n} \right) \log^2 n. \tag{2.26}$$

Thus by ignoring the constant factor, we have

$$\mathbb{P}((i,j) \in \Gamma_1) = \mathbb{P}((i,j) \in \Gamma_2) \geq C_0 \sqrt{\frac{\alpha_{ij}\mu_{ij}r}{n}} \log n, \tag{2.27}$$

$$\mathbb{P}((i,j) \in \Gamma_k) \geq C_0 \sqrt{\frac{\alpha_{ij}\mu_{ij}r}{n}}, \quad \text{for } k = 3, 4, \ldots, l, \tag{2.28}$$

$$\mathbb{P}((i,j) \in \Gamma_k) \geq C_0 \left(\frac{\mu_{ij}r}{n} \vee \frac{\sqrt{\mu_{ij}r}}{n} \right) \log n, \quad \text{for } k = 1, 2, \ldots, l. \tag{2.29}$$

Then following Lemma 2.8 and (2.29), we have

$$\|Z_k\|_F \leq \frac{1}{2} \|Z_{k-1}\|_F, \quad \text{for } k = 1, 2, \ldots, l. \tag{2.30}$$

Recalling the definition of $\|\cdot\|_{w(\infty)}$ in equation (2.13), following Lemma 2.9 and (2.29), we have with high probability

$$\|(I - \mathcal{R}_{\Gamma_k}) Z_{k-1}\| \leq \frac{C}{C_0} \|Z_{k-1}\|_{w(\infty)}, \quad \text{for } k = 1, 2, \ldots, l. \tag{2.31}$$

Due to the fact that $\sqrt{\alpha_{ij}} \geq C_0 \sqrt{\mu_{ij}r/n} \log n$ and (2.27), we further have

$$\mathbb{P}((i,j) \in \Gamma_1) = \mathbb{P}((i,j) \in \Gamma_2) \geq C_0 (\sqrt{\log n})^2 \left(\frac{\mu_{ij}r}{n} \log n \right). \tag{2.32}$$

Thus, following Lemma 2.10, with high probability, we have

$$\|Z_1\|_{w(\infty)} \leq \frac{1}{2\sqrt{\log n}} \|Z_0\|_{w(\infty)}, \tag{2.33}$$

$$\|Z_2\|_{w(\infty)} \leq \frac{1}{2\sqrt{\log n}} \|Z_1\|_{w(\infty)} \leq \frac{1}{2^2 \log n} \|Z_0\|_{w(\infty)}. \tag{2.34}$$

Moreover, following Lemma 2.10 and (2.29), we have

$$\|Z_k\|_{w(\infty)} \leq \frac{1}{2} \|Z_{k-1}\|_{w(\infty)} \leq \frac{1}{2^k \log n} \|Z_0\|_{w(\infty)} \quad \text{for } k = 3, \cdots, l. \tag{2.35}$$

We next bound $\|Z_0\|_F$ and $\|Z_0\|_{w(\infty)}$. Observe that $|(Z_0)_{ab}| \leq |(UV^*)_{ab}| + |[\mathcal{P}_T(\Lambda \circ \text{sgn}(S))]_{ab}|$ and $|(UV^*)_{ab}| = \sqrt{\frac{\mu_{ab} r}{n^2}}$. We need only to bound $|(\mathcal{P}_T \text{sgn}(S))_{ab}|$.

Proposition 2.12. *With the same assumptions on S and Λ as in the problem setup, for any given index (a, b), we have*

$$|[\mathcal{P}_T(\Lambda \circ sgn(S))]_{ab}| \leq \frac{C}{\sqrt{\log n}} \cdot \frac{\sqrt{\mu_{ab} r}}{n} \tag{2.36}$$

with high probability for some constant C.

Proof. See Section 2.6.3. □

By applying triangle inequality, we have

$$\|Z_0\|_F \leq \|UV^*\|_F + \|\mathcal{P}_T(\Lambda \circ \text{sgn}(S))\|_F \leq C\sqrt{\mu r},$$
$$\|Z_0\|_{w(\infty)} \leq \|UV^*\|_{w(\infty)} + \|\mathcal{P}_T(\Lambda \circ \text{sgn}(S))\|_{w(\infty)} \leq C,$$

for some constant C.

Next we verify that the constructed dual certificate Y satisfies the conditions in Proposition 2.11. Obviously, $\mathcal{P}_\Omega Y = 0$. Now we prove Y satisfies the following inequalities with high probability

$$\|\mathcal{P}_T(Y + \Lambda \circ \text{sgn}(S) - UV^*)\|_F \leq \frac{\lambda_{\min}}{n^2}, \tag{2.37}$$
$$\|\mathcal{P}_{T^\perp} Y\| \leq \frac{1}{8}, \tag{2.38}$$
$$\|\mathcal{P}_{T^\perp}(\Lambda \circ \text{sgn}(S))\| \leq \frac{1}{8}, \tag{2.39}$$
$$\|Y \circ \frac{1}{\Lambda}\|_\infty \leq \frac{1}{4}. \tag{2.40}$$

We first show (2.37) as follows.

$$\begin{aligned}
\|\mathcal{P}_T Y + \mathcal{P}_T(\Lambda \circ \mathrm{sgn}(S) - UV^*)\|_F &= \left\| Z_0 - (\sum_{k=1}^{l} \mathcal{P}_T \mathcal{R}_{\Gamma_k} Z_{k-1}) \right\|_F \\
&= \left\| (\mathcal{P}_T - \mathcal{P}_T \mathcal{R}_{\Gamma_1}) Z_0 - (\sum_{k=2}^{l} \mathcal{P}_T \mathcal{R}_{\Gamma_k} Z_{k-1}) \right\|_F = \left\| \mathcal{P}_T Z_1 - (\sum_{k=1}^{l} \mathcal{P}_T \mathcal{R}_{\Gamma_k} Z_{k-1}) \right\|_F \\
&= \cdots \\
&= \|Z_l\|_F \leq \left(\frac{1}{2}\right)^l \cdot \|Z_0\|_F \leq C \left(\frac{1}{2}\right)^l \sqrt{\mu r} \leq \frac{\lambda_{\min}}{n^2}.
\end{aligned}$$

We then show (2.38) as follows.

$$\begin{aligned}
\|\mathcal{P}_{T^\perp} Y\| = \left\| \mathcal{P}_{T^\perp} \sum_{k=1}^{l} \mathcal{R}_{\Gamma_k} Z_{k-1} \right\| &\leq \sum_{k=1}^{l} \|\mathcal{P}_{T^\perp} \mathcal{R}_{\Gamma_k} Z_{k-1}\| = \sum_{k=1}^{l} \|\mathcal{P}_{T^\perp}(\mathcal{R}_{\Gamma_k} Z_{k-1} - Z_{k-1})\| \\
&\leq \sum_{k=1}^{l} \|\mathcal{R}_{\Gamma_k} Z_{k-1} - Z_{k-1}\| \leq \sum_{k=1}^{l} \frac{C}{C_0} \|Z_{k-1}\|_{w(\infty)} \\
&\leq \frac{C}{C_0} \left(1 + \sum_{k=2}^{l} \frac{1}{\sqrt{\log n}} (\frac{1}{2})^{k-1} \right) \|Z_0\|_{w(\infty)} \leq \frac{2C}{C_0} \|Z_0\|_{w(\infty)} \leq \frac{1}{8},
\end{aligned}$$

provided that C_0 is sufficiently large.

We next derive the bound (2.39). It suffices to bound $\|\Lambda \circ \mathcal{P}_\Omega W\|$, where W is defined in Section 2.6.1. Following the way of proving Lemma 2.9, we first show that if $\mathbb{P}((i,j) \in \Omega) = \alpha_{ij}\rho_{ij}$, then

$$\|\Lambda \circ \mathcal{P}_\Omega W - \Lambda \circ \Delta \circ W\| \leq C \frac{\|W\|_\infty}{\sqrt{\log n}}, \tag{2.41}$$

where $\Lambda = [\frac{1}{\sqrt{\alpha_{ij} n \log n}}]$ and $\Delta = [\alpha_{ij}\rho_{ij}]$. Our proof applies the following Latala's Theorem [126].

Theorem 2.13 (Latala's Theorem)**.** *Let A be a random matrix whose entries a_{ij} are inde-*

pendent centered random variables with finite fourth moment. Then

$$\mathrm{E}s_{\max}(A) \le C\left[\max_i\left(\sum_j \mathrm{E}a_{ij}^2\right)^{1/2} + \max_j\left(\sum_i \mathrm{E}a_{ij}^2\right)^{1/2} + \left(\sum_{i,j}\mathrm{E}a_{ij}^4\right)^{1/4}\right].$$

By applying the Latala's Theorem, we obtain the bound

$$\|\Delta\circ\Lambda\circ W\| \le C'\frac{\|W\|_\infty}{\log n}. \tag{2.42}$$

Since $\|W\|_\infty = 1$,

$$\|\Lambda\circ\mathcal{P}_\Omega W\| \le \frac{C}{\sqrt{\log n}} + \frac{C'}{\log n} \le \frac{1}{8}, \tag{2.43}$$

provided n sufficiently large.

To show (2.41), for any matrix Z, we have

$$\Lambda\circ\mathcal{P}_\Omega Z - \Lambda\circ\Delta\circ Z = \sum_{i,j}(\delta_{ij}-\Delta_{ij})\Lambda_{ij}Z_{ij}e_ie_j^* := \sum_{i,j}X_{ij}, \tag{2.44}$$

where X_{ij} are independent zero-mean random matrices. Moreover,

$$\begin{aligned}\|X_{ij}\| &= \|(\delta_{ij}-\Delta_{ij})\Lambda_{ij}Z_{ij}e_ie_j^*\| \\ &\le \frac{\|Z\|_\infty}{\sqrt{n}\log n}\max_{i,j}\left(\frac{\delta_{ij}}{\sqrt{\alpha_{ij}}} - \sqrt{\alpha_{ij}}\rho_{ij}\right) \\ &\le \frac{\|Z\|_\infty}{\sqrt{n}\log n}\max_{i,j}\frac{1}{\sqrt{\alpha_{ij}}} \le \frac{1}{\log^2 n}\|Z\|_\infty,\end{aligned}$$

where the last inequality is due to the assumption $\alpha_{ij} \geq \frac{\log^2 n}{n}$. Furthermore,

$$\begin{aligned}
\left\| \sum_{i,j} \mathrm{E} X_{ij} X_{ij}^* \right\| &= \left\| \sum_{i,j} \mathrm{E}(\delta_{ij} - \Delta_{ij})^2 \Lambda_{ij}^2 Z_{ij}^2 e_i e_i^* \right\| \\
&= \left\| \sum_{i,j} (1 - \Delta_{ij}) \Delta_{ij} \Lambda_{ij}^2 Z_{ij}^2 e_i e_i^* \right\| \\
&\leq \max_i \sum_j (1 - \Delta_{ij}) \Delta_{ij} \Lambda_{ij}^2 Z_{ij}^2 \\
&\leq \frac{\|Z\|_\infty^2}{n \log^2 n} \max_i \sum_j (1 - \alpha_{ij}\rho_{ij}) \rho_{ij} \leq \frac{\|Z\|_\infty^2}{\log^2 n}.
\end{aligned}$$

By noncommutative Bernstein Inequality, we have

$$\|\Lambda \circ \mathcal{P} Z - \Lambda \circ \Delta \circ Z\| \leq C \left(\sqrt{\frac{\|Z\|_\infty^2}{\log^2 n} \cdot \log n} + \frac{\|Z\|_\infty}{\log^2 n} \cdot \log n \right) \leq C \frac{\|Z\|_\infty}{\sqrt{\log n}}.$$

This finishes the proof of (2.41) by letting $Z = W$.

Finally, we show that Y satisfies (2.40) as follows.

$$\begin{aligned}
\left\| Y \circ \frac{1}{\Lambda} \right\|_\infty &= \left\| \frac{1}{\Lambda} \circ \sum_{k=1}^{l} \mathcal{R}_{\Gamma_k} Z_{k-1} \right\|_\infty \\
&\leq \left\| \frac{1}{\Lambda} \circ \mathcal{R}_{\Gamma_1} Z_0 \right\|_\infty + \left\| \frac{1}{\Lambda} \circ \mathcal{R}_{\Gamma_2} Z_1 \right\|_\infty + \sum_{k=3}^{l} \left\| \frac{1}{\Lambda} \circ \mathcal{R}_{\Gamma_k} Z_{k-1} \right\|_\infty \\
&\leq \frac{1}{C_0} \|Z_0\|_{w(\infty)} + \frac{1}{C_0} \|Z_1\|_{w(\infty)} + \sum_{k=3}^{l} \frac{\log n}{C_0} \|Z_{k-1}\|_{w(\infty)} \\
&\leq \frac{1}{C_0} \|Z_0\|_{w(\infty)} + \frac{1}{2C_0\sqrt{\log n}} \|Z_0\|_{w(\infty)} + \sum_{k=3}^{l} \frac{\log n}{C_0} \cdot \left(\frac{1}{2} \right)^{k-1} \frac{1}{\log n} \|Z_0\|_{w(\infty)} \\
&\leq \frac{2}{C_0} \|Z_0\|_{w(\infty)} \leq \frac{2C}{C_0} \leq \frac{1}{4},
\end{aligned}$$

provided that C_0 is sufficiently large, where the second inequality is due to the fact (2.27)-(2.29) and the third inequality is due to the fact (2.33)-(2.35).

2.6.2 Proof of Theorem 2.2

The arguments here adapt elimination and derandomization arguments in [13], and are included here for the completeness. In contrast to the uniform sampling probability ρ, we need to deal with a nonuniform probability matrix ρ.

Elimination Procedure

Definition 2.14. *S' is said to be a* trimmed version *of S if supp$(S') \subset$ supp(S) and $S'_{ij} = S_{ij}$ whenever $S'_{ij} \neq 0$.*

The following theorem states that if PCP correctly recovers the low-rank and sparse components of $M_0 = L_0 + S_0$, it also correctly recovers the components of a matrix $M'_0 = L_0 + S'_0$ where S'_0 is a trimmed version of S_0.

Theorem 2.15 (**Theorem 2.2 in [13]**). *Suppose the solution to* (2.12) *with input data $M_0 = L_0 + S_0$ is unique and exact, and consider $M'_0 = L_0 + S'_0$, where S'_0 is a trimmed version of S_0. Then the solution to* (2.12) *with input M'_0 is exact as well.*

Derandomization Procedure

Let $\boldsymbol{\rho}$ be the matrix with each (i, j)-entry being ρ_{ij}. If PCP yields exact recovery with a certain probability for the random sign model with the parameter $2\boldsymbol{\rho}$, then it also yields exact recovery with at least the same probability for the fixed sign model with locations of non-zero entries sampled using Bernoulli model with the parameter $\boldsymbol{\rho}$.

Theorem 2.16. *Suppose that L_0 is a low-rank matrix with local coherence parameter $[\mu_{ij}]$ and S_0 follow the Bernoulli model with parameter $2\boldsymbol{\rho}$, and the signs of S_0 are independently distributed ± 1 as stated in* (2.10) *(and independent from the locations). Then, if the PCP solution is exact with high probability, then it is also exact with at least the same probability for the model in which the signs are fixed and the locations are sampled from the Bernoulli model with parameter $\boldsymbol{\rho}$.*

Proof. Consider the model with fixed-sign assumption. We view S_0 as $\mathcal{P}_\Omega S$ for some fixed matrix S, where Ω is sampled from the Bernoulli model with parameter $\boldsymbol{\rho}$. Therefore, S_0 has following distribution

$$(S_0)_{ij} = \begin{cases} S_{ij}, & \text{w. p. } \rho_{ij} \\ 0, & \text{w. p. } 1-\rho_{ij}. \end{cases}$$

Now consider a random sign matrix with each entry distributed independently as follows

$$E_{ij} = \begin{cases} 1, & \text{w. p. } \rho_{ij}, \\ 0, & \text{w. p. } 1-2\rho_{ij}, \\ -1, & \text{w. p. } \rho_{ij}, \end{cases}$$

and an "elimination" matrix η with entries defined by

$$\eta_{ij} = \begin{cases} 0, & \text{if } E_{ij}[\text{sgn}(S)]_{ij} = -1, \\ 1, & \text{otherwise.} \end{cases}$$

The entries of η are independent since they are functions of independent random variables.

Consider now $S_0' = \eta \circ (|S| \circ E)$, where $\circ$ denotes the componentwise product so that $[S_0']_{ij} = \eta_{ij} \circ (|S_{ij}| \circ E_{ij})$. Then, we claim that S_0' and S_0 have the same distribution. By independence of each entry, it suffices to check that their marginals match each other. For $S_{ij} \neq 0$, we have

$$\begin{aligned} \mathbb{P}([S_0']_{ij} = S_{ij}) &= \mathbb{P}(\eta_{ij} = 1 \text{ and } E_{ij} = [\text{sgn}(S)]_{ij}) \\ &= \mathbb{P}(E_{ij}[\text{sgn}(S)]_{ij} \neq -1 \text{ and } E_{ij} = [\text{sgn}(S)]_{ij}) \\ &= \mathbb{P}(E_{ij} = [\text{sgn}(S)]_{ij}) = \rho_{ij}, \end{aligned}$$

which establishes the claim.

Now, $|S| \circ E$ obeys the random sign model, and by assumption, PCP recovers $|S| \circ E$ with high probability. By the elimination procedure, PCP also recovers $S_0' = \eta \circ (|S| \circ E)$. Since S_0' and S_0 have the same distribution, the theorem follows. □

2.6.3 Proofs of Technical Lemmas

In this section, we prove the key lemmas provided in Section 2.6.1. The central technique used here is non-communicative Bernstein inequality [128].

Proof of Lemma 2.6

The proof of Lemma 2.6 follows the proof of [22, Lemma 9].

Proof of Lemma 2.7

We note that the condition $\|\mathcal{P}_T - \mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T\| \leq \frac{1}{2}$ implies that for any matrix Z

$$\frac{1}{2}\|\mathcal{P}_T Z\|_F \leq \|\mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T(Z)\|_F \leq \frac{3}{2}\|\mathcal{P}_T Z\|_F.$$

Thus, for any matrix Z, we have

$$\begin{aligned}
\left\|\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T(Z)\right\|_F^2 &= \langle \mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T(Z), \mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T(Z)\rangle \\
&= \langle Z, (\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T)^*\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T(Z)\rangle \\
&= \langle \mathcal{P}_T(Z), \mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T(Z)\rangle \\
&\leq \|\mathcal{P}_T Z\|_F \|\mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T(Z)\|_F \\
&\leq \frac{3}{2}\|\mathcal{P}_T Z\|_F^2.
\end{aligned}$$

Thus, $\left\|\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T\right\| \leq \sqrt{3/2}$ and hence $\left\|\mathcal{P}_T\mathcal{R}_{\Omega_0}^{1/2}\right\| \leq \sqrt{3/2}$ because $\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T$ and $\mathcal{P}_T\mathcal{R}_{\Omega_0}^{1/2}$

are adjoint operators. On the other hand, we show $\left\|\mathcal{R}_{\Omega_0}^{1/2}\right\| \leq 1/\sqrt{p_0}$. For any matrix Z,

$$\left\|\mathcal{R}_{\Omega_0}^{1/2}(Z)\right\|_F^2 = \left\|\sum_{i,j} \frac{1}{\sqrt{q_{ij}}} \mathbb{I}_{\{(i,j)\in\Omega_0\}} Z_{ij} e_i e_j^*\right\|_F^2 \leq \sum_{i,j} \frac{Z_{ij}^2}{q_{ij}} \leq \frac{1}{p_0}\|Z\|_F^2.$$

Thus, $\|\mathcal{R}_{\Omega_0}\mathcal{P}_T\| \leq \left\|\mathcal{R}_{\Omega_0}^{1/2}\right\| \cdot \left\|\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T\right\| \leq \sqrt{\frac{3}{2p_0}}$. Thus, $\|\mathcal{P}_T\mathcal{R}_{\Omega_0}\| \leq \sqrt{\frac{3}{2p_0}}$.

Since we have $\frac{1}{2}\|\mathcal{P}_T Z\|_F \leq \|\mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T(Z)\|_F \leq \frac{3}{2}\|\mathcal{P}_T Z\|_F$ for any matrix $Z \in T$, the operator $\mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T$ mapping T onto itself is well conditioned. Thus, $\mathcal{P}_{\Omega_0}\mathcal{P}_T$ is injective on T, i.e., for $Z \in T$, $\mathcal{P}_{\Omega_0}\mathcal{P}_T(Z) = 0$ if and only if $Z = 0$.

Proof of Lemma 2.8

This is a direct result of Lemma 2.6.

Proof of Lemma 2.9

Let δ_{ij} denote the Bernoulli random variable $\mathbb{I}((i,j) \in \Omega_0)$. We can derive

$$(\mathcal{R}_{\Omega_0} - I)Z = \sum_{i,j}\left(\frac{1}{q_{ij}}\delta_{ij} - 1\right)\langle e_i e_j^*, Z\rangle e_i e_j^* =: \sum_{i,j} X_{ij}.$$

We note that X_{ij} for all $i, j \in [n]$ are zero-mean independent random matrices. Furthermore,

$$\|X_{ij}\| \leq \frac{1}{q_{ij}}|Z_{ij}| \leq \frac{1}{C_0 \log n}\|Z\|_{w(\infty)},$$

and

$$\begin{aligned}\left\|\sum_{i,j} \mathrm{E}X_{ij}X_{ij}^*\right\| &= \left\|\sum_{i,j} \mathrm{E}\left(\frac{1}{q_{ij}}\delta_{ij} - 1\right)^2 Z_{ij}^2 e_i e_i^*\right\| = \left\|\sum_{i,j}\left(\frac{1}{q_{ij}} - 1\right) Z_{ij}^2 e_i e_i^*\right\| \\ &\leq \max_i \sum_j \frac{Z_{ij}^2}{q_{ij}} \leq n\|Z\|_{w(\infty)}^2 \cdot \max_{i,j}\frac{w_{ij}^2}{q_{ij}} \leq \frac{1}{C_0 \log n}\|Z\|_{w(\infty)}^2.\end{aligned}$$

Similarly, it can be shown that $\|\sum_{i,j} \mathrm{E}X_{ij}^* X_{ij}\| \leq \frac{1}{C_0 \log n}\|Z\|_{w(\infty)}^2$. Thus, applying the

non-commutative Bernstein inequality, we obtain with high probability,

$$\|(\mathcal{R}_{\Omega_0} - I)Z\| = \left\|\sum_{i,j} X_{ij}\right\| \leq C\left(\sqrt{\frac{1}{C_0 \log n}\|Z\|^2_{w(\infty)} \cdot \log n} + \frac{1}{C_0 \log n}\|Z\|_{w(\infty)} \cdot \log n\right)$$
$$\leq \frac{C}{\sqrt{C_0}}\|Z\|_{w(\infty)}.$$

Proof of Lemma 2.10

For any entry index pair (a, b), we have

$$[(\mathcal{P}_T\mathcal{R}_{\Omega_0} - \mathcal{P}_T)Z]_{ab} \cdot \frac{1}{w_{ab}} = \sum_{i,j}\left(\frac{1}{q_{ij}}\delta_{ij} - 1\right) Z_{ij}\langle \mathcal{P}_T(e_i e_j^*), e_a e_b^*\rangle \cdot \frac{1}{w_{ab}}$$
$$= \sum_{i,j}\left(\frac{1}{q_{ij}}\delta_{ij} - 1\right) Z_{ij}\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*)\rangle \cdot \frac{1}{w_{ab}} =: \sum_{i,j} x_{ij}.$$

We note that x_{ij} for $i, j \in [n]$ are independent random variables and $\mathrm{E}x_{ij} = 0$. Furthermore,

$$|x_{ij}| \leq \frac{1}{q_{ij}}|Z_{ij}| \cdot |\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*)\rangle| \cdot \frac{1}{w_{ab}}$$
$$\leq |Z_{ij}| \cdot \frac{1}{C_0\beta^{-2}(\frac{\mu_{ij}r}{n})\log n} \cdot \sqrt{\frac{2\mu_{ij}r}{n}} \cdot \sqrt{\frac{2\mu_{ab}r}{n}} \cdot \frac{1}{\sqrt{\frac{\mu_{ab}r}{n^2}}}$$
$$\leq \frac{2\beta^2}{C_0 \log n}\frac{|Z_{ij}|}{w_{ij}} \leq \frac{2\beta^2}{C_0 \log n}\|Z\|_{w_{(\infty)}},$$

and

$$\begin{aligned}\left|\sum_{i,j}\mathrm{E}x_{ij}^2\right| &\le \sum_{i,j}\mathrm{E}\left(\frac{1}{q_{ij}}\delta_{ij}-1\right)^2 Z_{ij}^2\cdot|\langle e_ie_j^*,\mathcal{P}_T(e_ae_b^*)\rangle|^2\cdot\frac{1}{w_{ab}^2}\\ &\le \sum_{i,j}\left(\frac{1}{q_{ij}}-1\right)\frac{Z_{ij}^2}{w_{ij}^2}\cdot\frac{w_{ij}^2}{w_{ab}^2}\cdot|\langle e_ie_j^*,\mathcal{P}_T(e_ae_b^*)\rangle|^2\\ &\le \frac{1}{C_0\beta^{-2}(\frac{\log n}{n})}\cdot\|Z\|_{w_{(\infty)}}^2\cdot\frac{1}{\mu_{ab}r}\|\mathcal{P}_T(e_ae_b^*)\|_F^2\\ &\le \frac{2\beta^2}{C_0\log n}\cdot\|Z\|_{w_{(\infty)}}^2,\end{aligned}$$

where we use the assumption $q_{ij}\ge C_0\beta^{-2}(\frac{\mu_{ij}r}{n})\log n$ and the fact $\|\mathcal{P}_T(e_ae_b^*)\|_F^2\le\frac{2\mu_{ab}r}{n}$.

Thus, applying the non-commutative Bernstein inequality, we have

$$\begin{aligned}\left|\sum_{i,j}x_{ij}\right| &\le C\left(\sqrt{\frac{2\beta^2}{C_0\log n}\|Z\|_{w_{(\infty)}}^2\cdot\log n}+\frac{2\beta^2}{C_0\log n}\|Z\|_{w_{(\infty)}}\cdot\log n\right)\\ &= C\left(\sqrt{\frac{2}{C_0}}\beta+\frac{2}{C_0}\beta^2\right)\|Z\|_{w_{(\infty)}}\le\frac{1}{2}\beta\|Z\|_{w_{(\infty)}},\end{aligned}$$

with high probability, provided that C_0 is sufficiently large.

Proof of Proposition 2.11

Suppose the adaptive PCP yields a solution $(\hat{L},\hat{S})$. Assume $\hat{L}=L+H$. Then we have $\mathcal{P}_O(H)=S-\hat{S}$ because of the relationship $\mathcal{P}_O(L)+S=\mathcal{P}_O(\hat{L})+\hat{S}$. It is clear that $\hat{S}$ is supported on O because S is supported on $\Omega\subset O$. By the definition of $(\hat{L},\hat{S})$, we have

$$\|\hat{L}\|_*+\|\Lambda\circ\hat{S}\|_1\le\|L\|_*+\|\Lambda\circ S\|_1. \tag{2.45}$$

By the definition of subgradient, we have

$$\|L+H\|_*\ge\|L\|_*+\langle\mathcal{P}_TH,UV^*\rangle+\|\mathcal{P}_{T^\perp}H\|_*$$

because we can always find a $W \in T^{\perp}$ and $\|W\| \leq 1$ such that $\|\mathcal{P}_{T^{\perp}} H\|_* = \langle \mathcal{P}_{T^{\perp}} H, W \rangle$.

Thus we obtain

$$\|\Lambda \circ S\|_1 - \|\Lambda \circ \hat{S}\|_1 \geq \langle \mathcal{P}_T H, UV^* \rangle + \|\mathcal{P}_{T^{\perp}} H\|_*,$$

which implies

$$\|\Lambda \circ S\|_1 - \|\Lambda \circ \mathcal{P}_{\Omega}(\hat{S})\|_1 \geq \langle H, UV^* \rangle + \|\mathcal{P}_{T^{\perp}} H\|_* + \|\Lambda \circ \mathcal{P}_{\Gamma}(\hat{S})\|_1,$$

because $\hat{S} = \mathcal{P}_{\Omega}(\hat{S}) + \mathcal{P}_{\Gamma}(\hat{S})$. Furthermore,

$$\begin{aligned} \|\Lambda \circ \mathcal{P}_{\Omega}(\hat{S})\|_1 &= \|\Lambda \circ (S + \mathcal{P}_{\Omega}(-H))\|_1 \\ &\geq \|\Lambda \circ S\|_1 + \langle \operatorname{sgn}(\Lambda \circ S), \Lambda \circ \mathcal{P}_{\Omega}(-H) \rangle \\ &= \|\Lambda \circ S\|_1 + \langle \Lambda \circ \operatorname{sgn}(S), -H \rangle . \end{aligned}$$

Combining the last two inequalities and using the fact $\mathcal{P}_{\Gamma} \hat{S} = \mathcal{P}_{\Gamma}(\hat{S} - S) = -\mathcal{P}_{\Gamma} H$, we have

$$\|\mathcal{P}_{T^{\perp}} H\|_* + \|\Lambda \circ \mathcal{P}_{\Gamma} H\|_1 \leq \langle H, \Lambda \circ \operatorname{sgn}(S) - UV^* \rangle .$$

By introducing a matrix Y which obeys the conditions in the theorem, we show that

$$\begin{aligned} &\langle H, \Lambda \circ \operatorname{sgn}(S) - UV^* \rangle = \langle H, Y + \Lambda \circ \operatorname{sgn}(S) - UV^* \rangle - \langle H, Y \rangle \\ &= \langle \mathcal{P}_T H, \mathcal{P}_T (Y + \Lambda \circ \operatorname{sgn}(S) - UV^*) \rangle + \langle \mathcal{P}_{T^{\perp}} H, \mathcal{P}_{T^{\perp}} (Y + \Lambda \circ \operatorname{sgn}(S)) \rangle \\ &\quad - \langle \mathcal{P}_{\Gamma} H, \mathcal{P}_{\Gamma} Y \rangle - \langle \mathcal{P}_{\Gamma^c} H, \mathcal{P}_{\Gamma^c} Y \rangle \\ &\leq \frac{\lambda_{\min}}{n^2} \|\mathcal{P}_T H\|_F + \frac{1}{4} \|\mathcal{P}_{T^{\perp}} H\|_* + \frac{1}{4} \|\Lambda \circ \mathcal{P}_{\Gamma} H\|_1. \end{aligned}$$

This implies

$$\frac{3}{4}\|\mathcal{P}_{T^\perp}H\|_* + \frac{3}{4}\|\Lambda \circ \mathcal{P}_\Gamma H\|_1 \leq \frac{\lambda_{\min}}{n^2}\|\mathcal{P}_T H\|_F.$$

Thus

$$\frac{3}{4}\|\mathcal{P}_{T^\perp}H\|_* + \frac{3\lambda_{\min}}{4}\|\mathcal{P}_\Gamma H\|_1 \leq \frac{\lambda_{\min}}{n^2}\|\mathcal{P}_T H\|_F.$$

Next we bound $\|\mathcal{P}_T H\|_F$ as follows:

$$\begin{aligned}
\|\mathcal{P}_T H\|_F &\leq 2\|\mathcal{P}_T\mathcal{R}_\Gamma\mathcal{P}_T(H)\|_F && \text{(Lemma 1)}\\
&\leq 2\|\mathcal{P}_T\mathcal{R}_\Gamma\mathcal{P}_{T^\perp}(H)\|_F + 2\|\mathcal{P}_T\mathcal{R}_\Gamma(H)\|_F && \text{(triangle inequality)}\\
&\leq \sqrt{\frac{6}{p_0}}\|\mathcal{P}_{T^\perp}(H)\|_F + \sqrt{\frac{6}{p_0}}\|\mathcal{P}_\Gamma(H)\|_F. && \text{(Lemma 2)}
\end{aligned}$$

Since $\|\cdot\|_F \leq \|\cdot\|_*$ and $\|\cdot\|_F \leq \|\cdot\|_1$, we have

$$\left(\frac{3}{4} - \frac{\lambda_{\min}}{n^2}\sqrt{\frac{6}{p_0}}\right)\|\mathcal{P}_{T^\perp}(H)\|_F + \left(\frac{3}{4}\lambda_{\min} - \frac{\lambda_{\min}}{n^2}\sqrt{\frac{6}{p_0}}\right)\|\mathcal{P}_\Gamma(H)\|_F \leq 0.$$

Given $p_0 \geq \frac{\log^2 n}{n^4}$, the above inequality implies $\mathcal{P}_{T^\perp}H = \mathcal{P}_\Gamma H = 0$, which further yields $\mathcal{P}_\Gamma\mathcal{P}_T(H) = 0$. Since $\mathcal{P}_\Gamma\mathcal{P}_T$ is injective on T, we have $\mathcal{P}_T H = 0$. Consequently, $H = 0$, and thus the solution is unique.

Proof of Proposition 2.12

We first observe that

$$\langle e_a e_b^*, \mathcal{P}_T(\Lambda \circ \operatorname{sgn}(S))\rangle = \sum_{i,j}\frac{\delta_{ij}}{\sqrt{\alpha_{ij}n}\log n}\left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*)\right\rangle =: \sum_{i,j} x_{ij}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{with prob.} \quad \alpha_{ij}\rho_{ij}/2 \\ 0 & \text{with prob.} \quad 1-\alpha_{ij}\rho_{ij} \\ -1 & \text{with prob.} \quad \alpha_{ij}\rho_{ij}/2. \end{cases}$$

Thus $\{x_{ij}\}$ are independent random variables and $\mathrm{E}x_{ij} = 0$. Furthermore,

$$|x_{ij}| \le \left| \frac{\delta_{ij}}{\sqrt{\alpha_{ij} n} \log n} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*) \right\rangle \right| \le \frac{1}{\sqrt{\alpha_{ij} n} \log n} \sqrt{\frac{2\mu_{ij} r}{n}} \cdot \sqrt{\frac{2\mu_{ab} r}{n}} \le \frac{2\sqrt{\mu_{ab} r}}{C_0 n \log^2 n},$$

where the last inequality is due to the assumption $\sqrt{\alpha_{ij}} \ge C_0 \sqrt{\mu_{ij} r/n} \log n$, and

$$\begin{aligned} \left| \sum_{i,j} \mathrm{E}x_{ij}^2 \right| &= \left| \sum_{i,j} \frac{\mathrm{E}\delta_{ij}^2}{\alpha_{ij} n \log^2 n} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*) \right\rangle^2 \right| = \left| \sum_{i,j} \frac{\rho_{ij}}{n \log^2 n} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*) \right\rangle^2 \right| \\ &\le \frac{1}{n \log^2 n} \left| \sum_{i,j} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*) \right\rangle^2 \right| = \frac{1}{n \log^2 n} \|\mathcal{P}_T(e_a e_b^*)\|_F^2 \le \frac{2\mu_{ab} r}{n^2 \log^2 n}. \end{aligned}$$

Thus applying the non-commutative Bernstein inequality, we obtain

$$\left| \sum_{i,j} x_{ij} \right| \le C \left(\sqrt{\frac{2\mu_{ab} r}{n^2 \log^2 n} \cdot \log n} + \frac{2\sqrt{\mu_{ab} r}}{C_0 n \log^2 n} \cdot \log n \right) \le \frac{2C}{\sqrt{\log n}} \frac{\sqrt{\mu_{ab} r}}{n},$$

where the last inequality follows that C_0 is sufficiently large.

CHAPTER 3
FAST PHASE RETRIEVAL: RESHAPED WIRTINGER FLOW

In this chapter, we present a novel algorithm *reshaped Wirtinger flow* (RWF) to solve the phase retrieval problem. We first formulate the phase retrieval problem in Section 3.1. We then present the RWF algorithm in Section 3.2. We establish the performance guarantee of RWF in Section 3.3. We introduce a stochastic algorithm IRWF and establish its performance guarantee in Section 3.4. We describe the numerical experiments in Section 3.5 and summarize this chapter in Section 3.6. At last, we provide technical proofs in Section 3.7.

Throughout this chapter, boldface lowercase letters such as $\boldsymbol{a}_i, \boldsymbol{x}, \boldsymbol{z}$ denote vectors, and boldface capital letters such as $\boldsymbol{A}, \boldsymbol{Y}$ denote matrices. For two matrices, $\boldsymbol{A} \preceq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive definite. For a complex matrix or vector, $\boldsymbol{A}^*$ and $\boldsymbol{z}^*$ denote conjugate transposes of $\boldsymbol{A}$ and $\boldsymbol{z}$ respectively. For a real matrix or vector, $\boldsymbol{A}^T$ and $\boldsymbol{z}^T$ denote transposes of $\boldsymbol{A}$ and $\boldsymbol{z}$ respectively. The indicator function $\mathbf{1}_A = 1$ if the event A is true, and $\mathbf{1}_A = 0$ otherwise.

3.1 Problem Formulation

As discussed in the Chapter 1, mathematically, phase retrieval amounts to recovering the signal from only measurements of magnitudes. Since our loss function is based on $|\boldsymbol{a}_i^T\boldsymbol{z}|$ rather than $|\boldsymbol{a}_i^T\boldsymbol{z}|^2$ in WF. For simplicity of notation, we adopt the following problem formulation of recovering a signal $\boldsymbol{x} \in \mathbb{R}^n$ based on m measurements y_i given by

$$y_i = |\langle \boldsymbol{a}_i, \mathbf{x}\rangle|, \quad \text{for } i = 1, \cdots, m, \tag{3.1}$$

where $\boldsymbol{a}_i \in \mathbb{R}^n$ for $i = 1, \cdots, m$ are known measurement vectors, independently generated by Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_{n\times n})$.

We note that the *Wirtinger flow* (WF) algorithm [1] guarantees signal recovery with $\mathcal{O}(n \log n)$ Gaussian measurements and attains $\epsilon-$accuracy within $\mathcal{O}(mn^2 \log 1/\epsilon)$ flops, via minimizing the following nonconvex loss function

$$\ell_{WF}(\boldsymbol{z}) := \frac{1}{4m}\sum_{i=1}^{m}(|\boldsymbol{a}_i^T\boldsymbol{z}|^2 - y_i^2)^2. \tag{3.2}$$

The WF algorithm uses the quadratic loss of $|\boldsymbol{a}_i^T\boldsymbol{z}|^2$ so that the optimization objective is a *smooth* function of $\boldsymbol{a}_i^T\boldsymbol{z}$ and the gradient step becomes simple. However, this comes with a cost of increasing the order of $\boldsymbol{a}_i^T\boldsymbol{z}$ to be four in the loss function. We adopt the quadratic loss of $|\boldsymbol{a}_i^T\boldsymbol{z}|$ as follows[1],

$$\ell(\boldsymbol{z}) := \frac{1}{2m}\sum_{i=1}^{m}\left(|\boldsymbol{a}_i^T\boldsymbol{z}| - y_i\right)^2. \tag{3.3}$$

Although the above loss function (3.3) is not smooth everywhere, it reduces the order of $\boldsymbol{a}_i^T\boldsymbol{z}$ to be two, and the general curvature can be more amenable to the convergence of the

[1]The loss function (3.3) was also used in [53] to derive a gradient-like update for the phase retrieval problem with Fourier magnitude measurements. However, the focus of this chapter is to characterize global convergence guarantee for such an algorithm with appropriate initialization, which was not studied in [53].

gradient method. For such a nonconvex and nonsmooth loss function, we develop a gradient descent-like algorithm, which sets zero for the "gradient" component corresponding to nonsmooth samples.

We show that the lower-order loss function has great advantage in both statistical and computational efficiency, although scarifying smoothness. In fact, the curvature of such a loss function behaves similarly to that of a least-squares loss function in the neighborhood of global optimums (see Section 3.3), and hence RWF converges fast. The nonsmoothness does not significantly affect the convergence of the algorithm because only with negligible probability the algorithm encounters nonsmooth points for some samples, which furthermore are set not to contribute to the gradient direction by the algorithm.

3.2 Reshaped Wirtinger Flow

It can be observed that if $\boldsymbol{z}$ is a solution, i.e., satisfying (3.1), then $\boldsymbol{z}e^{-j\phi}$ is also the solution of the problem. Thus, the recovery is up to a phase difference. We define the Euclidean distance between two vectors up to a global phase difference [1] as, for complex signals,

$$\text{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min_{\phi \in [0, 2\pi)} \|\boldsymbol{z}e^{-j\phi} - \boldsymbol{x}\|, \tag{3.4}$$

where it is simply $\min \|\boldsymbol{z} \pm \boldsymbol{x}\|$ for real case. We focus on the real-valued case in analysis, but the algorithm designed below is applicable to the complex-valued case and the case of *coded diffraction pattern* (CDP) as we demonstrate numerically in Section 3.5.

We design RWF (see Algorithm 1) for solving the above problem, which contains two stages: spectral initialization and gradient loop. Suggested values for parameters are $\alpha_l = 1, \alpha_u = 5$ and $\mu = 0.8$[2]. The scaling parameter in λ_0 and the conjugate transpose $\boldsymbol{a}_i^*$ allow the algorithm readily applicable to complex and CDP cases. We next describe the two stages of the algorithm in detail in Sections 3.2.1 and 3.2.2, respectively.

[2]For complex Gaussian case, we suggest $\mu = 1.2$.

Algorithm 1 Reshaped Wirtinger Flow

Input: $\boldsymbol{y} = \{y_i\}_{i=1}^m$, $\{\boldsymbol{a}_i\}_{i=1}^m$;

Parameters: Lower and upper thresholds α_l, α_u for truncation in initialization, stepsize μ;

Initialization: Let $\boldsymbol{z}^{(0)} = \lambda_0 \tilde{\boldsymbol{z}}$, where $\lambda_0 = \frac{mn}{\sum_{i=1}^m \|\boldsymbol{a}_i\|_1} \cdot \left(\frac{1}{m}\sum_{i=1}^m y_i\right)$ and $\tilde{\boldsymbol{z}}$ is the leading eigenvector of

$$\boldsymbol{Y} := \frac{1}{m}\sum_{i=1}^m y_i \boldsymbol{a}_i \boldsymbol{a}_i^* \mathbf{1}_{\{\alpha_l \lambda_0 < y_i < \alpha_u \lambda_0\}}. \tag{3.5}$$

Gradient loop: for $t = 0 : T-1$ do

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^* \boldsymbol{z}^{(t)} - y_i \cdot \frac{\boldsymbol{a}_i^* \boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^* \boldsymbol{z}^{(t)}|}\right) \boldsymbol{a}_i. \tag{3.6}$$

Output $\boldsymbol{z}^{(T)}$.

3.2.1 Initialization via Spectral Method

Differently from the spectral initialization methods for WF in [1] and for TWF in [61], both of which are based on $|\boldsymbol{a}_i^* \boldsymbol{x}|^2$, we propose an alternative initialization in Algorithm 1 that uses magnitude $|\boldsymbol{a}_i^* \boldsymbol{x}|$ instead, and truncates samples with both lower and upper thresholds as in (3.5). We show that such initialization achieves smaller sample complexity than WF and the same order-level sample complexity as TWF, and furthermore, performs better than both WF and TWF numerically.

Our initialization consists of estimation of both the norm and direction of $\boldsymbol{x}$. The norm estimation of $\boldsymbol{x}$ is given by λ_0 in Algorithm 1. Intuitively, with real Gaussian measurements, the scaling coefficient $\frac{mn}{\sum_{i=1}^m \|\boldsymbol{a}_i\|_1} \approx \sqrt{\frac{\pi}{2}}$. Moreover, $y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|$ are independent sub-Gaussian random variables for $i = 1, \ldots, m$ with mean $\sqrt{\frac{2}{\pi}}\|\boldsymbol{x}\|$, and thus $\frac{1}{m}\sum_{i=1}^m y_i \approx \sqrt{\frac{2}{\pi}}\|\boldsymbol{x}\|$. Combining these two facts yields the desired argument.

The direction of $\boldsymbol{x}$ is approximated by the leading eigenvector of $\boldsymbol{Y}$, because $\boldsymbol{Y}$ approaches $\mathrm{E}[\boldsymbol{Y}]$ by concentration of measure and the leading eigenvector of $\mathrm{E}[\boldsymbol{Y}]$ takes the form $c\boldsymbol{x}$ for some scalar $c \in \mathbb{R}$. We note that (3.5) involves truncation of samples from both sides, in contrast to truncation only by an upper threshold in [61]. This difference is due to the following reason. We note that in high dimension setting, two random vectors are

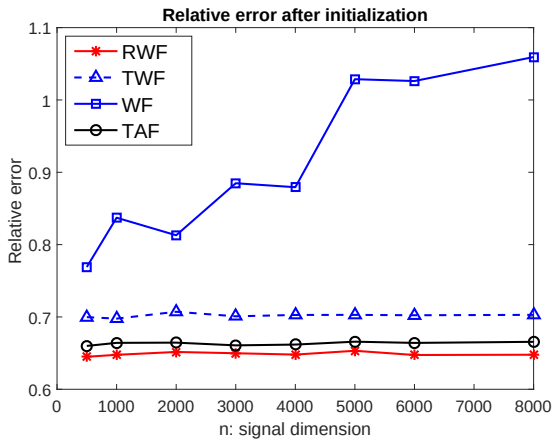


Fig. 3.1: Comparison of different initialization methods with $m = 6n$ and 50 iterations.

almost perpendicular to each other [129], which indicates there are considerable amount of $|\boldsymbol{a}_i^*\boldsymbol{x}|$ with small values, i.e., less than 1. These samples with small values deviate the direction of leading eigenvector of $\boldsymbol{Y}$ from $\boldsymbol{x}$, whose effect cannot be offset and neglected if there are only moderate number of samples. Specifically, [61] uses $y' = |\boldsymbol{a}_i^*\boldsymbol{x}|^2$ to weight the contribution of $\boldsymbol{a}_i\boldsymbol{a}_i^*$ in $\boldsymbol{Y}$ and the square power helps to reduce the contribution of bad directions (that samples with small $|\boldsymbol{a}_i^*\boldsymbol{x}|$ values). In contrast, we use $y_i = |\boldsymbol{a}_i^*\boldsymbol{x}|$ to weight the contribution of $\boldsymbol{a}_i\boldsymbol{a}_i^*$ and apply truncation from bellow to filter out bad directions directly.

We next provide the formal statement of the performance guarantee for the initialization step that we propose.

Proposition 3.1. *Fix $\delta > 0$. The initialization step in Algorithm 1 yields $\boldsymbol{z}^{(0)}$ satisfying dist$(\boldsymbol{z}^{(0)}, \boldsymbol{x}) \leq \delta\|\boldsymbol{x}\|$ with probability at least $1 - \exp(-cm\epsilon^2)$, if $m > C(\delta, \epsilon)n$, where c is some positive constant and C is a positive number only affected by δ and ϵ.*

Finally, Figure 3.1 demonstrates that RWF achieves better initialization accuracy in terms of the relative error $\frac{\text{dist}(\boldsymbol{z}^{(0)},\boldsymbol{x})}{\|\boldsymbol{x}\|}$ than WF and TWF. Furthermore, we also include the orthonormal promoting initialization method proposed for truncated amplitude flow (TAF) in the independent work [70], in the comparison. It can be seen that our initialization is slightly better.

3.2.2 Gradient Loop

The gradient loop of Algorithm 1 is based on the loss function (3.3). The update direction will be as follows:

$$\nabla \ell(\boldsymbol{z}) := \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z} - y_i \cdot \mathrm{sgn}(\boldsymbol{a}_i^T\boldsymbol{z})\right)\boldsymbol{a}_i = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z} - y_i \cdot \frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i, \quad (3.7)$$

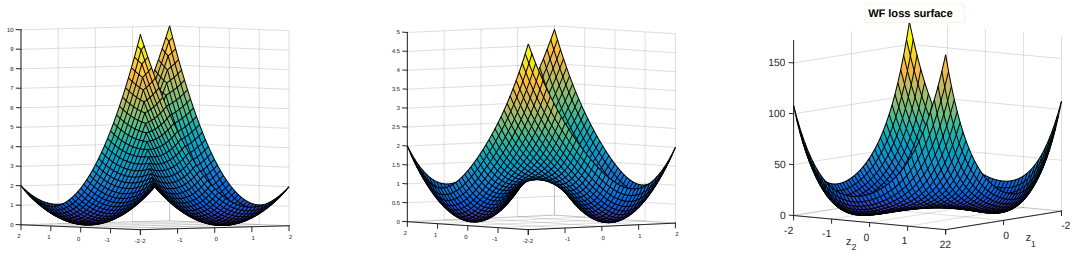
where $\mathrm{sgn}(\cdot)$ is the sign function for nonzero arguments. We further set $\mathrm{sgn}(0) = 0$ and $\frac{0}{|0|} = 0$. In fact, $\nabla \ell(\boldsymbol{z})$ equals the gradient of the loss function (3.3) if $\boldsymbol{a}_i^T\boldsymbol{z} \neq 0$ for all $i = 1, ..., m$. For samples with nonsmooth point, i.e., $\boldsymbol{a}_i^T\boldsymbol{z} = 0$, we adopt Fréchet superdifferential [130] for nonconvex function to set the corresponding gradient component to be zero (as zero is an element in Fréchet superdifferential). With abuse of terminology, we still refer to $\nabla \ell(\boldsymbol{z})$ in (3.7) as "gradient" for simplicity, which rather represents the update direction in the gradient loop of Algorithm 1.

3.3 Geometric Convergence of RWF

Before we present the theoretical analysis, let us first understand why RWF is fast intuitively. We argue that the curvature of loss function (3.3) behaves similarly to that of a least-squares problem with phase information in the neighborhood of global optimizers, and hence yields faster convergence. To provide further insights, consider the standard problem of solving $\boldsymbol{x}$ from *linear measurements* $\langle \boldsymbol{a}_i, \mathbf{x}\rangle$, $i = 1, \cdots, m$, where $\boldsymbol{a}_i$'s are composed of i.i.d. standard Gaussian entries. In this case, it is natural to use the least-squares loss function

$$\ell_{LS}(\boldsymbol{z}) := \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z} - \boldsymbol{a}_i^T\boldsymbol{x}\right)^2. \quad (3.8)$$

Examining the expected (with respect to $\boldsymbol{a}_i$'s) loss surface of $\min\{\ell_{LS}(\boldsymbol{z}), \ell_{LS}(-\boldsymbol{z})\}$ (to mimic sign ambiguity), $\ell(\boldsymbol{z})$, and $\ell_{WF}(\boldsymbol{z})$ in Figure 3.2, whose expressions can be found in Section 3.7.2, it can be seen that the loss of RWF, rather than the loss of WF, has a similar curvature to the quadratic least-squares loss around the global optimizers, which justifies its better performance than WF.



(a) Expected loss of LS (b) Expected loss of RWF (c) Expected loss of WF

Fig. 3.2: Surface of the expected loss function of (a) least-squares (mirrored symmetrically), (b) RWF, and (c) WF when $\boldsymbol{x} = [1, -1]^T$.

The nonsmoothness of the loss function (3.3) does not negatively impact the performance of RWF because only with negligible probability the algorithm encounters nonsmooth points for some samples, which furthermore are set not to contribute to the gradient direction by RWF. The gradient of the RWF loss (3.3) is given as

$$\nabla \ell(\boldsymbol{z}) := \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^T \boldsymbol{z} - |\boldsymbol{a}_i^T \boldsymbol{x}| \cdot \mathrm{sgn}(\boldsymbol{a}_i^T \boldsymbol{z}) \right) \boldsymbol{a}_i, \tag{3.9}$$

where $\mathrm{sgn}(0) = 0$ by convention. Comparing this with the gradient of the least-squares loss

$$\nabla \ell_{LS}(\boldsymbol{z}) = \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^T \boldsymbol{z} - \boldsymbol{a}_i^T \boldsymbol{x} \right) \boldsymbol{a}_i, \tag{3.10}$$

one can see that RWF uses estimated phase information $\mathrm{sgn}(\boldsymbol{a}_i^T \boldsymbol{z})$ to generate the gradient updates, and the convergence behavior of RWF is much similar to that of least-squares with phase information if initialized properly. Indeed, Figure 3.3 illustrates that RWF takes

almost the *same* number of iterations for recovering a signal (with only the magnitude information) as the least-squares gradient descent method for recovering a signal (with both the magnitude and the sign information).

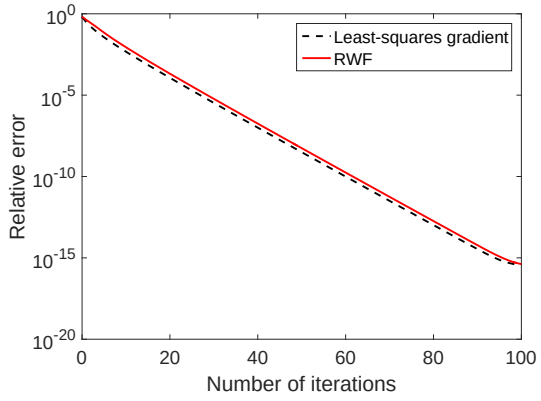


Fig. 3.3: Comparison of convergence behavior between RWF and the least-squares gradient descent with the same initialization, the same parameters $n = 1000$, $m = 6n$, and the same step size $\mu = 0.8$.

We next characterize the convergence of RWF in the following theorem.

Theorem 3.2. *Consider the problem of solving any given $\boldsymbol{x} \in \mathbb{R}^n$ from a system of equations* (3.1) *with Gaussian measurement vectors. There exist some universal constants $\mu_0 > 0$ (μ_0 can be set as 0.8 in practice), $0 < \rho, \nu < 1$ and $c_0, c_1, c_2 > 0$ such that if $m \geq c_0 n$ and $\mu < \mu_0$, then with probability at least $1 - c_1 \exp(-c_2 m)$, Algorithm 1 yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \leq \nu(1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N}. \tag{3.11}$$

Theorem 3.2 indicates that RWF recovers the true signal with $\mathcal{O}(n)$ samples, which is order-level optimal. Such an algorithm improves the sample complexity $\mathcal{O}(n \log n)$ of WF. Furthermore, RWF does not require truncation of weak samples in the gradient step to achieve the same sample complexity as TWF. This is mainly because RWF benefits from the lower-order loss function given in (3.3), the curvature of which behaves similarly to the least-squares loss function locally as we explain in Section 3.2.2.

Theorem 3.2 also suggests that RWF converges geometrically at a constant step size. To reach $\epsilon-$accuracy, it requires computational cost of $\mathcal{O}(mn\log 1/\epsilon)$ flops, which is better than WF ($\mathcal{O}(mn^2\log(1/\epsilon))$. Furthermore, it does not require truncation in gradient steps to reach the same computational cost as TWF. Numerically, as we demonstrate in Section 3.5, RWF is two times faster than TWF and four to six times faster than WF in terms of both iteration count and time cost in various examples.

We have established that RWF guarantees exact recovery with geometric convergence for noise-free case. We now study RWF in the presence of noise. Suppose the measurements are corrupted by bounded noise, and are given by

$$y_i = |\boldsymbol{a}_i^T\boldsymbol{x}| + w_i, \quad 1 \le i \le m, \tag{3.12}$$

where $\{w_i\}_{i=1}^m$ denote the additive noise. Then the following theorem shows that RWF is robust under such noise corruption.

Theorem 3.3. *Consider the model* (3.12). *Suppose that the measurement vectors are independently Gaussian, i.e.,* $\boldsymbol{a}_i \sim \mathcal{N}(0,\boldsymbol{I})$ *for* $1 \le i \le m$*, and the noise is bounded, i.e.,* $\|\boldsymbol{w}\|/\sqrt{m} \le c\|\boldsymbol{x}\|$. *Then there exist some universal constants* $\mu_0 > 0$ *(*μ_0 *can be set as* 0.8 *in practice),* $0 < \rho < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \ge c_0 n$ *and* $\mu < \mu_0$*, then with probability at least* $1 - c_1\exp(-c_2 m)$*, Algorithm 1 yields*

$$\textit{dist}(\boldsymbol{z}^{(t)},\boldsymbol{x}) \lesssim \frac{\|\boldsymbol{w}\|}{\sqrt{m}} + (1-\rho)^t\|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N}, \tag{3.13}$$

for some $\rho \in (0,1)$.

The numerical result under the Poisson noise model in Section 3.5 further corroborates the stability of RWF.

3.4 Incremental Reshaped Wirtinger Flow

In large-sample and online scenarios, stochastic algorithms are preferred due to their potential advantage of faster convergence and lower memory requirement. Thus, in this section, we develop stochastic versions of RWF, referred to as incremental reshaped Wirtinger flow (IRWF). We show that IRWF guarantees exact recovery at a linear convergence rate under the same sample complexity. We further draw the connection between IRWF and the randomized Kaczmarz method recently developed for phase retrieval [62, 63, 131], and establish its global convergence as a side product.

In oder to fully exploit the processing throughput of CPU/GPU, we develop a mini-batch IRWF, described in Algorithm 2. The mini-batch IRWF applies the same initialization step as in RWF, and uses a mini-batch of measurements for each gradient update.

Algorithm 2 Mini-batch Incremental Reshaped Wirtinger Flow (mini-batch IRWF)

Input: $\boldsymbol{y} = \{y_i\}_{i=1}^m$, $\{\boldsymbol{a}_i\}_{i=1}^m$, mini-batch size k;
Initialization: Same as in RWF (Algorithm 1);
Gradient loop: for $t = 0 : T-1$ do
Choose Γ_t uniformly at random from the subsets of $\{1, 2, \ldots, m\}$ with the cardinality k, and let

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \mu \cdot \boldsymbol{A}_{\Gamma_t}^* \left(\boldsymbol{A}_{\Gamma_t} \boldsymbol{z}^{(t)} - \boldsymbol{y}_{\Gamma_t} \odot \mathrm{Ph}(\boldsymbol{A}_{\Gamma_t} \boldsymbol{z}^{(t)}) \right), \tag{3.14}$$

where $\boldsymbol{A}_{\Gamma_t}$ is a matrix stacking $\boldsymbol{a}_i^*$ for $i \in \Gamma_t$ as its rows, $\boldsymbol{y}_{\Gamma_t}$ is a vector stacking y_i for $i \in \Gamma_t$ as its elements, and $\mathrm{Ph}(\boldsymbol{z})$ denotes the phase vector of $\boldsymbol{z}$.
Output $\boldsymbol{z}^{(T)}$.

If the gradient update uses only a single sample, i.e., $k = 1$, we refer to Algorithm 2 as IRWF, where the step (3.14) becomes

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \mu \left(\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)} - y_{i_t} \cdot \frac{\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)}|} \right) \boldsymbol{a}_{i_t}. \tag{3.15}$$

We characterize the convergence of mini-batch IRWF in the following theorem.

Theorem 3.4. *Consider the problem of solving any given $\boldsymbol{x} \in \mathbb{R}^n$ from a system of equa-*

tions (3.1) *with Gaussian measurement vectors. There exist some universal constants* $0 < \rho, \rho_0, \nu < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \geq c_0 n$ *and* $\mu = \rho_0/n$*, then with probability at least* $1 - c_1 \exp(-c_2 m)$*, Algorithm 2 yields*

$$\mathrm{E}_{\Gamma_{t+1}} \left[dist^2(\boldsymbol{z}^{(t+1)}, \boldsymbol{x}) \right] \leq \nu \left(1 - \frac{k\rho}{n} \right) \cdot dist^2(\boldsymbol{z}^{(t)}, \boldsymbol{x}), \tag{3.16}$$

for all $\boldsymbol{z}^{(t)}$ *satisfying* $\frac{dist(\boldsymbol{z}^{(t)}, \boldsymbol{x})}{\|\boldsymbol{z}\|} \leq \frac{1}{10}$*, where* $\mathrm{E}_{\Gamma_t}[\cdot]$ *denotes the expectation with respect to the randomly selected index set* Γ_t *conditioned on the high probability event of the random measurements* $\{\boldsymbol{a}_i\}_{i=1}^m$.

Proof. See Section 3.7.5. □

We suggest that $\rho_0 = 1$ and hence the step size $\mu = \frac{1}{n}$ in practice. Theorem 3.4 establishes that mini-batch IRWF achieves linear convergence to the global optimizer under the sample complexity $\mathcal{O}(n)$. For a generic optimization objective, it is not anticipated that incremental/stochastic first-order methods achieve linear convergence due to the variance of stochastic gradients. However, for our specific problem, the variance of stochastic gradients reduces as the estimate approaches the true signal, and hence a fixed step size can be employed and linear convergence can be established. A result similar in spirit was also established for the stochastic algorithm based on TWF (referred to as ITWF) [64]. We provide further comparisons between IRWF and ITWF in Section 3.4.2. On the other hand, it was shown in [132, 133] that stochastic gradient methods yield linear convergence to the minimizer $\boldsymbol{x}_\star$ if the objective $F(\boldsymbol{x}) = \sum_i f_i(\boldsymbol{x})$ is a smooth and strongly convex function and $\boldsymbol{x}_\star$ minimizes all components $f_i(\boldsymbol{x})$. The summands of our objective (3.3) also share the same minimizer (although it is neither convex nor smooth), which also helps to explain the linear convergence rate of IRWF.

3.4.1 Connection to the Kaczmarz Method for Phase Retrieval

The *Kaczmarz method* was originally developed for solving systems of linear equations [134]. In [62, 63], it was adapted to solve the phase retrieval problem, which we refer to as *Kaczmarz-PR*. It has been demonstrated in [62] that Kaczmarz-PR exhibits better empirical performance than error reduction (ER) [52, 53] and WF [1]. However, global convergence of Kaczmarz-PR has not been well established yet, although the randomized Kaczmarz method for the least-squares problem is established to converge at a linear rate [135, 136]. For instance, [62] obtained a bound on the estimation error which can be as large as the signal energy no matter how many iterations are taken. [63] established the asymptotic convergence in the regime when both m and n go to infinity but their ratio is fixed.

In this section, we draw a curious connection between IRWF and Kaczmarz-PR, which enables us to establish the theoretical guarantee of Kaczmarz-PR by adapting that of IRWF. This is analogous to the connection made in [133] between the Kaczmarz method and the stochastic gradient method for solving the least-squares problem. Here, the connection is made possible due to the lower-order loss function of RWF, which was not evident in previous studies of WF and TWF.

To be more specific, the Kaczmarz-PR [62, Algorithm 3] employs the following update rule

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{1}{\|\boldsymbol{a}_{i_t}\|^2} \left(\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)} - y_{i_t} \cdot \frac{\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^* \boldsymbol{z}^{(t)}|} \right) \boldsymbol{a}_{i_t}, \tag{3.17}$$

where i_t is selected either in a deterministic manner or randomly. We focus on the randomized case where i_t is selected uniformly at random from $\{1, \ldots, m\}$.

Comparing (3.17) and (3.15), the update rule of Kaczmarz-PR becomes equivalent to IRWF, if we replace the step size μ by $\frac{1}{\|\boldsymbol{a}_{i_t}\|^2}$. Moreover, these two update rules are close if μ is set as suggested, i.e., $\mu = \frac{1}{n}$, because for Gaussian measurements, $\|\boldsymbol{a}_{i_t}\|^2$ concentrates around n by the law of large numbers. As we demonstrate in the numerical experiments

(see Table 3.1), Kaczmarz-PR and IRWF have similar performance as anticipated. Thus, following the convergence guarantee for IRWF in Theorem 3.4, we establish the convergence guarantee for the randomized Kaczmarz-PR as follows.

Theorem 3.5. *Assume the measurement vectors are independent and each* $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I})$. *There exist some universal constants* $0 < \rho < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \geq c_0 n$, *then with probability at least* $1 - c_1 m \exp(-c_2 n)$, *the randomized Kaczmarz-PR update rule* (3.17) *yields*

$$\mathrm{E}_{i_t}\left[dist^2(\boldsymbol{z}^{(t+1)}, \boldsymbol{x})\right] \leq \left(1 - \frac{\rho}{n}\right) \cdot dist^2(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \tag{3.18}$$

for all $\boldsymbol{z}^{(t)}$ *satisfying* $\frac{dist(\boldsymbol{z}^{(t)}, \boldsymbol{x})}{\|\boldsymbol{z}\|} \leq \frac{1}{10}$.

Proof. See Section 3.7.6. □

The above theorem implies that once the estimate $\boldsymbol{z}^{(t)}$ enters the neighborhood of the true signal, the error diminishes at a linear rate in expectation.

Furthermore, [62] also provided a *block* Kaczmarz-PR (similar to the mini-batch version), whose update rule is given by

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \boldsymbol{A}_{\Gamma_t}^{\dagger}\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)} - \boldsymbol{y}_{\Gamma_t} \odot \mathrm{Ph}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right), \tag{3.19}$$

where Γ_t is a selected block at iterate t containing row indices, and $\dagger$ represents *Moore-Penrose pseudoinverse*, which is computed as follows:

$$\boldsymbol{A}^{\dagger} = \begin{cases} (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*, & \text{if } \boldsymbol{A} \text{ has linearly independent columns;} \\ \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}, & \text{if } \boldsymbol{A} \text{ has linearly independent rows.} \end{cases} \tag{3.20}$$

Comparing (3.19) and the mini-batch IRWF update in (3.14), these two update rules are similar to each other if $\boldsymbol{A}_{\Gamma_t}\boldsymbol{A}_{\Gamma_t}^*$ approaches $\frac{n}{\rho_0}\boldsymbol{I}_{|\Gamma_t|}$. For the case with Gaussian measure-

ments, $\boldsymbol{A}_{\Gamma_t}$ has linearly independent rows with high probability if $|\Gamma_t| \leq n$ and hence $\boldsymbol{A}_{\Gamma_t}\boldsymbol{A}_{\Gamma_t}^*$ is not far from $n\boldsymbol{I}_{|\Gamma_t|}$. Our numerical experiments (see Table 3.1) further suggest similar convergence rates for these two algorithms with the same block/mini-batch size.

Next, we argue that for the CDP setting, block Kaczmarz-PR is the same as the mini-batch IRWF with $\mu = 1$. The CDP measurements are collected in the following form

$$\boldsymbol{y}^{(l)} = |\boldsymbol{F}\boldsymbol{D}^{(l)}\boldsymbol{x}|, \quad 1 \leq l \leq L, \tag{3.21}$$

where $\boldsymbol{F}$ represents the discrete Fourier transform (DFT) matrix, $\boldsymbol{D}^{(l)}$ denotes a diagonal matrix (mask), and L denotes the number of masks. We choose the block size $|\Gamma_t|$ to be the dimension n of the signal for the convenience of Fourier transform. Then $\boldsymbol{A}_{\Gamma_t}$ becomes the Fourier transform composed with $\boldsymbol{D}^{(l)}$ (mask effect) and $\boldsymbol{A}_{\Gamma_t}^*$ becomes $\boldsymbol{D}^{(l)*}$ multiplied by the inverse Fourier transform. Therefore, $(\boldsymbol{A}_{\Gamma_t}\boldsymbol{A}_{\Gamma_t}^*) = \boldsymbol{I}$ if the diagonal elements of $\boldsymbol{D}^{(l)}$ have unit magnitude. Taking the step size $\mu = 1$, the two algorithms are identical.

On the other hand, since the block Kaczmarz-PR needs to calculate the matrix inverse or to solve an inverse problem, the block size cannot be too large. However, mini-batch IRWF works well for a wide range of the mini-batch sizes, which can even grow with the signal dimension n as long as a batch of data is loadable into the memory.

3.4.2 Comparison with Incremental Truncated Wirtinger Flow (ITWF)

Recently, [64] designed and analyzed an incremental algorithm based on TWF, which is referred to as ITWF. More specifically, ITWF employs the same initialization procedure as TWF and randomly chooses one sample with the index i_t selected uniformly at random from $\{1, 2, \ldots, m\}$ for the gradient update as follows:

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \mu \cdot \frac{|\boldsymbol{a}_{i_t}^T\boldsymbol{z}|^2 - y_{i_t}^2}{\boldsymbol{a}_{i_t}^T\boldsymbol{z}}\boldsymbol{a}_{i_t}\mathbf{1}_{\mathcal{E}_{1,t}^{i_t}\cap\mathcal{E}_3^{i_t}}, \tag{3.22}$$

where $\mathbf{1}_{\mathcal{E}_{1,t}^{i_t} \cap \mathcal{E}_3^{i_t}}$ represents the truncation rule determined by the events $\mathcal{E}_{1,t}^{i_t}$ and $\mathcal{E}_3^{i_t}$. As a comparison, the update rule of IRWF is much simpler due to the use of the lower-order loss function and does not require any truncation in the gradient loop. [64] proved that ITWF converges linearly to the true signal as long as m/n is large enough. Compared to ITWF, IRWF also achieves the same linear convergence, but runs faster than ITWF numerically as demonstrated in Section 3.5.

3.5 Numerical Results

In this section, we demonstrate the numerical efficiency of RWF and (mini-batch) IRWF by comparing their performance with other competitive algorithms. Our experiments are conducted not only for the real Gaussian case but also for the complex Gaussian and the CDP cases. All the experiments are implemented in Matlab 2015b and conducted on a computer equipped with Intel Core i7 3.4GHz CPU and 12GB RAM.

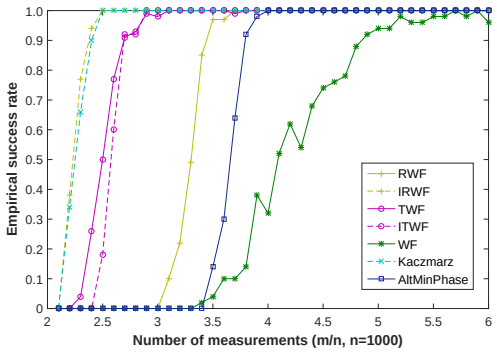
We first compare the sample complexity of RWF and IRWF with those of TWF, WF, Kaczmarz-PR and AltMinPhase via the empirical successful recovery rate versus the number of measurements. For RWF, we follow Algorithm 1 with the suggested parameters. For IRWF, we adopt a block size 64 for efficiency and set the step size $\mu = 1/n$. For WF, TWF, we use the code provided in the original papers with the suggested parameters. For ITWF, we also adopt a block size 64 and set the step size $\mu = 0.6/n$ (optimal step size). We conduct the experiments for the real Gaussian, complex Gaussian and CDP cases respectively. For the real and complex cases, we set the signal dimension n to be 1000, and set the ratio m/n to take values from 2 to 6 with a step 0.1. For each m, we run 100 trials and count the number of successful trials. For each trial, we run a maximal number of iterations/passes $T = 10000$ for all algorithms, and a trial is declared to be successful whenever the iterate satisfies $\text{dist}(\boldsymbol{z}^{(T)}, \boldsymbol{x})/\|\boldsymbol{x}\| \leq 10^{-5}$. For the real Gaussian case, we generate the signal $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{n\times n})$, and generate the mea-

surement vectors $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n\times n})$ i.i.d. for $i = 1, \ldots, m$. For the complex Gaussian case, we generate the signal $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_{n\times n}) + j\mathcal{N}(0, \boldsymbol{I}_{n\times n})$ and the measurement vectors $\boldsymbol{a}_i \sim \frac{1}{2}\mathcal{N}(0, \boldsymbol{I}_{n\times n}) + j\frac{1}{2}\mathcal{N}(0, \boldsymbol{I}_{n\times n})$ i.i.d. for $i = 1, \ldots, m$. For the CDP case (3.21), we set $n = 1024$ for the convenience of FFT and $m/n = L = 1, 2, \ldots, 8$. All other settings are the same as those for the real case.

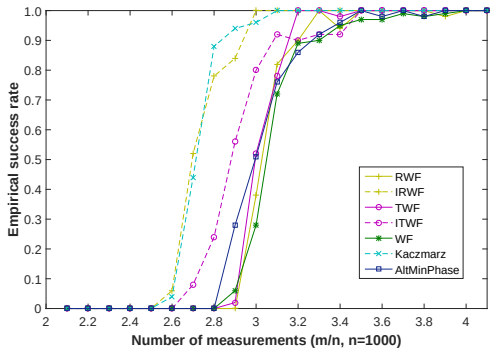
We note that for the CDP case, the Kaczmarz-PR algorithm is identical to the IRWF with step size $\mu = 1/n$ due to the argument in Section 3.4.1. Moreover under the CDP case, the AltMinPhase algorithm is identical to the RWF with step size $\mu = 1$ because the inverse of the Fourier measurement matrix is nothing but its conjugate transpose. In the following experiments for the CDP case, we choose the step size $\mu = 1/n$ for the IRWF and $\mu = 1$ for the RWF, under which the Kaczmarz-PR algorithm coincides with the IRWF and the AltMinPhase algorithm coincides with the RWF.

Figure 3.4 plots the fraction of successful trials out of 100 trials for all algorithms, with respect to m/n. It can be seen that IRWF and Kaczmarz-PR exhibit a similar sample complexity, which is the best for all three cases, and is close to the theoretical limit [137]. It can also be seen that the two incremental methods (IRWF and ITWF) outperform the batch methods (RWF, TWF, AltMinPhase and WF). This can be due to the inherent noise in incremental methods, which helps to escape bad local minima. This can be extremely helpful in the regime with a small number of samples, where local minima do exist near the global minima. Comparing among the batch methods (RWF, TWF AltMinPhase and WF), it can be seen that although RWF outperforms only WF and AltMinPhase (not TWF) for the real Gaussian case, it has a comparable performance for the complex case and outperforms TWF and WF in the CDP case. An intuitive explanation for the real case is that a substantial number of samples with small $|\boldsymbol{a}_i^T\boldsymbol{z}|$ can deviate the gradient direction so that truncation indeed helps to stabilize the algorithm if the number of measurements is not large.

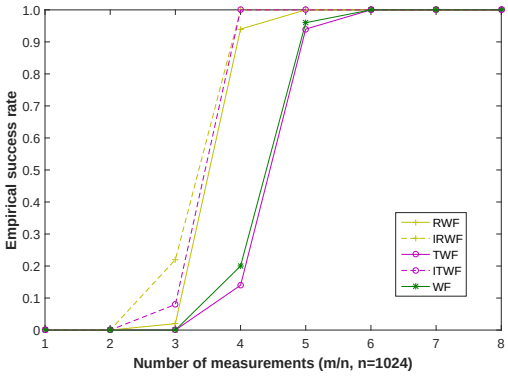
We next compare the convergence rate of RWF, IRWF with those of TWF, ITWF, WF, Kaczmarcz and AltMinPhase. We run all of the algorithms with the suggested parameters



(a) Real Gaussian case



(b) Complex Gaussian case



(c) CDP case

Fig. 3.4: Comparison of sample complexity among RWF, IRWF, TWF, ITWF, WF, Karzmarz-PR and AltMinPhase.

in the original code. We generate the signal and measurements in the same way as those in the first experiment with $n = 5000, m = 8n$. All algorithms are seeded with the RWF initialization. In Table 3.1, we list the number of passes and the time cost for all the algorithms to achieve a relative error of 10^{-14} averaged over 10 trials. For the incremental methods, one update passes k samples and one pass amounts to m/k updates. Clearly, IRWF with mini-batch size 64 runs the fastest for both the real and complex cases. Moreover, among the batch (deterministic) algorithms, RWF takes much fewer passes as well as runs much faster than TWF and WF. Although RWF takes more iterations than AltMinPhase, it runs much faster than AltMinPhase due to the fact that each iteration of AltMinPhase needs to solve a least-squares problem that takes much longer than a simple gradient update in RWF.

We also compare the performance of the above algorithms on the recovery of a real

Table 3.1: Comparison of iteration count and time cost among algorithms ($n = 5000, m = 8n$).

		Real Gaussian		Complex Gaussian	
		#passes	time(s)	# passes	time(s)
Batch methods	RWF	72	**12.66**	176	**122.4**
	TWF	186	32.36	487	395.1
	WF	319	54.83	932	887.8
	AltMinPhase	**6**	79.58	**159**	9637
Incremental methods	IRWF	9	44.77	21	233.2
	mini-batch IRWF (64)	9	**8.076**	**21**	**48.58**
	mini-batch ITWF (64)	16	37.38	29	149.5
	Kaczmarz-PR	9	50.68	21	248.4
	block Kaczmarz-PR (64)	**8**	28.50	22	89.31

image from the Fourier intensity measurements (the two dimensional CDP case). The image (see Figure 3.5) is the Milky Way Galaxy with resolution 1920×1080. Table 3.2 lists the number of passes and the time cost for the above six algorithms to achieve the relative error of 10^{-15} for one R/G/B channel. All algorithms are seeded with the RWF initialization. To explore the advantage of FFT, we run the incremental/stochastic methods with the mini-batch size equal to the number of pixels for one R/G/B channel. We note that with such a mini-batch size, IRWF is equivalent to block Kaczmarz-PR from the discussion in Section 3.4.1. It can be seen that in general, the incremental/stochastic methods (IRWF and ITWF) run faster than the batch methods (RWF, TWF, WF). Moreover, among the batch methods, RWF outperforms the other three algorithms in both the number of passes and the computational time. In particular, RWF runs two times faster than TWF and six times faster than WF in terms of both the number of iterations and the computational time.

We next demonstrate the robustness of RWF to noise and compare it with TWF. We consider the phase retrieval problem in imaging applications, where Poisson noise is often used to model the sensor and electronic noise [138]. Specifically, the noisy measurements

Fig. 3.5: Milky way Galaxy.

Table 3.2: Comparison of iterations and time cost among algorithms on recovery of Galaxy image (shown in Figure 3.5), where $L = m/n$ denotes the number of CDP masks.

	Algorithms	RWF	IRWF	TWF	ITWF	WF
$L = 6$	#passes	140	**24**	410	41	fail
	time cost(s)	110	**21.2**	406	43	fail
$L = 12$	#passes	70	**8**	190	12	315
	time cost(s)	107	**13.7**	363.6	25.9	426

of intensity can be expressed as $y_i = \sqrt{\alpha \cdot \text{Poisson}\left(|\boldsymbol{a}_i^T \boldsymbol{x}|^2/\alpha\right)}$, for $i = 1, 2, ...m$ where α denotes the level of the input noise, and Poisson(λ) denotes a random sample generated by the Poisson distribution with mean λ. It can be observed from Figure 3.6 that RWF performs better than TWF in terms of the recovery accuracy under two different noise levels.

3.6 Conclusion

In this chapter, we study RWF and its stochastic version IRWF to recover a signal from a quadratic systems of equations, based on a *nonconvex and nonsmooth* quadratic loss function of magnitude measurements. This loss function sacrifices the smoothness but

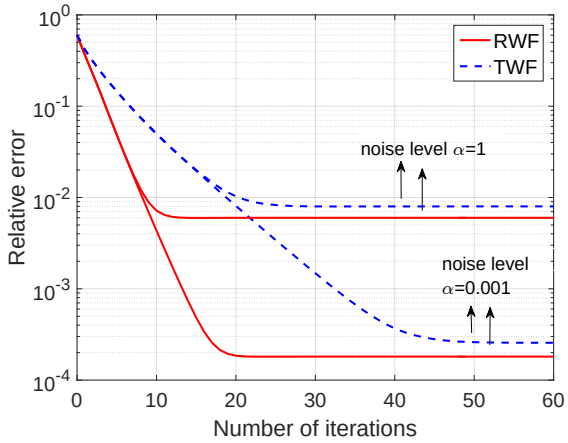


Fig. 3.6: Comparison of relative error under Poisson noise between RWF and TWF.

enjoys advantages in statistical and computational efficiency. It has potential to be extended in various scenarios. One interesting direction is to extend such an algorithm to exploit signal structures (e.g., non-negativity, sparsity, etc) to assist the recovery. The lower-order loss function may offer great simplicity to prove performance guarantee in such cases.

Another interesting direction is to study the convergence of algorithms from random initialization. In the regime of large sample size ($m \gg n$), the empirical loss surface approaches the asymptotic loss (Figure 3.2(b)) and hence has no spurious local minima. Due to the result [139], it is conceivable that gradient descent converges from random starting point. Similar phenomenons have been observed in [43, 80]. However, under moderate number of measurements ($m < 10n$), authentic local minima do exist and often locate not far from the global ones. In this regime, the batch gradient method often fails with random initialization. As always believed, stochastic algorithms are efficient in escaping bad local minima or saddle points in nonconvex optimization because of the inherent noise [82, 140]. We observe numerically that IRWF and block IRWF from random starting point still converge to global minimum even with very small sample size which is close to the theoretical limits [141]. It is of interest to analyze theoretically that stochastic methods escape these local minima (not just saddle points) efficiently.

3.7 Technical Proofs

We first introduce some notations here. We let $\mathcal{A} : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^m$ be a linear map

$$\boldsymbol{M} \in \mathbb{R}^{n\times n} \mapsto \mathcal{A}(\boldsymbol{M}) := \{\boldsymbol{a}_i^T \boldsymbol{M} \boldsymbol{a}_i\}_{1\le i\le m}.$$

We let $\|\cdot\|_1$ and $\|\cdot\|$ denote the l_1 norm and l_2 norm of a vector, respectively. Moreover, let $\|\cdot\|_F$ and $\|\cdot\|$ denote the Frobenius norm and the spectral norm of a matrix, respectively. We note that the constants c, C, c_0, c_1, c_2 may be different from line to line, for the sake of notational simplicity.

3.7.1 Proof of Proposition 3.1: Initialization

The idea of using truncation to bound some non-sub-Gaussian sequences has appeared in previous works [45, Lemma2.3] and TWF [61]. Compared to the proof for TWF, this proof has new technical developments to address the magnitude measurements and truncation from both sides.

We first estimate the norm of $\boldsymbol{x}$ as

$$\lambda_0 = \frac{mn}{\sum_{i=1}^m \|\boldsymbol{a}_i\|_1} \cdot \left(\frac{1}{m}\sum_{i=1}^m y_i\right). \tag{3.23}$$

Since $\boldsymbol{a}_i \sim \mathcal{N}(0, \boldsymbol{I}_{n\times n})$, by Hoeffding-type inequality, it can be shown that

$$\left|\frac{\sum_{i=1}^m \|\boldsymbol{a}_i\|_1}{mn} - \sqrt{\frac{2}{\pi}}\right| < \frac{\epsilon}{3} \tag{3.24}$$

holds with probability at least $1 - 2\exp(-c_1 mn\epsilon^2)$ for some constant $c_1 > 0$.

Moreover, given $\boldsymbol{x}$, y_i's are independent sub-Gaussian random variables. Thus, by

Hoeffding-type inequality, it can be shown that

$$\left|\sqrt{\frac{\pi}{2}}\left(\frac{1}{m}\sum_{i=1}^{m} y_i\right) - \|\boldsymbol{x}\|\right| < \frac{\epsilon}{3}\|\boldsymbol{x}\| \tag{3.25}$$

holds with probability at least $1 - 2\exp(-c_1 m\epsilon^2)$ for some constant $c_1 > 0$.

On the event $E_1 = \{\text{both (3.24) and (3.25) hold}\}$, it can be argued that

$$|\lambda_0 - \|\boldsymbol{x}\|| < \epsilon\|\boldsymbol{x}\|. \tag{3.26}$$

Without loss of generality, we let $\|\boldsymbol{x}\| = 1$. Then on the event E_1, the truncation function satisfies the following bounds

$$\mathbf{1}_{\{\alpha_l(1+\epsilon)<|\boldsymbol{a}_i^T\boldsymbol{x}|<\alpha_u(1-\epsilon)\}} \le \mathbf{1}_{\{\alpha_l\lambda_0<y_i<\alpha_u\lambda_0\}} \le \mathbf{1}_{\{\alpha_l(1-\epsilon)<|\boldsymbol{a}_i^T\boldsymbol{x}|<\alpha_u(1+\epsilon)\}}.$$

Thus, by defining

$$\begin{aligned}\boldsymbol{Y}_1 :=& \frac{1}{m}\sum \boldsymbol{a}_i\boldsymbol{a}_i^T|\boldsymbol{a}_i^T\boldsymbol{x}|\mathbf{1}_{\{\alpha_l(1+\epsilon)<|\boldsymbol{a}_i^T\boldsymbol{x}|<\alpha_u(1-\epsilon)\}}\\ \boldsymbol{Y}_2 :=& \frac{1}{m}\sum \boldsymbol{a}_i\boldsymbol{a}_i^T|\boldsymbol{a}_i^T\boldsymbol{x}|\mathbf{1}_{\{\alpha_l(1-\epsilon)<|\boldsymbol{a}_i^T\boldsymbol{x}|<\alpha_u(1+\epsilon)\}},\end{aligned}$$

we have $\boldsymbol{Y}_1 \prec \boldsymbol{Y} \prec \boldsymbol{Y}_2$. We further compute the expectations of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ and obtain

$$\mathrm{E}[\boldsymbol{Y}_1] = (\beta_1\boldsymbol{x}\boldsymbol{x}^T + \beta_2\boldsymbol{I}), \quad \mathrm{E}[\boldsymbol{Y}_2] = (\beta_3\boldsymbol{x}\boldsymbol{x}^T + \beta_4\boldsymbol{I}), \tag{3.27}$$

where

$$\begin{aligned}\beta_1 &:= \mathrm{E}[|\xi|^3\mathbf{1}_{\{\alpha_l(1+\epsilon)<|\xi|<\alpha_u(1-\epsilon)\}}] - \mathrm{E}[|\xi|\mathbf{1}_{\{\alpha_l(1+\epsilon)<|\xi|<\alpha_u(1-\epsilon)\}}],\\ \beta_3 &:= \mathrm{E}[|\xi|^3\mathbf{1}_{\{\alpha_l(1-\epsilon)<|\xi|<\alpha_u(1+\epsilon)\}}] - \mathrm{E}[|\xi|\mathbf{1}_{\{\alpha_l(1-\epsilon)<|\xi|<\alpha_u(1+\epsilon)\}}],\\ \beta_2 &:= \mathrm{E}[|\xi|\mathbf{1}_{\{\alpha_l(1+\epsilon)<|\xi|<\alpha_u(1-\epsilon)\}}], \quad \beta_4 := \mathrm{E}[|\xi|\mathbf{1}_{\{\alpha_l(1-\epsilon)<|\xi|<\alpha_u(1+\epsilon)\}}]\end{aligned}$$

where $\xi \sim \mathcal{N}(0,1)$. For given α_l and α_u, small value of ϵ yields arbitrarily close β_1 and β_3, as well as arbitrarily close β_2 and β_4. For example, taking $\alpha_l = 1, \alpha_u = 5$ and $\epsilon = 0.01$, we have $\beta_1 = 0.9678, \beta_2 = 0.4791, \beta_3 = 0.9688, \beta_4 = 0.4888$.

Now applying standard results on random matrices with non-isotropic sub-Gaussian rows [142, equation (5.26)] and noticing that $\boldsymbol{a}_i\boldsymbol{a}_i^T|\boldsymbol{a}_i^T\boldsymbol{x}|\mathbf{1}_{\{\alpha_l(1+\epsilon)<|\boldsymbol{a}_i^T\boldsymbol{x}|<\alpha_u(1-\epsilon)\}}$ can be rewritten as $\boldsymbol{b}_i\boldsymbol{b}_i^T$ for sub-Gaussian vector $\boldsymbol{b}_i := \boldsymbol{a}_i\sqrt{|\boldsymbol{a}_i^T\boldsymbol{x}|}\mathbf{1}_{\{\alpha_l(1+\epsilon)<|\boldsymbol{a}_i^T\boldsymbol{x}|<\alpha_u(1-\epsilon)\}}$, one can derive

$$\|\boldsymbol{Y}_1 - \mathrm{E}[\boldsymbol{Y}_1]\| \le \delta, \quad \|\boldsymbol{Y}_2 - \mathrm{E}[\boldsymbol{Y}_2]\| \le \delta \tag{3.28}$$

with probability $1 - 4\exp(-c_1(\delta)m)$ for some positive c_1 which is only affected by δ, provided that m/n exceeds a certain constant. Furthermore, when ϵ is sufficiently small, one further has $\|\mathrm{E}[\boldsymbol{Y}_1] - \mathrm{E}[\boldsymbol{Y}_2]\| \le \delta$. Combining the above facts together, one can show that

$$\|\boldsymbol{Y} - (\beta_1\boldsymbol{x}\boldsymbol{x}^T + \beta_2\boldsymbol{I})\| \le 3\delta. \tag{3.29}$$

Let $\tilde{\boldsymbol{z}}^{(0)}$ be the normalized leading eigenvector of $\boldsymbol{Y}$. Following the arguments in [1, Section 7.8] and taking δ and ϵ to be sufficiently small, one has

$$\mathrm{dist}(\tilde{\boldsymbol{z}}^{(0)}, \boldsymbol{x}) \le \tilde{\delta}, \tag{3.30}$$

for a given $\tilde{\delta} > 0$, as long as m/n exceeds a certain constant.

3.7.2 Proof in Section 3.2.2: Expectation of loss functions

The expectation of the loss function (3.2) of WF is given by [80] as

$$\mathrm{E}[\ell_{WF}(\boldsymbol{z})] = \frac{3}{4}\|\boldsymbol{x}\|^4 + \frac{3}{4}\|\boldsymbol{z}\|^4 - \frac{1}{2}\|\boldsymbol{x}\|^2\|\boldsymbol{z}\|^2 - |\boldsymbol{z}^T\boldsymbol{x}|^2. \tag{3.31}$$

We next show that the expectation of the loss function (3.3) of RWF has the following form:

$$\mathrm{E}[\ell(\boldsymbol{z})] = \frac{1}{2}\|\boldsymbol{x}\|^2 + \frac{1}{2}\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|\|\boldsymbol{z}\| \cdot \mathrm{E}\left[\frac{|\boldsymbol{a}_i^T\boldsymbol{z}|}{\|\boldsymbol{z}\|} \cdot \frac{|\boldsymbol{a}_i^T\boldsymbol{x}|}{\|\boldsymbol{x}\|}\right], \tag{3.32}$$

where

$$\mathrm{E}\left[\frac{|\boldsymbol{a}_i^T\boldsymbol{z}|}{\|\boldsymbol{z}\|} \cdot \frac{|\boldsymbol{a}_i^T\boldsymbol{x}|}{\|\boldsymbol{x}\|}\right] = \begin{cases} \frac{(1-\rho^2)^{3/2}}{\pi}\int_0^\infty t(e^{\rho t} + e^{-\rho t})K_0(t)dt, & \text{if } |\rho| < 1; \\ 1, & \text{if } |\rho| = 1; \end{cases} \tag{3.33}$$

where $\rho = \frac{\boldsymbol{z}^T\boldsymbol{x}}{\|\boldsymbol{x}\|\|\boldsymbol{z}\|}$ and $K_0(\cdot)$ is the modified Bessel function of the second kind.

In order to derive (3.33), we first define

$$u := \frac{\boldsymbol{a}_i^T\boldsymbol{z}}{\|\boldsymbol{z}\|} \text{ and } v := \frac{\boldsymbol{a}_i^T\boldsymbol{x}}{\|\boldsymbol{x}\|},$$

and it suffices to drive $\mathrm{E}[|uv|]$. Note that $(u, v) \sim \mathcal{N}(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \text{and} \quad \rho = \frac{\boldsymbol{z}^T\boldsymbol{x}}{\|\boldsymbol{x}\|\|\boldsymbol{z}\|}.$$

Following [143], the density function of $u \cdot v$ is given by

$$\phi_{uv}(x) = \frac{1}{\pi\sqrt{1-\rho^2}} \exp\left(\frac{\rho x}{1-\rho^2}\right) K_0\left(\frac{|x|}{1-\rho^2}\right), \quad x \neq 0.$$

Thus, the density of $|uv|$ is given by

$$\psi_{|uv|}(x) = \frac{1}{\pi\sqrt{1-\rho^2}} \left[\exp\left(\frac{\rho x}{1-\rho^2}\right) + \exp\left(-\frac{\rho x}{1-\rho^2}\right)\right] K_0\left(\frac{|x|}{1-\rho^2}\right), \quad x > 0,$$

for $|\rho| < 1$. Therefore, if $|\rho| < 1$, then

$$\begin{aligned}
\mathrm{E}[|uv|] &= \int_0^\infty x \cdot \psi_\rho(x) dx \\
&= \int_0^\infty x \cdot \frac{1}{\pi\sqrt{1-\rho^2}} \left[\exp\left(\frac{\rho x}{1-\rho^2}\right) + \exp\left(-\frac{\rho x}{1-\rho^2}\right)\right] K_0\left(\frac{|x|}{1-\rho^2}\right) dx \\
&= \frac{(1-\rho^2)^{3/2}}{\pi} \int_0^\infty t(e^{\rho t} + e^{-\rho t}) K_0(t) dt
\end{aligned}$$

where the last step follows by changing variables.

If $|\rho| = 1$, then $|uv|$ becomes a χ_1^2 random variable, with the density

$$\psi_{|uv|}(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} \exp(-x/2), \quad x > 0,$$

and hence $\mathrm{E}[|uv|] = 1$.

3.7.3 Proof of Theorem 3.2: Geometric Convergence of RWF

The general structure of the proof follows that for WF in [1] and TWF in [61]. However, the proof requires development of new bounds due to the nonsmoothness of the loss function and absolute value based measurements. On the other hand the proof is much simpler due to the lower-order loss function adopted in RWF.

We first introduce a global phase notation for real case as follows:

$$\Phi(z) := \begin{cases} 0, & \text{if } \|\boldsymbol{z} - \boldsymbol{x}\| \le \|\boldsymbol{z} + \boldsymbol{x}\|, \\ \pi, & \text{otherwise.} \end{cases} \tag{3.34}$$

For the sake of simplicity, we let $\boldsymbol{z}$ be $e^{-j\Phi(\boldsymbol{z})}\boldsymbol{z}$, which indicates that $\boldsymbol{z}$ is always in the neighborhood of $\boldsymbol{x}$. Furthermore, we denote $\boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$.

The idea of the proof is to show that within the neighborhood of global optima, RWF

satisfies the *Regularity Condition* $\mathsf{RC}(\mu, \lambda, c)$, i.e.,

$$\langle \nabla \ell(\boldsymbol{z}), \boldsymbol{h} \rangle \geq \frac{\mu}{2} \|\nabla \ell(\boldsymbol{z})\|^2 + \frac{\lambda}{2} \|\boldsymbol{h}\|^2 \tag{3.35}$$

for all $\boldsymbol{z}$ and $\boldsymbol{h} = \boldsymbol{z} - \boldsymbol{x}$ obeying $\|\boldsymbol{h}\| \leq c\|\boldsymbol{x}\|$, where $0 < c < 1$ is some constant. Then, as shown in [61], once the initialization lands into this neighborhood, linear convergence can be guaranteed, i.e.,

$$\text{dist}^2 \left(\boldsymbol{z} - \mu \nabla \ell(\boldsymbol{z}), \boldsymbol{x} \right) \leq (1 - \mu\lambda) \text{dist}^2(\boldsymbol{z}, \boldsymbol{x}), \tag{3.36}$$

for any $\boldsymbol{z}$ with $\|\boldsymbol{z} - \boldsymbol{x}\| \leq c\|\boldsymbol{x}\|$.

To show the regularity condition, we first define a set $\mathcal{S} := \{i : 1 \leq i \leq m, (\boldsymbol{a}_i^T \boldsymbol{z})(\boldsymbol{a}_i^T \boldsymbol{x}) < 0\}$, and then derive the following bound:

$$\begin{aligned}
\langle \nabla \ell(\boldsymbol{z}), \boldsymbol{h} \rangle &= \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^T \boldsymbol{z} - |\boldsymbol{a}_i^T \boldsymbol{x}| \text{sgn}(\boldsymbol{a}_i^T \boldsymbol{z}) \right) (\boldsymbol{a}_i^T \boldsymbol{h}) \\
&= \frac{1}{m} \left[\sum_{i=1}^{m} (\boldsymbol{a}_i^T \boldsymbol{h})^2 + 2 \sum_{i \in \mathcal{S}} (\boldsymbol{a}_i^T \boldsymbol{x})(\boldsymbol{a}_i^T \boldsymbol{h}) \right] \\
&\geq \frac{1}{m} \left[\sum_{i=1}^{m} (\boldsymbol{a}_i^T \boldsymbol{h})^2 - 2 \left| \sum_{i \in \mathcal{S}} (\boldsymbol{a}_i^T \boldsymbol{x})(\boldsymbol{a}_i^T \boldsymbol{h}) \right| \right] \\
&\geq \frac{1}{m} \left[\sum_{i=1}^{m} (\boldsymbol{a}_i^T \boldsymbol{h})^2 - \sum_{i \in \mathcal{S}} 2 \left| (\boldsymbol{a}_i^T \boldsymbol{x})(\boldsymbol{a}_i^T \boldsymbol{h}) \right| \right].
\end{aligned} \tag{3.37}$$

The first term in (3.37) can be bounded using Lemma 3.1 in [45], which we state below.

Lemma 3.6. *For any $0 < \epsilon < 1$, if $m > c_0 n \epsilon^{-2}$, then with probability at least $1 - 2\exp(-c_1 \epsilon^2 m)$,*

$$(1 - \epsilon)\|\boldsymbol{h}\|^2 \leq \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{a}_i^T \boldsymbol{h})^2 \leq (1 + \epsilon)\|\boldsymbol{h}\|^2 \tag{3.38}$$

holds for all non-zero vectors $\boldsymbol{h} \in \mathbb{R}^n$. Here, $c_0, c_1 > 0$ are some universal constants.

For the second term in (3.37), we derive

$$\begin{aligned}\sum_{i\in\mathcal{S}} 2\left|\boldsymbol{a}_i^T\boldsymbol{x}\right|\left|\boldsymbol{a}_i^T\boldsymbol{h}\right| &\le \sum_{i\in\mathcal{S}}\left[(\boldsymbol{a}_i^T\boldsymbol{x})^2+(\boldsymbol{a}_i^T\boldsymbol{h})^2\right]\\ &= \sum_{i=1}^{m}[(\boldsymbol{a}_i^T\boldsymbol{x})^2+(\boldsymbol{a}_i^T\boldsymbol{h})^2]\cdot\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<0\}}\\ &= \sum_{i=1}^{m}[(\boldsymbol{a}_i^T\boldsymbol{x})^2+(\boldsymbol{a}_i^T\boldsymbol{h})^2]\cdot\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})^2+(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{h})<0\}}\\ &\le \sum_{i=1}^{m}[(\boldsymbol{a}_i^T\boldsymbol{x})^2+(\boldsymbol{a}_i^T\boldsymbol{h})^2]\cdot\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}}\\ &\le 2\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2\cdot\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}}. \end{aligned}\tag{3.39}$$

The above equation can be further upper bounded by the following lemma.

Lemma 3.7. *For any $\epsilon > 0$, if $m > c_0 n\epsilon^{-2}\log\epsilon^{-1}$, then with probability at least $1 - C\exp(-c_1\epsilon^2 m)$,*

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2\cdot\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \le (0.13+\epsilon)\left\|\boldsymbol{h}\right\|^2 \tag{3.40}$$

holds for all non-zero vectors $\boldsymbol{h}\in\mathbb{R}^n$ satisfying $\|\boldsymbol{h}\|\le\frac{1}{10}\|\boldsymbol{x}\|$. Here, $c_0, c_1, C > 0$ are some universal constants.

Proof. We first prove bounds for any fixed $\boldsymbol{h}\le\frac{1}{10}\|\boldsymbol{x}\|$, and then develop a uniform bound later on. We introduce a series of auxiliary random Lipschitz functions to approximate the indicator functions. For $i = 1,\dots,m$, define

$$\chi_i(t) := \begin{cases} t, & \text{if } t > (\boldsymbol{a}_i^T\boldsymbol{x})^2;\\ \frac{1}{\delta}(t-(\boldsymbol{a}_i^T\boldsymbol{x})^2)+(\boldsymbol{a}_i^T\boldsymbol{x})^2, & \text{if } (1-\delta)(\boldsymbol{a}_i^T\boldsymbol{x})^2\le t\le(\boldsymbol{a}_i^T\boldsymbol{x})^2;\\ 0, & \text{else;}\end{cases}\tag{3.41}$$

and then $\chi_i(t)$'s are random Lipschitz functions with Lipschitz constant $\frac{1}{\delta}$. We further have

$$|\boldsymbol{a}_i^T\boldsymbol{h}|^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \le \chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2) \le |\boldsymbol{a}_i^T\boldsymbol{h}|^2\mathbf{1}_{\{(1-\delta)|\boldsymbol{a}_i^T\boldsymbol{x}|^2<|\boldsymbol{a}_i^T\boldsymbol{h}|^2\}}. \tag{3.42}$$

For convenience, we denote $\gamma_i := \frac{|\boldsymbol{a}_i^T\boldsymbol{h}|^2}{\|\boldsymbol{h}\|^2}\mathbf{1}_{\{(1-\delta)|\boldsymbol{a}_i^T\boldsymbol{x}|^2<|\boldsymbol{a}_i^T\boldsymbol{h}|^2\}}$ and $\theta := \|\boldsymbol{h}\|/\|\boldsymbol{x}\|$. We next estimate the expectation of γ_i, by conditional expectation,

$$\mathrm{E}[\gamma_i] = \int_\Omega \gamma_i d\mathbb{P} = \iint_{-\infty}^{\infty} \mathrm{E}\left[\gamma_i \middle| \boldsymbol{a}_i^T\boldsymbol{x} = \tau_1\|\boldsymbol{x}\|, \boldsymbol{a}_i^T\boldsymbol{h} = \tau_2\|\boldsymbol{h}\|\right] \cdot f(\tau_1,\tau_2)d\tau_1 d\tau_2, \tag{3.43}$$

where $f(\tau_1,\tau_2)$ is the density of two joint Gaussian random variables with correlation $\rho = \frac{\boldsymbol{h}^T\boldsymbol{x}}{\|\boldsymbol{h}\|\|\boldsymbol{x}\|} \neq \pm 1$. We then continue to derive

$$\begin{aligned}
\mathrm{E}[\gamma_i] &= \iint_{-\infty}^{\infty} \tau_2^2 \cdot \mathbf{1}_{\{\sqrt{1-\delta}|\tau_1|<|\tau_2|\theta\}} \cdot f(\tau_1,\tau_2)d\tau_1 d\tau_2 \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \int_{\frac{-|\tau_2|\theta}{\sqrt{1-\delta}}}^{\frac{|\tau_2|\theta}{\sqrt{1-\delta}}} \exp\left(-\frac{(\tau_1-\rho\tau_2)^2}{2(1-\rho^2)}\right) d\tau_1 d\tau_2 \qquad (3.44) \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \int_{\frac{-\frac{|\tau_2|\theta}{\sqrt{1-\delta}}-\rho\tau_2}{\sqrt{1-\rho^2}}}^{\frac{\frac{|\tau_2|\theta}{\sqrt{1-\delta}}-\rho\tau_2}{\sqrt{1-\rho^2}}} \exp\left(-\frac{\tau^2}{2}\right) d\tau d\tau_2 \qquad \text{by changing variables} \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \sqrt{\frac{\pi}{2}} \left(\mathrm{erf}\left(\frac{\frac{|\tau_2|\theta}{\sqrt{1-\delta}}-\rho\tau_2}{\sqrt{1-\rho^2}}\right) - \mathrm{erf}\left(\frac{-\frac{|\tau_2|\theta}{\sqrt{1-\delta}}-\rho\tau_2}{\sqrt{1-\rho^2}}\right)\right) d\tau_2 \\
&= \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \tau_2^2 \exp\left(-\frac{\tau_2^2}{2}\right) \cdot \left(\mathrm{erf}\left(\frac{(\frac{\theta}{\sqrt{1-\delta}}-\rho)\tau_2}{\sqrt{1-\rho^2}}\right) + \mathrm{erf}\left(\frac{(\frac{\theta}{\sqrt{1-\delta}}+\rho)\tau_2}{\sqrt{1-\rho^2}}\right)\right) d\tau_2. \qquad (3.45)
\end{aligned}$$

For $|\rho| < 1$, $\mathrm{E}[\gamma_i]$ is a continuous function of ρ. For $|\rho| = 1$, $\mathrm{E}[\gamma_i] = 0$. The last integral (3.45) can be calculated numerically. Figure 3.7 plots $\mathrm{E}[\gamma_i]$ for $\theta = 0.1$ and $\delta = 0.01$ over $\rho \in [-1, 1]$. Furthermore, (3.44) indicates that $\mathrm{E}[\gamma_i]$ is monotonically increasing with both θ and δ. Thus, we obtain a universal bound

$$\mathrm{E}[\gamma_i] \le 0.13 \quad \text{for } \theta < 0.1 \text{ and } \delta = 0.01, \tag{3.46}$$

which further implies $\mathrm{E}[\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2)] \leq 0.13\|\boldsymbol{h}\|^2$ for $\theta < 0.1$ and $\delta = 0.01$.

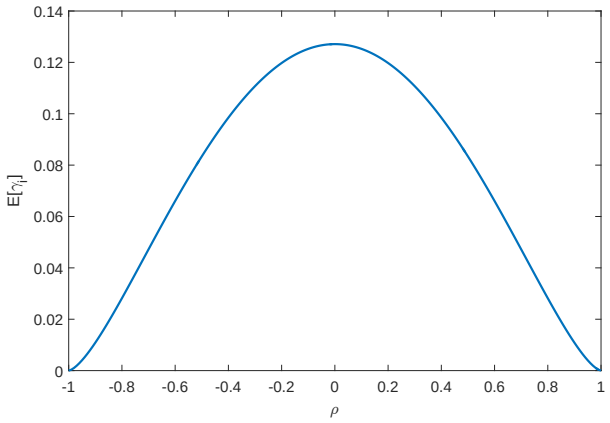


Fig. 3.7: $\mathrm{E}[\gamma_i]$ with respect to ρ

Furthermore, $\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2)$'s are sub-exponential with sub-exponential norm $\mathcal{O}(\|\boldsymbol{h}\|^2)$. By the sub-exponential tail bound (Bernstein type) [142], we have

$$\mathcal{P}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2)}{\|\boldsymbol{h}\|^2} > (0.13+\epsilon)\right] < \exp(-cm\epsilon^2), \tag{3.47}$$

for some universal constant c, as long as $\|\boldsymbol{h}\| \leq \frac{1}{10}\|\boldsymbol{x}\|$.

We have proved so far that the claim holds for a fixed $\boldsymbol{h}$. We next obtain a uniform bound over all $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\| \leq \frac{1}{10}\|\boldsymbol{x}\|$. We first show the claim holds for all $\boldsymbol{h}$ with $\|\boldsymbol{h}\| = \frac{1}{10}\|\boldsymbol{x}\|$ and then argue the claim holds when $\|\boldsymbol{h}\| < \frac{1}{10}\|\boldsymbol{x}\|$ towards the end of the proof. Let $\epsilon' = \epsilon\frac{\|\boldsymbol{x}\|}{10}$ and we construct an $\epsilon'-$net $\mathcal{N}_{\epsilon'}$ covering the sphere with radius $\frac{1}{10}\|\boldsymbol{x}\|$ in $\mathbb{R}^n$ with cardinality $|\mathcal{N}_{\epsilon'}| \leq (1+\frac{2}{\epsilon})^n$. Then for any $\|\boldsymbol{h}\| = \frac{1}{10}\|\boldsymbol{x}\|$, there exists a $\boldsymbol{h}_0 \in \mathcal{N}_{\epsilon'}$ such that $\|\boldsymbol{h}-\boldsymbol{h}_0\| \leq \epsilon\|\boldsymbol{h}\|$. Taking the union bound for all the points on the net, we claim that

$$\frac{1}{m}\sum_{i=1}^{m}\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2\right) \leq (0.13+\epsilon)\,\|\boldsymbol{h}_0\|^2, \quad \forall \boldsymbol{h}_0 \in \mathcal{N}_{\epsilon'} \tag{3.48}$$

holds with probability at least $1-(1+2/\epsilon)^n\exp(-cm\epsilon^2)$.

Since $\chi_i(t)$'s are Lipschitz functions with constant $1/\delta$, we have the following bound

$$\left|\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2) - \chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2)\right| \le \frac{1}{\delta}\left||\boldsymbol{a}_i^T\boldsymbol{h}|^2 - |\boldsymbol{a}_i^T\boldsymbol{h}_0|^2\right|. \tag{3.49}$$

Moreover, by [61, Lemma 1], we have

$$\frac{1}{m}\|\mathcal{A}(\boldsymbol{M})\|_1 \le c_2\|\boldsymbol{M}\|_F, \qquad \text{for all symmetric rank-2 matrices } \boldsymbol{M} \in \mathbb{R}^{n\times n}, \tag{3.50}$$

holds with probability at least $1 - C\exp(-c_1 m)$ as long as $m > c_0 n$ for some constants $C, c_0, c_1, c_2 > 0$. Consequently, on the event that (3.50) holds, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^m \chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}|^2\right) - \frac{1}{m}\sum_{i=1}^m \chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2\right)\right| \\
&\le \frac{1}{m}\sum_{i=1}^m \left|\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}|^2\right) - \chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}_0|^2\right)\right| \\
&\le \frac{1}{\delta}\cdot\frac{1}{m}\|\mathcal{A}(\boldsymbol{h}\boldsymbol{h}^T - \boldsymbol{h}_0\boldsymbol{h}_0^T)\|_1 \qquad \text{because of (3.49)} \\
&\le \frac{1}{\delta}\cdot c_2\|\boldsymbol{h}\boldsymbol{h}^T - \boldsymbol{h}_0\boldsymbol{h}_0^T\|_F \qquad \text{because of (3.50)} \\
&\le \frac{1}{\delta}\cdot 3c_2\|\boldsymbol{h} - \boldsymbol{h}_0\|\cdot\|\boldsymbol{h}\| \le 3c_3\epsilon/\delta\|\boldsymbol{h}\|^2,
\end{aligned}$$

where the last inequality is due to the Lemma 2 in [61].

On the intersection of events that (3.48) and (3.50) hold, we have

$$\frac{1}{m}\sum_{i=1}^m \chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}|^2\right) \le (0.13 + \epsilon + 3c_3\epsilon/\delta)\,\|\boldsymbol{h}\|^2, \tag{3.51}$$

for all $\boldsymbol{h}$ with $\|\boldsymbol{h}\| = \frac{1}{10}\|\boldsymbol{x}\|$.

For the case when $\|\boldsymbol{h}'\| < \frac{1}{10}\|\boldsymbol{x}\|$, $\boldsymbol{h}' = \omega\boldsymbol{h}$ for some $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\| = \frac{1}{10}\|\boldsymbol{x}\|$ and $0 < \omega < 1$. By the definition of $\chi_i(\cdot)$, it can be verified that

$$\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}'|^2) = \chi_i(|\boldsymbol{a}_i^T(\omega\boldsymbol{h})|^2) \le \omega^2\chi_i(|\boldsymbol{a}_i^T\boldsymbol{h}|^2). \tag{3.52}$$

Applying (3.51), on the same event that (3.48) and (3.50) hold, we have

$$\frac{1}{m}\sum_{i=1}^{m}\chi_i\left(|\boldsymbol{a}_i^T\boldsymbol{h}'|^2\right)\le(0.13+\epsilon+3c_3\epsilon/\delta)\,\|\boldsymbol{h}'\|^2, \tag{3.53}$$

for all $\|\boldsymbol{h}'\|<\frac{1}{10}\|\boldsymbol{x}\|$. Since ϵ can be arbitrarily small, the proof is completed. □

Therefore, combining Lemmas 3.6 and 3.7 with (3.37) yields

$$\langle\nabla\ell(\boldsymbol{z}),\boldsymbol{h}\rangle\ge(1-0.26-2\epsilon)\|\boldsymbol{h}\|^2=(0.74-2\epsilon)\|\boldsymbol{h}\|^2. \tag{3.54}$$

We further provide an upper bound on $\|\nabla\ell(\boldsymbol{z})\|$ in the following lemma.

Lemma 3.8. *Fix $\delta>0$, and assume $y_i=|\boldsymbol{a}_i^T\boldsymbol{x}|$. Suppose that $m\ge c_0n$ for a certain constant $c_0>0$. There exist some universal constants $c,C>0$ such that with probability at least $1-C\exp(-cm)$,*

$$\|\nabla\ell(\boldsymbol{z})\|\le(1+\delta)\cdot2\|\boldsymbol{h}\| \tag{3.55}$$

holds for all non-zero vectors $\boldsymbol{h},\boldsymbol{z}\in\mathbb{R}^n$ satisfying $\boldsymbol{z}=\boldsymbol{x}+\boldsymbol{h}$ and $\frac{\|\boldsymbol{h}\|}{\|\boldsymbol{x}\|}\le\frac{1}{10}$.

Proof. Denote $v_i:=\boldsymbol{a}_i^T\boldsymbol{z}-|\boldsymbol{a}_i^T\boldsymbol{x}|\mathrm{sgn}(\boldsymbol{a}_i^T\boldsymbol{z})$. Then

$$\nabla\ell(\boldsymbol{z})=\frac{1}{m}\boldsymbol{A}^T\boldsymbol{v}, \tag{3.56}$$

where $\boldsymbol{A}$ is a matrix with each row being $\boldsymbol{a}_i^T$ and $\boldsymbol{v}$ is a $m-$dimensional vector with each entry being v_i. Thus,

$$\|\nabla\ell(\boldsymbol{z})\|=\left\|\frac{1}{m}\boldsymbol{A}^T\boldsymbol{v}\right\|\le\frac{1}{m}\|\boldsymbol{A}\|\cdot\|\boldsymbol{v}\|\le(1+\delta)\frac{\|\boldsymbol{v}\|}{\sqrt{m}} \tag{3.57}$$

as long as $m\ge c_1n$ for some sufficiently large $c_1>0$, where the spectral norm bound $\|\boldsymbol{A}\|\le\sqrt{m}(1+\delta)$ follows from [142, Theorem 5.32].

We next bound $\|\boldsymbol{v}\|$. Let $\boldsymbol{v} = \boldsymbol{v}^{(1)} + \boldsymbol{v}^{(2)}$, where $v_i^{(1)} = \boldsymbol{a}_i^T \boldsymbol{h}$ and $v_i^{(2)} = 2\boldsymbol{a}_i^T \boldsymbol{x} \mathbf{1}_{\{(\boldsymbol{a}_i^T \boldsymbol{z})(\boldsymbol{a}_i^T \boldsymbol{x})<0\}}$. By triangle inequality, we have $\|\boldsymbol{v}\| \leq \|\boldsymbol{v}^{(1)}\| + \|\boldsymbol{v}^{(2)}\|$. Furthermore, given $m > c_0 n$, by [45, Lemma 3.1] with probability $1 - \exp(-cm)$, we have

$$\frac{1}{m}\|\boldsymbol{v}^{(1)}\|^2 = \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T \boldsymbol{h})^2 \leq (1+\delta)\|\boldsymbol{h}\|^2. \tag{3.58}$$

By Lemma 3.7, we have with probability $1 - C\exp(-c_1 m)$

$$\frac{1}{m}\|\boldsymbol{v}^{(2)}\|^2 = \frac{1}{m}\sum_{i=1}^{m} 4(\boldsymbol{a}_i^T \boldsymbol{x})^2 \cdot \mathbf{1}_{\{(\boldsymbol{a}_i^T \boldsymbol{x})(\boldsymbol{a}_i^T \boldsymbol{z})<0\}} \leq 4(0.13+\epsilon)\|\boldsymbol{h}\|^2. \tag{3.59}$$

Hence,

$$\frac{\|\boldsymbol{v}\|}{\sqrt{m}} \leq [\sqrt{1+\delta} + 2\sqrt{0.13+\epsilon}]\|\boldsymbol{h}\|. \tag{3.60}$$

This concludes the proof. □

Thus, applying Lemma 3.8 to (3.54), we conclude that *Regularity Condition* (3.35) holds for μ and λ satisfying

$$0.74 - 2\epsilon \geq \frac{\mu}{2} \cdot 4(1+\delta)^2 + \frac{\lambda}{2}, \tag{3.61}$$

which concludes the proof.

We note that (3.61) implies an upper bound $\mu \leq \frac{0.74}{2} = 0.37$, by taking ϵ and δ to be sufficiently small. This suggests a range to set the step size in Algorithm 1. However, in practice, μ can be set much larger than such a bound, say 0.8, while still keeping the algorithm convergent. This is because the coefficients in the proof are set for convenience of proof rather than being tightly chosen.

3.7.4 Proof of Theorem 3.3: Stability to Bounded Noise

We consider the model (3.12) with bounded noise, i.e., $y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle| + w_i$ for $i = 1, \cdots, m$. The initialization analysis is similar to Section 3.7.1. To analyze the gradient loop, we consider two regimes.

• **Regime 1**: $c_4\|\boldsymbol{z}\| \geq \|\boldsymbol{h}\| \geq c_3\frac{\|\boldsymbol{w}\|}{\sqrt{m}}$. In this regime, error contraction by each gradient step is given by

$$\text{dist}\left(\boldsymbol{z} - \mu\nabla\ell(\boldsymbol{z}), \boldsymbol{x}\right) \leq (1-\rho)\text{dist}(\boldsymbol{z}, \boldsymbol{x}). \tag{3.62}$$

It suffices to justify that $\nabla\ell(\boldsymbol{z})$ satisfies the RC. We have

$$\begin{aligned}\nabla\ell(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z} - y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i \\ &= \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z} - |\boldsymbol{a}_i^T\boldsymbol{x}|\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i}_{\nabla^{clean}\ell(\boldsymbol{z})} - \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(w_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i}_{\nabla^{noise}\ell(\boldsymbol{z})}.\end{aligned} \tag{3.63}$$

All the proofs for Lemma 3.6, 3.7 and 3.8 are still valid for $\nabla^{clean}\ell(\boldsymbol{z})$, and thus we have

$$\langle\nabla^{clean}\ell(\boldsymbol{z}), \boldsymbol{h}\rangle \geq 0.74\|\boldsymbol{h}\|^2, \tag{3.64}$$

$$\left\|\nabla^{clean}\ell(\boldsymbol{z})\right\| \leq 2(1+\delta)\|\boldsymbol{h}\|. \tag{3.65}$$

Next, we analyze the contribution of the noise. Let $\tilde{w}_i = w_i\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}$, and then for sufficient large m/n, we have

$$\|\nabla^{noise}\ell(\boldsymbol{z})\| = \left\|\frac{1}{m}\boldsymbol{A}^T\tilde{\boldsymbol{w}}\right\| \leq \left\|\frac{1}{\sqrt{m}}\boldsymbol{A}^T\right\|\left\|\frac{\tilde{\boldsymbol{w}}}{\sqrt{m}}\right\| \leq (1+\delta)\frac{\|\tilde{\boldsymbol{w}}\|}{\sqrt{m}} \leq (1+\delta)\frac{\|\boldsymbol{w}\|}{\sqrt{m}}, \tag{3.66}$$

where the second inequality is because the spectral norm bound $\|\boldsymbol{A}\| \leq \sqrt{m}(1+\delta)$ following from [142, Theorem 5.32]. Given the regime condition $\|\boldsymbol{h}\| \geq c_3 \frac{\|\boldsymbol{w}\|}{\sqrt{m}}$, we further have

$$\|\nabla^{noise}\ell(\boldsymbol{z})\| \leq \frac{(1+\delta)}{c_3}\|\boldsymbol{h}\|, \tag{3.67}$$

$$\left|\left\langle \nabla^{noise}\ell(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \leq \left\|\nabla^{noise}\ell(\boldsymbol{z})\right\| \cdot \|\boldsymbol{h}\| \leq \frac{(1+\delta)}{c_3}\|\boldsymbol{h}\|^2. \tag{3.68}$$

Combining these together, one has

$$\langle \nabla\ell(\boldsymbol{z}), \boldsymbol{h}\rangle \geq \left\langle \nabla^{clean}\ell(\boldsymbol{z}), \boldsymbol{h}\right\rangle - \left|\left\langle \nabla^{noise}\ell(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \geq \left(0.74 - \frac{(1+\delta)}{c_3}\right)\|\boldsymbol{h}\|^2, \tag{3.69}$$

and

$$\|\nabla\ell(\boldsymbol{z})\| \leq \left\|\nabla^{clean}\ell(\boldsymbol{z})\right\| + \left\|\nabla^{noise}\ell(\boldsymbol{z})\right\| \leq (1+\delta)\left(2 + \frac{1}{c_3}\right)\|\boldsymbol{h}\|. \tag{3.70}$$

The RC is guaranteed if μ, λ, ϵ are chosen properly, c_3 is sufficiently large, and s is sufficiently small.

• **Regime 2**: Once the iterate enters the regime with $\|\boldsymbol{h}\| \leq \frac{c_3\|\boldsymbol{w}\|}{\sqrt{m}}$, gradient update may not reduce the estimation error. However, in this regime, each move size $\mu\nabla\ell(\boldsymbol{z})$ is at most $\mathcal{O}(\|\boldsymbol{w}\|/\sqrt{m})$. Then the estimation error cannot increase by more than $\|\boldsymbol{w}\|/\sqrt{m}$ with a constant factor. Thus, one has

$$\text{dist}\left(\boldsymbol{z} + \mu\nabla\ell(\boldsymbol{z}), \boldsymbol{x}\right) \leq c_5\frac{\|\boldsymbol{w}\|}{\sqrt{m}} \tag{3.71}$$

for some constant c_5. As long as $\|\boldsymbol{w}\|/\sqrt{m}$ is sufficiently small, it is guaranteed that $c_5\frac{\|\boldsymbol{w}\|}{\sqrt{m}} \leq c_4\|\boldsymbol{x}\|$. If the iterate jumps out of *Regime 2*, it falls into *Regime 1*.

3.7.5 Proof of Theorem 3.4: Mini-batch IRWF

Since the initialization is the same as that in Algorithm 1, it suffices to show the convergence of gradient loops given that the initial point lands into the neighborhood of global minima. To prove Theorem 3.4, the major step is to prove the following Proposition 3.9 which characterizes how the error of an estimate decays upon one iteration of Algorithm 2. Once Proposition 3.9 is established, we take expectation on both sides of (3.73) with respect to i_{t-1}, and apply Proposition 3.9 one more time to obtain

$$\mathrm{E}_{\{i_{t-1},i_t\}}\left[\mathrm{dist}^2(\boldsymbol{z}^{(t+1)},\boldsymbol{x})\right] \le \left(1-\frac{k\rho}{n}\right)^2 \mathrm{dist}^2(\boldsymbol{z}^{(t-1)},\boldsymbol{x}). \tag{3.72}$$

Continuing this process until the initialization point $\boldsymbol{z}^{(0)}$ yields Theorem 3.4. We next focus on proving Proposition 3.9 stated bellow.

Proposition 3.9. *Assume the measurement vectors are independent and each $\boldsymbol{a}_i \sim \mathcal{N}(0,\boldsymbol{I})$. There exist some universal constants $0<\rho,\rho_0<1$ and $c_0,c_1,c_2>0$ such that if $m \ge c_0 n$ and $\mu=\rho_0/n$ for the update rule* (3.14)*, then with probability at least $1-c_1\exp(-c_2 m)$, we have*

$$\mathrm{E}_{\Gamma_t}\left[\mathit{dist}^2(\boldsymbol{z}^{(t+1)},\boldsymbol{x})\right] \le \left(1-\frac{k\rho}{n}\right)\cdot \mathit{dist}^2(\boldsymbol{z}^{(t)},\boldsymbol{x}) \tag{3.73}$$

to hold for all $\boldsymbol{z}^{(t)}$ satisfying $\frac{\mathit{dist}(\boldsymbol{z}^{(t)},\boldsymbol{x})}{\|\boldsymbol{z}\|} \le \frac{1}{10}$.

Proof. Without loss of generality, we assume $\boldsymbol{z}^{(t)}$ is in the neighborhood of $\boldsymbol{x}$ (otherwise it is in the neighborhood of $-\boldsymbol{x}$). Let $\boldsymbol{h}=\boldsymbol{z}^{(t)}-\boldsymbol{x}$.

We follow the notations in Section 3.7.3 and let $\mathcal{S}=\{i:(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)})<0\}$. Then

we have

$$
\begin{aligned}
&\mathrm{E}_{\Gamma_t}\left[\mathrm{dist}^2\left(\boldsymbol{z}^{(t+1)},\boldsymbol{x}\right)\right]\\
&=\mathrm{E}_{\Gamma_t}\left[\left\|\boldsymbol{z}^{(t)}-\mu\boldsymbol{A}_{\Gamma_t}^T\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)}-y_{\Gamma_t}\odot\mathrm{sgn}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right)-\boldsymbol{x}\right\|^2\right]\\
&=\|\boldsymbol{h}\|^2-2\mu\mathrm{E}_{\Gamma_t}\left[\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)}-y_{\Gamma_t}\odot\mathrm{sgn}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right)^T\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{h}\right)\right]\\
&\quad+\mu^2\mathrm{E}_{\Gamma_t}\left[\left(\boldsymbol{A}_{\Gamma_t}^T\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)}-y_{\Gamma_t}\odot\mathrm{sgn}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right)\right)^T\left(\boldsymbol{A}_{\Gamma_t}^T\left(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)}-y_{\Gamma_t}\odot\mathrm{sgn}(\boldsymbol{A}_{\Gamma_t}\boldsymbol{z}^{(t)})\right)\right)\right]\\
&\overset{(a)}{=}\|\boldsymbol{h}\|^2-\frac{2\mu k}{m}\sum_{i=1}^m\left[\boldsymbol{a}_i^T\boldsymbol{h}\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}-y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)\right]\\
&\quad+\frac{\mu^2k}{m}\sum_{i=1}^m\left[\|\boldsymbol{a}_i\|^2\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}-y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)^2\right]\\
&=\|\boldsymbol{h}\|^2-\frac{2\mu k}{m}\left(\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{h})^2+\sum_{i\in\mathcal{S}}2(\boldsymbol{a}_i^T\boldsymbol{h})(\boldsymbol{a}_i^T\boldsymbol{x})\right)\\
&\quad+\frac{\mu^2k}{m}\left(\sum_{i=1}^m\|\boldsymbol{a}_i\|^2(\boldsymbol{a}_i^T\boldsymbol{h})^2+4\sum_{i\in\mathcal{S}}\|\boldsymbol{a}_i\|^2(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)})\right)\\
&\le\|\boldsymbol{h}\|^2-\frac{2\mu k}{m}\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{h})^2+\frac{4\mu k}{m}\sum_{i\in\mathcal{S}}\left|(\boldsymbol{a}_i^T\boldsymbol{h})(\boldsymbol{a}_i^T\boldsymbol{x})\right|+\frac{\mu^2k}{m}\sum_{i=1}^m\|\boldsymbol{a}_i\|^2(\boldsymbol{a}_i^T\boldsymbol{h})^2, \qquad (3.74)
\end{aligned}
$$

where (a) is due to the fact that Γ_t is uniformly chosen from all subsets of $\{1,2,\ldots,m\}$ with cardinality k.

By Lemma 3.6, we have that if $m\ge c_0\epsilon^{-2}n$ with probability $1-2\exp(-c_1m\epsilon^2)$

$$(1-\epsilon)\|\boldsymbol{h}\|^2\le\frac{1}{m}\sum_{i=1}^m(\boldsymbol{a}_i^T\boldsymbol{h})^2\le(1+\epsilon)\|\boldsymbol{h}\|^2.$$

holds for all vectors $\boldsymbol{h}$. By Lemma 3.7, we have that with probability $1-C\exp(-c_1m\epsilon^2)$

$$\frac{1}{m}\sum_{i\in\mathcal{S}}\left|(\boldsymbol{a}_i^T\boldsymbol{h})(\boldsymbol{a}_i^T\boldsymbol{x})\right|\le(0.13+\epsilon)\|\boldsymbol{h}\|^2$$

holds for all $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\|/\|\boldsymbol{x}\|\le\frac{1}{10}$.

Define an event $E_1:=\{\max_{1\le i\le m}\|\boldsymbol{a}_i\|^2\le 6n\}$. We claim that E_1 holds with proba-

bility $1 - m\exp(-1.5n)$. Then on the event E_1, (3.74) is further upper bounded by

$$\begin{aligned}\mathrm{E}_{\Gamma_t}\left[\mathrm{dist}^2\left(\boldsymbol{z}^{(t+1)},\boldsymbol{x}\right)\right] &\le \left(1-2\mu k(1-\epsilon)+4\mu k(0.13+\epsilon)+\mu^2 k\cdot 6n(1+\epsilon)\right)\|\boldsymbol{h}\|^2 \\ &\le \left(1-2\mu k(0.74-3\epsilon-3n\mu(1+\epsilon))\right)\|\boldsymbol{h}\|^2. \end{aligned} \tag{3.75}$$

By choosing the step size $\mu \le \frac{0.24}{n}$, the proposition is proved. □

3.7.6 Proof of Theorem 3.5: Kaczmarz-PR Algorithm

Without loss of generality, we assume $\boldsymbol{z}^{(t)}$ is in the neighborhood of $\boldsymbol{x}$ (otherwise it is in the neighborhood of $-\boldsymbol{x}$). Let $\boldsymbol{h} = \boldsymbol{z}^{(t)} - \boldsymbol{x}$.

We follow the notations in Section 3.7.3 and let $\mathcal{S} = \{i : (\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}) < 0\}$. Then we have

$$\begin{aligned}\mathrm{E}_{i_t}\left[\mathrm{dist}^2\left(\boldsymbol{z}^{(t+1)},\boldsymbol{x}\right)\right] &= \mathrm{E}_{i_t}\left[\left\|\left(\boldsymbol{z}^{(t)}-\frac{1}{\|\boldsymbol{a}_{i_t}\|^2}\left(\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}-y_{i_t}\cdot\frac{\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}|}\right)\boldsymbol{a}_{i_t}\right)-\boldsymbol{x}\right\|^2\right] \\ &= \|\boldsymbol{h}\|^2 - 2\mathrm{E}_{i_t}\left[\frac{1}{\|\boldsymbol{a}_{i_t}\|^2}\left(\boldsymbol{a}_{i_t}^T\boldsymbol{h}\right)\left(\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}-y_{i_t}\cdot\frac{\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}|}\right)\right] \\ &\quad + \mathrm{E}_{i_t}\left[\frac{1}{\|\boldsymbol{a}_{i_t}\|^2}\left(\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}-y_{i_t}\cdot\frac{\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_{i_t}^T\boldsymbol{z}^{(t)}|}\right)^2\right] \\ &\overset{(a)}{=} \|\boldsymbol{h}\|^2 - \frac{2}{m}\sum_{i=1}^m\left[\frac{1}{\|\boldsymbol{a}_i\|^2}\left(\boldsymbol{a}_i^T\boldsymbol{h}\right)\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}-y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)\right] \\ &\quad + \frac{1}{m}\sum_{i=1}^m\left[\frac{1}{\|\boldsymbol{a}_i\|^2}\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}-y_i\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)^2\right] \\ &= \|\boldsymbol{h}\|^2 - \frac{2}{m}\left(\sum_{i=1}^m\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2}+\sum_{i\in\mathcal{S}}\frac{2(\boldsymbol{a}_i^T\boldsymbol{h})(\boldsymbol{a}_i^T\boldsymbol{x})}{\|\boldsymbol{a}_i\|^2}\right) \\ &\quad + \frac{1}{m}\left(\sum_{i=1}^m\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2}+4\sum_{i\in\mathcal{S}}\frac{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)})}{\|\boldsymbol{a}_i\|^2}\right) \\ &= \|\boldsymbol{h}\|^2 - \frac{1}{m}\sum_{i=1}^m\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2}+\frac{4}{m}\sum_{i\in\mathcal{S}}\frac{(\boldsymbol{a}_i^T\boldsymbol{x})^2}{\|\boldsymbol{a}_i\|^2} \end{aligned} \tag{3.76}$$

where (a) is due to the fact that i_t is sampled uniformly at random from $\{1, 2, \cdots, m\}$. By the spectral case of Lemma 5.20 in [142], $\{\sqrt{n}\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|}\}_{i=1}^m$ are independent isotropic random vectors in $\mathbb{R}^n$ and hence

$$\mathrm{E}\left[n\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2}\right] = \|\boldsymbol{h}\|^2.$$

Moreover, $\{\sqrt{n}\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|}\}_{i=1}^m$ are sub-Gaussian and the sub-Gaussian norm is bounded by an absolute constant. Thus, we have that if $m \geq c_0\epsilon^{-2}n$, then with probability $1-2\exp(-c_1m\epsilon^2)$,

$$\frac{1}{m}\sum_{i=1}^m \frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\|\boldsymbol{a}_i\|^2} \geq \frac{(1-\epsilon)}{n}\|\boldsymbol{h}\|^2.$$

holds for all vectors $\boldsymbol{h}$. By Lemma 3.7, we have that with probability $1 - C\exp(-c_1m\epsilon^2)$

$$\frac{1}{m}\sum_{i\in\mathcal{S}}\left|\boldsymbol{a}_i^T\boldsymbol{x}\right|^2 \leq \frac{1}{m}\sum_{i=1}^m\left|\boldsymbol{a}_i^T\boldsymbol{h}\right|^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \leq (0.13+\epsilon)\|\boldsymbol{h}\|^2$$

holds for all $\boldsymbol{h}$ satisfying $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| \leq \frac{1}{10}$.

Define an event $E_2 := \{\min_{1\leq i\leq m}\|\boldsymbol{a}_i\|^2 \geq \frac{2}{3}n\}$. It can be shown that $\mathbb{P}\{E_2\} \geq 1 - m\exp(-n/12)$. Then on the event E_2, (3.76) is further upper bounded by

$$\mathrm{E}_{i_t}\left[\mathrm{dist}^2\left(\boldsymbol{z}^{(t+1)},\boldsymbol{x}\right)\right] \leq \left(1 - \frac{1-\epsilon}{n} + \frac{6(0.13+\epsilon)}{n}\right)\|\boldsymbol{h}\|^2 \leq \left(1 - \frac{0.22-7\epsilon}{n}\right)\|\boldsymbol{h}\|^2,$$

which concludes the proof.

CHAPTER 4
ROBUST PHASE RETRIEVAL: MEDIAN TRUNCATION APPROACH

In this chapter, we study the case when the observations of phase retrieval are corrupted by sparse outliers. Section 4.1 provides the problem formulation. Section 4.2 describes two algorithms, median-TWF and median-RWF. Section 4.3 provides their performance guarantees. Section 4.4 presents the numerical experiments. Section 4.6 includes supplemental proofs.

Throughout this chapter, boldface lowercase letters such as $\boldsymbol{a}_i, \boldsymbol{x}, \boldsymbol{z}$ denote vectors, and boldface capital letters such as $\boldsymbol{A}, \boldsymbol{Y}$ denote matrices. For two matrices, $\boldsymbol{A} \preceq \boldsymbol{B}$ means that $\boldsymbol{B} - \boldsymbol{A}$ is positive definite. For a complex matrix or vector, $\boldsymbol{A}^*$ and $\boldsymbol{z}^*$ denote conjugate transposes of $\boldsymbol{A}$ and $\boldsymbol{z}$ respectively. For a real matrix or vector, $\boldsymbol{A}^T$ and $\boldsymbol{z}^T$ denote transposes of $\boldsymbol{A}$ and $\boldsymbol{z}$ respectively. The indicator function $\mathbf{1}_A = 1$ if the event A is true, and $\mathbf{1}_A = 0$ otherwise.

4.1 Problem Formulation

As we discuss in the Chapter 1, the ability to handle outliers is of great importance for phase retrieval algorithms, because outliers arise frequently from the phase imaging applications [71] due to various reasons such as detector failures, recording errors, and missing data. However, the performance of WF, TWF and RWF can be very sensitive to outliers that take arbitrary values and can introduce anomalous search directions. Even for TWF, since the sample mean can be arbitrarily perturbed, the truncation rule based on such sample mean cannot control the gradient well.

Mathematically, suppose the observations are given by

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 + \eta_i, \quad i = 1, \cdots, m, \tag{4.1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is the unknown signal,[1] $\boldsymbol{a}_i \in \mathbb{R}^n$ for $i = 1, \ldots, m$ are measurement vectors with each $\boldsymbol{a}_i$ having $i.i.d.$ Gaussian entries distributed as $\mathcal{N}(0, 1)$, and $\eta_i \in \mathbb{R}$ for $i = 1, \ldots, m$ are outliers with arbitrary values. We assume that outliers are sparse with sm nonzero values, i.e., $\|\boldsymbol{\eta}\|_0 \leq sm$, where $\boldsymbol{\eta} = \{\eta_i\}_{i=1}^m \in \mathbb{R}^m$. Here, s is a nonzero constant, representing the faction of measurements that are corrupted by outliers.

We are also interested in the model when the measurements are corrupted by not only sparse arbitrary outliers but also dense bounded noise. Under such a model, the measurements are given by

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 + w_i + \eta_i, \quad i = 1, \cdots, m, \tag{4.2}$$

where the bounded noise $\boldsymbol{w} = \{w_i\}_{i=1}^m$ satisfies $\|\boldsymbol{w}\|_\infty \leq c_1 \|\boldsymbol{x}\|^2$ for some universal constant c_1, and as before, the outliers satisfy $\|\boldsymbol{\eta}\|_0 \leq sm$.

The goal is to recover the signal $\boldsymbol{x}$ (up to a global sign difference) from the measure-

[1] We focus on real signals here, but our analysis can be extended to complex signals.

ments $\boldsymbol{y} = \{y_i\}_{i=1}^m$ and measurement vectors $\{\boldsymbol{a}_i\}_{i=1}^m$. It can be observed that if $\boldsymbol{z}$ is a solution, then $-\boldsymbol{z}$ is also the solution of the problem. Thus, the recovery is up to a sign difference. We define the Euclidean distance between two vectors up to a global sign difference [1] as,

$$\text{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min \|\boldsymbol{z} \pm \boldsymbol{x}\|. \tag{4.3}$$

In this chapter, we develop non-convex phase retrieval algorithms with both statistical and computational efficiency, and provable robustness to even a constant proportion of outliers. We hope that the algorithm has three folds of properties. First, it should recover the true signal as long as the number of outliers is not too large. Moreover, the algorithm does not have to require prior knowledge about the outliers. At last, the algorithm should be efficient in terms of sample complexity and convergence rate.

Our strategy is to use sample median as truncation thresholds to eliminate the contribution of bad samples. The robustness property of median lies in the fact that the median cannot be arbitrarily perturbed unless the outliers dominate the inliers [73]. This is in sharp contrast to the mean, which can be made arbitrarily large even by a single outlier. Thus, using the sample median in the truncation rule can effectively remove the impact of outliers. By applying median truncation strategy to modify TWF and RWF, we obtain two new algorithms: median-TWF and median-RWF. These two algorithms share many similarities and also involve different aspects. To the best of the authors' knowledge, our work is the first application of the median to robustify high-dimensional statistical estimation in the presence of arbitrary outliers with rigorous non-asymptotic performance guarantees.

4.2 Median-based Algorithms

If some measurements are corrupted by outliers as in (4.1), then WF, RWF and TWF can fail. This is because the gradient of the loss function typically contains the term $|y_i - |\boldsymbol{a}_i^T \boldsymbol{z}|^2|$

or $\left|\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}|\right|$. With y_i being corrupted by arbitrarily large η_i, the gradient can deviate the search direction to the signal arbitrarily. In TWF, the truncation rule is based on the sample mean of the gradient, which can be affected significantly by outliers. It is not anticipated that TWF converges globally in the presence of arbitrary outliers.

To handle outliers, our central idea is to prune the samples in both the initialization and each gradient descent iteration via the *sample median* related quantities. Compared to the sample mean used to set truncation thresholds in TWF, the sample median is much less affected by outliers, and thus the algorithms are more robust in the presence of outliers.

The Poisson loss function adopted in TWF [61] models well the physical behavior of photons while the reshaped loss function used in RWF [68] is shown to have advantages on convergence rate. It is worthy to see whether the idea of the median truncation works for both loss functions. We apply the median truncation to TWF and RWF respectively and obtain two algorithms median-TWF and median-RWF.

The difference of median-TWF and median-RWF comes from the difference of TWF and RWF, which mainly lies in the different loss functions and that truncation is needed or not in gradient loop. Specifically, median-TWF employs the median of $\left|y_i - |\boldsymbol{a}_i^T \boldsymbol{z}|^2\right|$ to set the truncation threshold in gradient loop while median-RWF adopts the median of $\left|\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}|\right|$ to set the corresponding threshold. Empirically, median-TWF performs a little better than median-RWF in terms of sample complexity for real Gaussian measurements while median-RWF can tolerate more outliers than median-TWF. Another empirical fact is that median-RWF converges faster while median-TWF achieves better accuracy under the dense noise.

However, the two algorithms share similar properties. Both algorithms resist outliers in a oblivious fashion, which means we do not have to know the knowledge of outliers before running the algorithms. Moreover, the performance guarantees of median-TWF and median-RWF turn out to be almost the same except for different choices of constants, as presented in Section 4.3.

Algorithm 3 Median Truncated Wirtinger Flow (Median-TWF)

Input: $\boldsymbol{y} = \{y_i\}_{i=1}^m$, $\{\boldsymbol{a}_i\}_{i=1}^m$;
Parameters: thresholds α_y, α_h, α_l, and α_u, stepsize μ_t;
Initialization: Let $\boldsymbol{z}^{(0)} = \lambda_0 \tilde{\boldsymbol{z}}$, where $\lambda_0 = \sqrt{\text{med}(\boldsymbol{y})/0.455}$ and $\tilde{\boldsymbol{z}}$ is the leading eigenvector of

$$\boldsymbol{Y} := \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T \mathbf{1}_{\{|y_i| \le \alpha_y^2 \lambda_0^2\}}. \tag{4.5}$$

Gradient loop: for $t = 0 : T-1$ do

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{m}\sum_{i=1}^{m} \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2 - y_i}{\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}} \boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i}, \tag{4.6}$$

where

$$\begin{aligned}
\mathcal{E}_1^i &:= \left\{ \alpha_l \|\boldsymbol{z}^{(t)}\| \le |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}| \le \alpha_u \|\boldsymbol{z}^{(t)}\| \right\}, \\
\mathcal{E}_2^i &:= \left\{ |y_i - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2| \le \alpha_h K_t \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|}{\|\boldsymbol{z}^{(t)}\|} \right\}, \\
K_t &:= \text{med}\Big(\{ |y_i - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2| \}_{i=1}^m \Big).
\end{aligned}$$

Output $\boldsymbol{z}_T$.

The technical proofs of two algorithms follow the same structure. The crux is to use the median statistical properties to show that the median-trimmed gradient satisfy the so-called *Regularity Condition*, which guarantees the linear convergence of gradient update. We provide separate proofs for two algorithms because they involve different bounding techniques due to different loss functions.

4.2.1 Median-TWF Algorithm

In the following, we describe the median-TWF in details. We adopt the following Poisson loss function,

$$\ell(\boldsymbol{z}) := \frac{1}{2m}\sum_{i=1}^{m} \left(|\boldsymbol{a}_i^T \boldsymbol{z}|^2 - y_i \log |\boldsymbol{a}_i^T \boldsymbol{z}|^2 \right). \tag{4.4}$$

Median-TWF algorithm (summarized in Algorithm 3) minimizes (4.4) via an initializa-

tion step and gradient descent.

1. **Initialization**: We initialize $\boldsymbol{z}^{(0)}$ by the spectral method with a truncated set of samples, where the threshold is determined by the median of $\{y_i\}_{i=1}^m$. In comparison, WF does not truncate samples, and the truncation in TWF is based on the mean of $\{y_i\}_{i=1}^m$, which is not robust to outliers. As will be shown, as long as the portion of outliers is not too large, our initialization (4.5) is guaranteed to be within a small neighborhood of the true signal.

2. **Gradient loop**: for each iteration $0 \le t \le T-1$, median-TWF uses an iteration-varying truncated gradient given as

$$\nabla \ell_{tr}(\boldsymbol{z}^{(t)}) = \frac{1}{m}\sum_{i=1}^{m} \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2 - y_i}{\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}} \boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i}. \tag{4.7}$$

It is clear from the definition of the set $\mathcal{E}_2^i$ (see Algorithm 3), that samples are truncated by the sample median of gradient components evaluated at the current iteration, as opposed to the sample mean in TWF.

We set the step size in the median-TWF to be a fixed small constant, i.e., $\mu = 0.4$. The rest of the parameters $\{\alpha_y, \alpha_h, \alpha_l, \alpha_u\}$ are set to satisfy

$$\begin{aligned}
&\zeta_1 := \max\left\{ \mathrm{E}\left[\xi^2 \mathbf{1}_{\left\{|\xi|<\sqrt{1.01}\alpha_l \text{ or } |\xi|>\sqrt{0.99}\alpha_u\right\}}\right], \mathrm{E}\left[\mathbf{1}_{\left\{|\xi|<\sqrt{1.01}\alpha_l \text{ or } |\xi|>\sqrt{0.99}\alpha_u\right\}}\right] \right\}, \\
&\zeta_2 := \mathrm{E}\left[\xi^2 \mathbf{1}_{\{|\xi|>0.248\alpha_h\}}\right], \\
&2(\zeta_1+\zeta_2) + \sqrt{8/\pi}\alpha_h^{-1} < 1.99 \\
&\alpha_y \ge 3,
\end{aligned} \tag{4.8}$$

where $\xi \sim \mathcal{N}(0,1)$. For example, we set $\alpha_l = 0.3, \alpha_u = 5, \alpha_y = 3$ and $\alpha_h = 12$, and consequently $\zeta_1 \approx 0.24$ and $\zeta_2 \approx 0.032$.

4.2.2 Median-RWF Algorithm

In this subsection, we apply the idea of median truncation to reshaped Wirtinger flow loss [68]

$$\mathcal{R}(\boldsymbol{z}) = \frac{1}{2m}\sum_{i=1}^{m}\left(\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}|\right)^2, \tag{4.9}$$

which has been shown to have advantage over Wirtinger flow loss (squared loss of squared measurements) and truncated-WF loss (Poisson loss of squared measurements). We call this new algorithm *median reshaped Wirtinger flow* (median-RWF). It uses median-based threshold to truncate the measurements in both initialization and gradient loop, as illustrated in Algorithm 4.

In the following, we discuss the median-RWF algorithm in detail.

1. **Initialization**: For simplicity, we here use the same initialization as in median-TWF (Algorithm 3).

2. **Gradient loop**: Median-RWF uses the following iteration-varying truncated gradient

$$\nabla\mathcal{R}_{tr}(\boldsymbol{z}^{(t)}) = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)} - \sqrt{y_i}\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i}, \tag{4.10}$$

From the definition of the set $\mathcal{T}^i$ (see Algorithm 4), it is clear that samples are truncated by the sample median of gradient components evaluated at the current iteration. We set $\frac{0}{0} = 0$ when calculating (4.11).

We set the step size in the median-RWF to be a fixed small constant, i.e., $\mu' = 0.8$. We set $\alpha_h' = 5$.

Algorithm 4 Median Reshaped Wirtinger Flow (median-RWF)

Input: $\boldsymbol{y} = \{y_i\}_{i=1}^m$, $\{\boldsymbol{a}_i\}_{i=1}^m$;

Parameters: thresholds α_y, α_h, and step size μ;
Initialization: Same as median-TWF (see Algorithm 3).
Gradient loop: for $t = 0 : T-1$ do

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu'}{m} \sum_{i=1}^{m} \left(\boldsymbol{a}_i^T \boldsymbol{z}^{(t)} - \sqrt{y_i} \cdot \frac{\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|} \right) \boldsymbol{a}_i \mathbf{1}_{\mathcal{T}^i}, \tag{4.11}$$

where

$$\mathcal{T}^i := \left\{ \left| \sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}| \right| \le \alpha_h' M_t \right\}, \quad \text{and} \quad M_t := \text{med}\left(\left\{ \left| \sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}| \right| \right\}_{i=1}^m \right).$$

Output $\boldsymbol{z}_T$.

4.3 Performance Guarantees

In this section, we characterize the performance guarantees of median-TWF and median-RWF. The theoretical guarantees for both algorithms are almost the same. Thus, in order to avoid repetition, we present the results in a way that works for both algorithms. However, the proofs for median-TWF and median-RWF involve different techniques and are shown separately.

We first show that median-TWF/median-RWF works well for the noise-free model in the following proposition, which lends support to the model with outliers. This also justifies that we can run median-TWF/median-RWF without having to know whether the underlying measurements are corrupted.

Proposition 4.1 (**Exact recovery for noise-free model**). *Suppose that the measurements are noise-free, i.e.,* $\eta_i = 0$ *for* $i = 1, \cdots, m$ *in the model* (4.1). *There exist constants* $\mu_0 > 0$, $0 < \rho, \nu < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \ge c_0 n \log n$ *and* $\mu \le \mu_0$, *then with probability at least* $1 - c_1 \exp(-c_2 m)$, *the median-TWF/median-RWF yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \le \nu (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \tag{4.12}$$

simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$.

Furthermore, as the median is quite stable as long as the number of outliers is not so large, the following theorem describes that median-TWF/median-RWF still works well even in the presence of sparse outliers.

Theorem 4.2 (**Exact recovery with sparse arbitrary outliers**). *Consider the phase retrieval problem with sparse outliers given in* (4.1). *There exist constants* μ_0, $s_0 > 0$, $0 < \rho, \nu < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \geq c_0 n \log n$, $s < s_0$, $\mu \leq \mu_0$, *then with probability at least* $1 - c_1 \exp(-c_2 m)$, *the median-TWF/median-RWF yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \leq \nu(1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \tag{4.13}$$

simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$.

Theorem 4.2 indicates that median-TWF/median-RWF admits exact recovery for *all* signals in the presence of sparse outliers with arbitrary magnitudes even when the number of outliers scales linearly with the number of measurements, as long as the number of samples satisfies $m \gtrsim n \log n$. This is near-optimal up to a logarithmic factor.

Moreover, median-TWF/median-RWF converges at a geometric rate using a constant step size, with per-iteration cost $O(mn)$ (note that the median can be computed in linear time [144]). To reach ϵ-accuracy, i.e., $\text{dist}(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \leq \epsilon$, only $O(\log 1/\epsilon)$ iterations are needed, and the total computational cost is $O(mn \log 1/\epsilon)$, which is highly efficient. Empirically in the experiments, median-RWF converges faster than median-TWF and median-RWF tolerates larger fraction of outliers than median-TWF, which may be due to that the lower-order model is more stable.

We next consider the model when the measurements are corrupted by both sparse arbitrary outliers and dense bounded noise. Our following theorem characterizes that median-TWF/median-RWF is robust to coexistence of the two types of noises.

Theorem 4.3 (**Stability to sparse arbitrary outliers and dense bounded noises**). *Consider the phase retrieval problem given in* (4.2) *in which measurements are corrupted by both sparse arbitrary and dense bounded noises. There exist constants* $\mu_0, s_0 > 0$, $0 < \rho < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \geq c_0 n \log n$, $s < s_0$, $\mu \leq \mu_0$, *then with probability at least* $1 - c_1 \exp(-c_2 m)$, *median-TWF/median-RWF respectively yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|} + (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \tag{4.14}$$

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \sqrt{\|\boldsymbol{w}\|_\infty} + (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \tag{4.15}$$

simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$.

Theorem 4.3 immediately implies the stability of median-TWF/median-RWF for the model corrupted only by dense bounded noise.

Corollary 4.4. *Consider the phase retrieval problem in which measurements are corrupted only by dense bounded noises, i.e.,* $\eta_i = 0$ *for* $i = 1, \cdots, m$ *in the model* (4.2). *There exist constants* $\mu_0 > 0$, $0 < \rho < 1$ *and* $c_0, c_1, c_2 > 0$ *such that if* $m \geq c_0 n \log n$, $\mu \leq \mu_0$, *then with probability at least* $1 - c_1 \exp(-c_2 m)$, *median-TWF/median-RWF respectively yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|} + (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \tag{4.16}$$

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \sqrt{\|\boldsymbol{w}\|_\infty} + (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \tag{4.17}$$

simultaneously for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$.

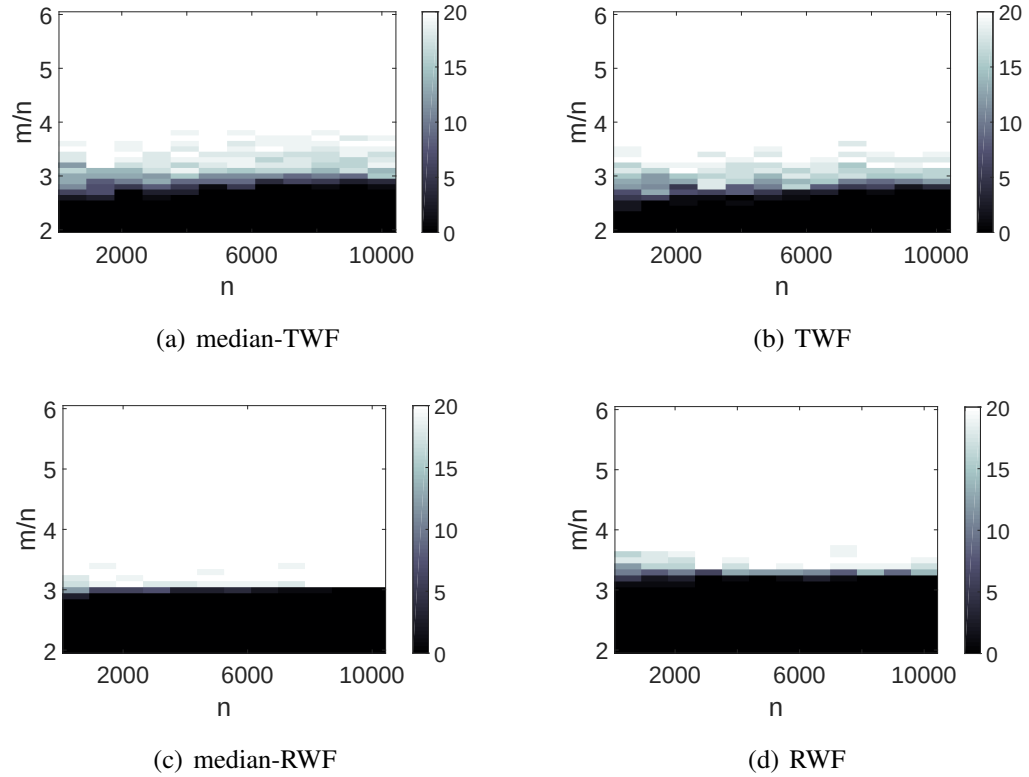
Thus, Theorem 4.3 and Corollary 4.4 imply that median-TWF/median-RWF for the model with both sparse arbitrary outliers and dense bounded noises achieves the same convergence rate and the same level of estimation error as the model with only bounded noise. In fact, together with Theorem 4.2 and Proposition 4.1, it can be seen that applying median-TWF/median-RWF does not require the knowledge of the noise corruption models. When

there do exist outliers, median-TWF/median-RWF achieves almost the same performance *as if the outliers do not exist*. Empirically, under dense noise, the median-TWF reaches better accuracy than median-RWF because of the delicate truncation rule.

4.4 Numerical Experiments

In this section, we provide numerical experiments to demonstrate the effectiveness of median-TWF and median-RWF, which corroborates with our theoretical findings.We first show that, in the noise-free case, the median-TWF and median-RWF perform similarly as TWF [61] for exact recovery. We set the parameters of median-TWF and median-RWF as specified in Section 4.2.1 and Section 4.2.2 respectively, and those of TWF and RWF as suggested in [61] and [68] respectively. Let the signal length n take values from 1000 to 10000 by a step size of 1000, and the ratio of the sample complexity to the signal length, m/n, take values from 2 to 6 by a step size of 0.1. For each pair of (m, n), we generate a signal $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{n\times n})$, and the measurement vectors $\boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{n\times n})$ i.i.d. for $i = 1, \ldots, m$. For three algorithms, a fixed number of iterations $T = 500$ are run, and the trial is declared successful if $\boldsymbol{z}^{(T)}$, the output of the algorithm, satisfies $\text{dist}(\boldsymbol{z}^{(T)}, \boldsymbol{x})/\|\boldsymbol{x}\| \leq 10^{-8}$. Figure 4.1 shows the number of successful trials out of 20 trials for both algorithms, with respect to m/n and n. It can be seen that for all three algorithms, as soon as m is above $4n$, exact recovery is achieved for both algorithms. Around the phase transition boundary, the performance of median-TWF is slightly worse than that of TWF, which is possibly due to the inefficiency of median compared to mean in the noise-free case [73]. The empirical sample complexity of median-RWF is slightly better than RWF because the truncation improves the stableness of RWF.

We next examine the performance of median-TWF and median-RWF in the presence of sparse outliers. We compare the performance of median-TWF and median-RWF with not only TWF but also an alternative which we call the *trimean-TWF*, based on replacing



(a) median-TWF (b) TWF

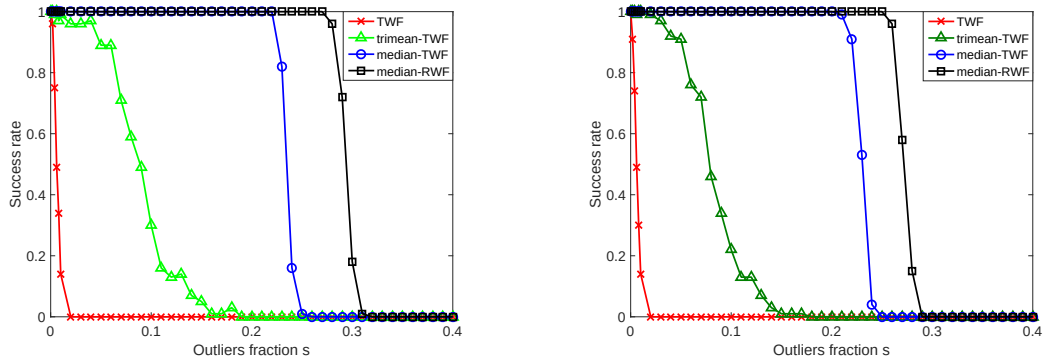
(c) median-RWF (d) RWF

Fig. 4.1: Sample complexity of median-TWF, median-RWF, RWF and TWF for noise-free data: the gray scale of each cell $(m/n, n)$ indicates the number of successful recovery out of 20 trials.

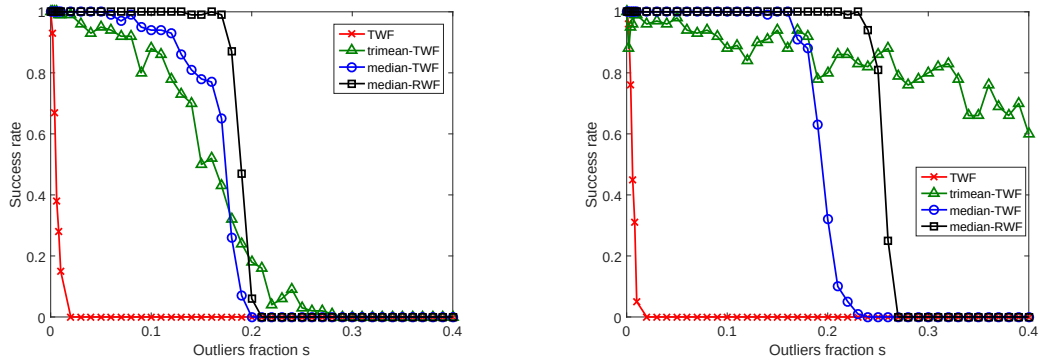
the sample mean in TWF by the *trimmed mean*. More specifically, trimean-TWF requires knowing the fraction s of outliers so that samples corresponding to sm largest gradient values are removed, and truncation is then based on the mean of remaining samples.

We fix the signal length $n = 1000$ and the number of measurements $m = 8000$. We assume each measurement y_i is corrupted with probability $s \in [0, 0.4]$ independently, where the corruption value $\eta_i \sim \mathcal{U}(0, \|\boldsymbol{\eta}\|_\infty)$ is randomly generated from a uniform distribution. Figure 4.2 shows the success rate of exact recovery over 100 trials as a function of s at different levels of outlier magnitudes $\|\boldsymbol{\eta}\|_\infty / \|\boldsymbol{x}\|^2 = 0.1, 1, 10, 100$, for the four algorithms median-TWF, median-RWF, trimean-TWF and TWF.

From Figure 4.2, it can be seen that median-TWF and median-RWF allow exact recovery as long as s is not too large for all levels of outlier magnitudes, without any knowledge of the outliers, which validates our theoretical analysis. Unsurprisingly, TWF fails quickly



(a) $\|\boldsymbol{\eta}\|_\infty = 0.1\|\boldsymbol{x}\|^2$ (b) $\|\boldsymbol{\eta}\|_\infty = \|\boldsymbol{x}\|^2$



(c) $\|\boldsymbol{\eta}\|_\infty = 10\|\boldsymbol{x}\|^2$ (d) $\|\boldsymbol{\eta}\|_\infty = 100\|\boldsymbol{x}\|^2$

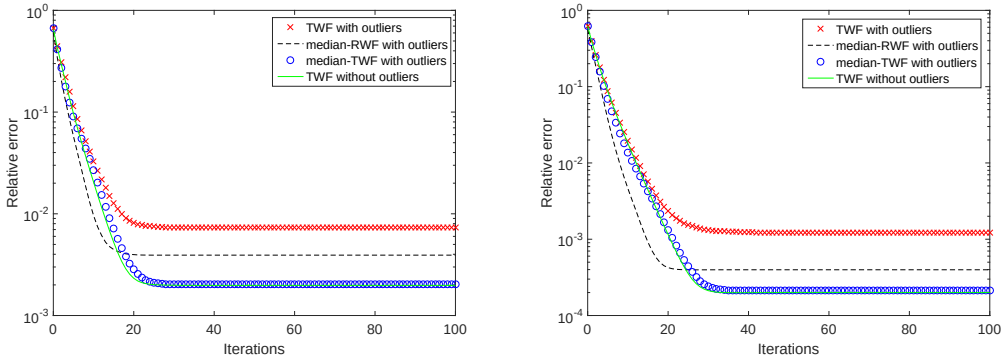
Fig. 4.2: Success rate of exact recovery with outliers for median-RWF, median-TWF, trimean-TWF, and TWF at different levels of outlier magnitudes.

even with very small fraction of outliers. No successful instance is observed for TWF when $s \geq 0.02$ irrespective of the value of $\|\boldsymbol{\eta}\|_\infty$. Trimean-TWF requires knowing the number of outliers and does not exhibit as sharp phase transition as median-TWF, and in general underperforms our median-TWF, except when both $\|\boldsymbol{\eta}\|_\infty$ and s gets very large, see Figure 4.2(d). This is because in this range with large outliers, knowing the fraction s of outliers provides substantial advantage for trimean-TWF to eliminate them, while the sample median can be deviated significantly from the true median for large s. Moreover, it is worth mentioning that exact recovery is more challenging for median-TWF and median-RWF when the magnitudes of most outliers are comparable to the measurements, as in Figure 4.2(c). In such a case, the outliers are more difficult to exclude as opposed to the case with very large outlier magnitudes as in Figure 4.2(d); and meanwhile, the outlier

magnitudes in Figure 4.2(c) are large enough to affect the accuracy heavily in contrast to the cases in Figure 4.2(a) and 4.2(b) where outliers are less prominent. In general, median-RWF can tolerate larger fraction of outliers than median-TWF. This could be due to that the lower-order objective reduces the variance and brings more stable search direction.

We now examine the performance of median-TWF and median-RWF in the presence of both sparse outliers and dense bounded noise. The entries of the dense bounded noise $\boldsymbol{w}$ is generated independently from $\mathcal{U}(0, \|\boldsymbol{w}\|_\infty)$, with $\|\boldsymbol{w}\|_\infty/\|\boldsymbol{x}\|^2 = 0.001, 0.01$ respectively. The outliers are then generated as $\eta_i \sim \|\boldsymbol{w}\| \cdot \text{Bernoulli}(0.1)$ independently. Figure 4.3(a) and Figure 4.3(b) depict the relative error $\text{dist}(\boldsymbol{z}^{(t)}, \boldsymbol{x})/\|\boldsymbol{x}\|$ with respect to the iteration count t, for uniform noise at different levels. It can be seen that median-TWF under outlier corruption clearly outperforms TWF under the same situation, and acts as if the outliers do not exist by achieving almost the same accuracy as TWF under no outliers. Moreover, the solution accuracy of median-TWF has 10 times gain from Figure 4.3(a) to Figure 4.3(b) as $\|\boldsymbol{w}\|_\infty$ shrinks by $1/10$, which corroborates Theorem 4.3 nicely. Furthermore, it can be seen that median-RWF converges faster than other algorithms, which is due to good curvature of low-oder objective and corroborates the result in [68]. It can also be seen that the solution returned by median-RWF is not as accurate as median-TWF. This is because that median-TWF employs more delicate truncation rule ($\mathbf{1}_{\mathcal{E}_1^i}$) which may reduce the contribution of dense noises.

Finally, we consider when the measurements are corrupted both by Poisson noise and outliers, which models photon detection in optical imaging applications. We generate each measurement as $y_i \sim \text{Poisson}(|\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2)$, for $i = 1, \cdots, m$, which is then corrupted with probability $s = 0.1$ by outliers. The entries of the outlier are obtained by first generating $\eta_i \sim \|\boldsymbol{x}\|^2 \cdot \mathcal{U}(0, 1)$ independently, and then rounding it to the nearest integer. Figure 4.4 depicts the relative error $\text{dist}(\boldsymbol{z}^{(t)}, \boldsymbol{x})/\|\boldsymbol{x}\|$ with respect to the iteration count t, where median-TWF and median-RWF under both Poisson noise and sparse outlier noise has almost the same accuracy as, if not better than, TWF under only Poisson noise.



(a) $\|\boldsymbol{w}\|_\infty = 0.01\|\boldsymbol{x}\|^2$ (b) $\|\boldsymbol{w}\|_\infty = 0.001\|\boldsymbol{x}\|^2$

Fig. 4.3: The relative error with respect to the iteration count for median-TWF, median-RWF and TWF with both dense noise and sparse outliers, and TWF with only dense noise.(a) and (b): Uniform noise with different levels.

4.5 Conclusions

In this chapter, we study provably effective approaches, median-TWF and median-RWF, for phase retrieval when the measurements are corrupted by sparse outliers that can take arbitrary values. Our strategy is to apply gradient descent with respect to carefully chosen loss functions, where both the initialization and the search directions are pruned guided by the sample median. We show that both algorithms allow exact recovery even with a constant proportion of arbitrary outliers for robust phase retrieval using a near-optimal number of measurements up to a logarithmic factor. Our algorithm performs well for phase retrieval problem under sparse corruptions. We anticipate that the technique developed here will be useful for designing provably robust algorithms for other inference problems under sparse corruptions.

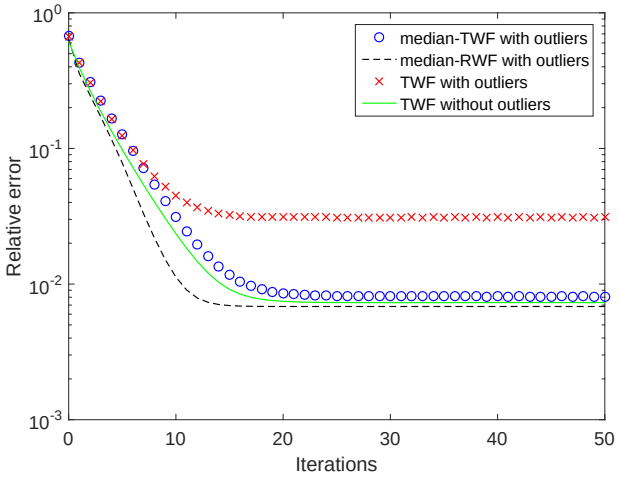


Fig. 4.4: The relative error with respect to the iteration count for median-TWF, median-RWF and TWF with both Poisson noise and sparse outliers, and TWF with only Poisson noise.

4.6 Technical Proofs

4.6.1 Proof Roadmap

Broadly speaking, the proofs for median-TWF and median-RWF follow the same roadmap. The crux is to use the statistical properties of the median to show that the median-truncated gradients satisfy the so-called *Regularity Condition* [1], which guarantees the linear convergence of the update rule, provided the initialization provably lands in a small neighborhood of the true signal.

We first develop a few statistical properties of median. We then analyzes the initialization that is used in both algorithms. We then state the definition of Regularity Condition and explain how it leads to the linear convergence rate. We provide separate detailed proofs for two algorithms in Section 4.6.2 and Section 4.6.3, respectively, because they involve different bounding techniques that may be of independent interest due to different loss functions.

We define the quantile of a population distribution and its sample version.

Definition 4.5 (Generalized quantile function)**.** *Let* $0 < p < 1$. *For a cumulative distribu-*

tion function (CDF) F, the generalized quantile function is defined as

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}. \tag{4.18}$$

*For simplicity, denote $\theta_p(F) = F^{-1}(p)$ as the p-*quantile *of F. Moreover for a sample sequence $\{X_i\}_{i=1}^m$, the sample p-quantile $\theta_p(\{X_i\})$ means $\theta_p(\hat{F})$, where $\hat{F}$ is the empirical distribution of the samples $\{X_i\}_{i=1}^m$.*

Remark 1. *We note that the median med$(\{X_i\}) = \theta_{1/2}(\{X_i\})$, and we use both notations interchangeably.*

Next, we show that as long as the sample size is large enough, the sample quantile concentrates around the population quantile (motivated from [145]), as in Lemma 4.6.

Lemma 4.6. *Suppose $F(\cdot)$ is cumulative distribution function (i.e., non-decreasing and right-continuous) with continuous density function $F'(\cdot)$. Assume the samples $\{X_i\}_{i=1}^m$ are i.i.d. drawn from F. Let $0 < p < 1$. If $l < F'(\theta) < L$ for all θ in $\{\theta : |\theta - \theta_p| \leq \epsilon\}$, then*

$$|\theta_p(\{X_i\}_{i=1}^m) - \theta_p(F)| < \epsilon \tag{4.19}$$

holds with probability at least $1 - 2\exp(-2m\epsilon^2 l^2)$.

Proof. See Section 4.6.4. □

Lemma 4.7 bounds the distance between the median of two sequences.

Lemma 4.7. *Given a vector $\boldsymbol{X} = (X_1, X_2, ..., X_n)$, reorder the entries in a non-decreasing manner*

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n-1)} \leq X_{(n)}.$$

Given another vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$*, then*

$$|X_{(k)} - Y_{(k)}| \leq \|\boldsymbol{X} - \boldsymbol{Y}\|_\infty, \tag{4.20}$$

holds for all $k = 1, ..., n$.

Proof. See Section 4.6.4. □

Lemma 4.8, as a key robustness property of median, suggests that in the presence of outliers, one can bound the sample median from both sides by neighboring quantiles of the corresponding clean samples.

Lemma 4.8. *Consider* clean samples $\{\tilde{X}_i\}_{i=1}^m$*. If a fraction* s *(*$s < \frac{1}{2}$*) of them are corrupted by outliers, one obtains* contaminated samples $\{X_i\}_{i=1}^m$ *which contain* sm *corrupted samples and* $(1-s)m$ *clean samples. Then for a quantile* p *such that* $s < p < 1 - s$*, we have*

$$\theta_{p-s}(\{\tilde{X}_i\}) \leq \theta_p(\{X_i\}) \leq \theta_{p+s}(\{\tilde{X}_i\}).$$

Proof. See Section 4.6.4. □

Finally, Lemma 4.9 is related to bound the value of the median, as well as the density at the median for the product of two possibly correlated standard Gaussian random variables.

Lemma 4.9. *Let* $u, v \sim \mathcal{N}(0, 1)$ *which can be correlated with the correlation coefficient* $|\rho| \leq 1$*. Let* $r = |uv|$*, and* $\psi_\rho(x)$ *represent the density of* r*. Denote* $\theta_{\frac{1}{2}}(\psi_\rho)$ *as the median of* r*, and the value of* $\psi_\rho(x)$ *at the median as* $\psi_\rho(\theta_{1/2})$*. Then for all* ρ*,*

$$0.348 < \theta_{1/2}(\psi_\rho) < 0.455,$$
$$0.47 < \psi_\rho(\theta_{1/2}) < 0.76.$$

Proof. See Section 4.6.4. □

We next consider the model that the measurements are corrupted by both bounded noise and sparse outliers given by (4.2), and show that the initialization provided by the median-truncated spectral method in (4.5) is close enough to the ground truth, i.e., $\mathrm{dist}(\boldsymbol{z}^{(0)}, \boldsymbol{x}) \leq \delta\|\boldsymbol{x}\|$.

Proposition 4.10. *Fix $\delta > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$, and consider the model given by* (4.2)*. Suppose that $\|\mathbf{w}\|_\infty \leq c\|\boldsymbol{x}\|^2$ for some sufficiently small constant $c > 0$ and that $\|\eta\|_0 \leq sm$ for some sufficiently small constant s. With probability at least $1 - \exp(-\Omega(m))$, the initialization given by the median-truncated spectral method obeys*[2]

$$dist(\boldsymbol{z}^{(0)}, \boldsymbol{x}) \leq \delta\|\boldsymbol{x}\|, \tag{4.21}$$

provided that $m > c_0 n$ for some constant $c_0 > 0$.

Proof. See Section 4.6.4.

Once the initialization is guaranteed to be within a small neighborhood of the ground truth, we only need to show that the truncated gradient (4.7) and (4.10) satisfy the *Regularity Condition* (RC) [1, 61], which guarantees the geometric convergence of median-TWF/median-RWF once the initialization lands into this neighborhood.

Definition 4.11. *The gradient $\nabla\ell(\boldsymbol{z})$ satisfies the Regularity Condition $\mathsf{RC}(\mu, \lambda, c)$ if*

$$\langle \nabla\ell(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}\rangle \geq \frac{\mu}{2}\|\nabla\ell(\boldsymbol{z})\|^2 + \frac{\lambda}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 \tag{4.22}$$

for all $\boldsymbol{z}$ obeying $\|\boldsymbol{z} - \boldsymbol{x}\| \leq c\|\boldsymbol{x}\|$.

The above RC guarantees that the gradient descent update $\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \mu\nabla\ell(\boldsymbol{z})$ converges to the true signal $\boldsymbol{x}$ geometrically [61] if $\mu\lambda < 1$. We repeat this argument below

[2]Notation $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \geq c|g(n)|$.

for completeness.

$$\begin{aligned}\text{dist}^2(\boldsymbol{z}-\mu\nabla\ell(\boldsymbol{z}),\boldsymbol{x}) &\le \|\boldsymbol{z}-\mu\nabla\ell(\boldsymbol{z})-\boldsymbol{x}\|^2 \\ &= \|\boldsymbol{z}-\boldsymbol{x}\|^2+\|\mu\nabla\ell(\boldsymbol{z})\|^2-2\mu\left\langle \boldsymbol{z}-\boldsymbol{x},\nabla\ell(\boldsymbol{z})\right\rangle \\ &\le \|\boldsymbol{z}-\boldsymbol{x}\|^2+\|\mu\nabla\ell(\boldsymbol{z})\|^2-\mu^2\|\nabla\ell(\boldsymbol{z})\|^2-\mu\lambda\|\boldsymbol{z}-\boldsymbol{x}\|^2 \\ &= (1-\mu\lambda)\text{dist}^2(\boldsymbol{z},\boldsymbol{x}).\end{aligned}$$

4.6.2 Proofs for Median-TWF

We first show that $\nabla\ell_{tr}(\boldsymbol{z})$ in (4.7) satisfies the RC for the noise-free case, and then extend it to the model with only sparse outliers, thus together with Proposition 4.10 establishing the global convergence of median-TWF in both cases. At last we prove Theorem 4.3 in the presence of both sparse outliers and dense bounded noise.

Proof of Proposition 4.1

We consider the noise-free model. The central step to establish the RC is to show that the sample median used in the truncation rule of median-TWF concentrates at the level $\|\boldsymbol{z}-\boldsymbol{x}\|\|\boldsymbol{z}\|$ as stated in the following proposition.

Proposition 4.12. *If $m > c_0 n\log n$, then with probability at least $1-c_1\exp(-c_2 m)$,*

$$0.6\|\boldsymbol{z}\|\|\boldsymbol{z}-\boldsymbol{x}\| \le \theta_{0.49},\theta_{0.5},\theta_{0.51}(\left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}|^2-|\boldsymbol{a}_i^T\boldsymbol{z}|^2\right|\right\}_{i=1}^m) \le \|\boldsymbol{z}\|\|\boldsymbol{z}-\boldsymbol{x}\|, \tag{4.23}$$

holds for all $\boldsymbol{z},\boldsymbol{x}$ satisfying $\|\boldsymbol{z}-\boldsymbol{x}\|<1/11\|\boldsymbol{z}\|$.

Proof. Detailed proof is provided in Section 4.6.4. □

We note that a similar property for the sample mean has been shown in [61] as long as the number m of measurements is on the order of n. In fact, the sample median is much more challenging to bound due to its non-linearity, which also causes slightly more

measurements compared to the sample mean.

Then we can establish that $\langle \nabla \ell_{tr}(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x} \rangle$ is lower bounded on the order of $\|\boldsymbol{z} - \boldsymbol{x}\|^2$, as in Proposition 4.13, and that $\|\nabla \ell_{tr}(\boldsymbol{z})\|$ is upper bounded on the order of $\|\boldsymbol{z} - \boldsymbol{x}\|$, as in Proposition 4.14.

Proposition 4.13 (Adapted version of Proposition 2 of [61])**.** *Consider the noise-free case* $y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|^2$ *for* $i = 1, \cdots, m$*, and any fixed constant* $\epsilon > 0$*. Under the condition* (4.8)*, if* $m > c_0 n \log n$*, then with probability at least* $1 - c_1 \exp(-c_2 \epsilon^{-2} m)$*,*

$$\langle \nabla \ell_{tr}(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x} \rangle \geq \left\{ 1.99 - 2(\zeta_1 + \zeta_2) - \sqrt{8/\pi} \alpha_h^{-1} - \epsilon \right\} \|\boldsymbol{z} - \boldsymbol{x}\|^2 \tag{4.24}$$

holds uniformly over all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$ *satisfying*

$$\frac{\|\boldsymbol{z} - \boldsymbol{x}\|}{\|\boldsymbol{z}\|} \leq \min \left\{ \frac{1}{11}, \frac{\alpha_l}{\alpha_h}, \frac{\alpha_l}{6}, \frac{\sqrt{98/3}(\alpha_l)^2}{2\alpha_u + \alpha_l} \right\}, \tag{4.25}$$

where $c_0, c_1, c_2 > 0$ *are some universal constants, and* $\zeta_1, \zeta_2, \alpha_l, \alpha_u$ *and* α_h *are defined in* (4.8)*.*

The proof of Proposition 4.13 adapts the proof of Proposition 2 of [61], by properly setting parameters based on the properties of sample median. For completeness, we include a short outline of the proof in Section 4.6.

Proposition 4.14 (Lemma 7 of [61])**.** *Under the same condition as in Proposition 4.13, if* $m > c_0 n$*, then there exist some constants* $c_1, c_2 > 0$ *such that with probability at least* $1 - c_1 \exp(-c_2 m)$*,*

$$\|\nabla \ell_{tr}(\boldsymbol{z})\| \leq (1 + \delta) \cdot 2\sqrt{1.02 + 2/\alpha_h} \|\boldsymbol{z} - \boldsymbol{x}\| \tag{4.26}$$

holds uniformly over all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$ *satisfying*

$$\frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{z}\|} \leq \min\left\{\frac{1}{11}, \frac{\alpha_l}{\alpha_h}, \frac{\alpha_l}{6}, \frac{\sqrt{98/3}(\alpha_l)^2}{2\alpha_u+\alpha_l}\right\}, \tag{4.27}$$

where δ *can be arbitrarily small as long as* m/n *sufficiently large, and* α_l, α_u *and* α_h *are given in* (4.8).

Proof. See the proof of Lemma 7 in [61]. □

With these two propositions and (4.8), RC is guaranteed by setting

$$\mu < \mu_0 := \frac{(1.99 - 2(\zeta_1+\zeta_2) - \sqrt{8/\pi}\alpha_h^{-1}}{2(1+\delta)^2 \cdot (1.02 + 2/\alpha_h)},$$
$$\lambda + \mu \cdot 4(1+\delta)^2 \cdot (1.02 + 2/\alpha_h) < 2\left\{1.99 - 2(\zeta_1+\zeta_2) - \sqrt{8/\pi}\alpha_h^{-1} - \epsilon\right\}.$$

Proof of Theorem 4.2

We next consider the model (4.2) with only sparse outliers. It suffices to show that $\nabla \ell_{tr}(\boldsymbol{z})$ continues to satisfy the RC. The critical step is to bound the sample median of the corrupted measurements. Lemma 4.8 yields

$$\theta_{\frac{1}{2}-s}(\{|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2|\}) \leq \theta_{\frac{1}{2}}(\{|y_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2|\}) \leq \theta_{\frac{1}{2}+s}(\{|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2|\}. \tag{4.28}$$

For simplicity of notation, we let $\boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$. Then for the instance of $s = 0.01$, by Proposition 4.12, we have with probability at least $1 - 2\exp(-\Omega(m))$,

$$0.6\|\boldsymbol{z}\|\|\boldsymbol{h}\| \leq \theta_{\frac{1}{2}}(\{|y_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2|\}) \leq \|\boldsymbol{z}\|\|\boldsymbol{h}\|. \tag{4.29}$$

To differentiate from $\mathcal{E}_2^i$, we define $\tilde{\mathcal{E}}_2^i := \left\{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \leq \alpha_h \text{med}\left\{\left|y_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} \frac{|\boldsymbol{a}_i^T\boldsymbol{z}|}{\|\boldsymbol{z}\|}\right\}$.

We then have

$$\nabla \ell_{tr}(\boldsymbol{z}) = \frac{1}{m}\sum_{i=1}^{m} \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - y_i}{\boldsymbol{a}_i^T\boldsymbol{z}}\boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i}$$
$$= \underbrace{\frac{1}{m}\sum_{i=1}^{m} \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}}\boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \tilde{\mathcal{E}}_2^i}}_{\nabla^{clean}\ell_{tr}(\boldsymbol{z})} + \underbrace{\frac{1}{m}\sum_{i\in S}\left(\frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - y_i}{\boldsymbol{a}_i^T\boldsymbol{z}}\mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i} - \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}}\mathbf{1}_{\mathcal{E}_1^i \cap \tilde{\mathcal{E}}_2^i}\right)\boldsymbol{a}_i}_{\nabla^{extra}\ell_{tr}(\boldsymbol{z})}.$$

Choosing ϵ small enough, it is easy to verify that Propositions 4.13 and 4.14 are still valid on $\nabla^{clean}\ell_{tr}(\boldsymbol{z})$. Thus, one has

$$\langle \nabla^{clean}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle \geq \left\{1.99 - 2(\zeta_1+\zeta_2) - \sqrt{8/\pi}\alpha_h^{-1} - \epsilon\right\}\|\boldsymbol{h}\|^2,$$
$$\left\|\nabla^{clean}\ell_{tr}(\boldsymbol{z})\right\| \leq (1+\delta)\cdot 2\sqrt{1.02 + 2/\alpha_h}\|\boldsymbol{h}\|.$$

We next bound the contribution of $\nabla^{extra}\ell_{tr}(\boldsymbol{z})$. Introduce $\boldsymbol{q} = [q_1, \ldots, q_m]^T$, where

$$q_i := \left(\frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - y_i}{\boldsymbol{a}_i^T\boldsymbol{z}}\mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i} - \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}}\mathbf{1}_{\mathcal{E}_1^i \cap \tilde{\mathcal{E}}_2^i}\right)\mathbf{1}_{\{i\in S\}}.$$

It can be seen that $|q_i| \leq 2\alpha_h\|\boldsymbol{h}\|$. Thus $\|\boldsymbol{q}\| \leq \sqrt{sm}\cdot 2\alpha_h\|\boldsymbol{h}\|$, and

$$\left\|\nabla^{extra}\ell_{tr}(\boldsymbol{z})\right\| = \frac{1}{m}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\| \leq 2(1+\delta)\sqrt{s}\alpha_h\|\boldsymbol{h}\|,$$
$$\left|\left\langle \nabla^{extra}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \leq \|\boldsymbol{h}\|\cdot\left\|\frac{1}{m}\nabla^{extra}\ell_{tr}(\boldsymbol{z})\right\| \leq 2(1+\delta)\sqrt{s}\alpha_h\|\boldsymbol{h}\|^2,$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]^T$. Then, we have

$$-\langle \nabla \ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle \geq \left\langle \nabla^{clean}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle - \left|\left\langle \nabla^{extra}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right|$$
$$\geq \left(1.99 - 2(\zeta_1+\zeta_2) - \sqrt{8/\pi}\alpha_h^{-1} - \epsilon - 2(1+\delta)\sqrt{s}\alpha_h\right)\|\boldsymbol{h}\|^2,$$

and

$$\begin{aligned}\|\nabla \ell_{tr}(\boldsymbol{z})\| &\le \left\|\nabla^{clean}\ell_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{extra}\ell_{tr}(\boldsymbol{z})\right\| \\ &\le 2(1+\delta)\left(\sqrt{1.02+2/\alpha_h}+\sqrt{s}\alpha_h\right)\|\boldsymbol{h}\|. \end{aligned} \tag{4.30}$$

Therefore, the RC is guaranteed if μ, λ, ϵ are chosen properly and s is sufficiently small.

Proof of Theorem 4.3

We consider the model (4.2), and split our analysis of the gradient loop into two regimes.

• **Regime 1**: $c_4\|\boldsymbol{z}\| \ge \|\boldsymbol{h}\| \ge c_3\frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{z}\|}$. In this regime, error contraction by each gradient step is given by

$$\text{dist}\left(\boldsymbol{z}-\mu\nabla\ell_{tr}(\boldsymbol{z}),\boldsymbol{x}\right) \le (1-\rho)\text{dist}(\boldsymbol{z},\boldsymbol{x}).$$

It suffices to justify that $\nabla\ell_{tr}(\boldsymbol{z})$ satisfies the RC. Denote $\tilde{y}_i := (\boldsymbol{a}_i^T\boldsymbol{x})^2 + w_i$. Then by Lemma 4.8, we have

$$\theta_{\frac{1}{2}-s}\left\{\left|\tilde{y}_i-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} \le \text{med}\left\{\left|y_i-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} \le \theta_{\frac{1}{2}+s}\left\{\left|\tilde{y}_i-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\}.$$

Moreover, by Lemma 4.7 we have

$$\begin{aligned}\left|\theta_{\frac{1}{2}+s}\left\{\left|\tilde{y}_i-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\}-\theta_{\frac{1}{2}+s}\left\{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\}\right| &\le \|\boldsymbol{w}\|_\infty, \\ \left|\theta_{\frac{1}{2}-s}\left\{\left|\tilde{y}_i-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\}-\theta_{\frac{1}{2}-s}\left\{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\}\right| &\le \|\boldsymbol{w}\|_\infty.\end{aligned}$$

Assume that $s = 0.01$. By Proposition 4.12, if c_3 is sufficiently large (i.e., $c_3 > 100$), we still shave

$$0.6\|\boldsymbol{x}-\boldsymbol{z}\|\|\boldsymbol{z}\| \le \text{med}\left\{\left|y_i-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} \le \|\boldsymbol{x}-\boldsymbol{z}\|\|\boldsymbol{z}\|. \tag{4.31}$$

Furthermore, recall $\tilde{\mathcal{E}}_2^i := \left\{ \left|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \le \alpha_h \text{med}\left\{\left|(\boldsymbol{a}_i^T\boldsymbol{z})^2 - y_i\right|\right\} \frac{|\boldsymbol{a}_i^T\boldsymbol{z}|}{\|\boldsymbol{z}\|} \right\}$. Then,

$$\begin{aligned}
\nabla \ell_{tr}(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^m \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - y_i}{\boldsymbol{a}_i^T\boldsymbol{z}} \boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i} \\
&= \underbrace{\frac{1}{m}\left(\sum_{i\notin S} \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}} \boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i} + \sum_{i\in S} \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}} \boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \tilde{\mathcal{E}}_2^i} \right)}_{\nabla^{clean}\ell_{tr}(\boldsymbol{z})} \\
&\quad \underbrace{- \frac{1}{m}\sum_{i\notin S} \frac{\boldsymbol{w}_i}{\boldsymbol{a}_i^T\boldsymbol{z}} \boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i}}_{\nabla^{noise}\ell_{tr}(\boldsymbol{z})} + \underbrace{\frac{1}{m}\sum_{i\in S} \left(\frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - y_i}{\boldsymbol{a}_i^T\boldsymbol{z}} \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i} - \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}} \mathbf{1}_{\mathcal{E}_1^i \cap \tilde{\mathcal{E}}_2^i} \right) \boldsymbol{a}_i}_{\nabla^{extra}\ell_{tr}(\boldsymbol{z})} .
\end{aligned}$$

For $i \notin S$, the inclusion property (i.e. $\mathcal{E}_3^i \subseteq \mathcal{E}_2^i \subseteq \mathcal{E}_4^i$) holds because

$$\left|y_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \in \left|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \pm |w_i|$$

and $|w_i| \le \frac{1}{c_3}\|\boldsymbol{h}\|\|\boldsymbol{z}\|$ for some sufficient large c_3. For $i \in S$, the inclusion $\mathcal{E}_3^i \subseteq \tilde{\mathcal{E}}_2^i \subseteq \mathcal{E}_4^i$ holds because of (4.31). All the proof arguments for Propositions 4.13 and 4.14 are also valid for $\nabla^{clean}\ell_{tr}(\boldsymbol{z})$, and thus we have

$$\begin{aligned}
&\langle \nabla^{clean}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle \ge \left\{1.99 - 2(\zeta_1+\zeta_2) - \sqrt{8/\pi}\alpha_h^{-1} - \epsilon\right\}\|\boldsymbol{h}\|^2, \\
&\left\|\nabla^{clean}\ell_{tr}(\boldsymbol{z})\right\| \le (1+\delta)\cdot 2\sqrt{1.02 + 2/\alpha_h}\|\boldsymbol{h}\|.
\end{aligned}$$

Next, we turn to control the contribution of the noise. Let $\tilde{w}_i = \frac{w_i}{\boldsymbol{a}_i^T\boldsymbol{z}}\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i}$, and then we have

$$\|\nabla^{noise}\ell_{tr}(\boldsymbol{z})\| = \left\|\frac{1}{m}\boldsymbol{A}^T\tilde{\boldsymbol{w}}\right\| \le \left\|\frac{1}{\sqrt{m}}\boldsymbol{A}^T\right\| \left\|\frac{\tilde{\boldsymbol{w}}}{\sqrt{m}}\right\| \le (1+\delta)\|\tilde{\boldsymbol{w}}\|_\infty \le (1+\delta)\frac{\|\boldsymbol{w}\|_\infty}{\alpha_l\|\boldsymbol{z}\|},$$

when m/n is sufficiently large. Given the regime condition $\|\boldsymbol{h}\| \ge c_3\frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{z}\|}$, we further

have

$$\|\nabla^{noise}\ell_{tr}(\boldsymbol{z})\| \le \frac{(1+\delta)}{c_3\alpha_l}\|\boldsymbol{h}\|,$$
$$\left|\left\langle \nabla^{noise}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \le \left\|\nabla^{noise}\ell_{tr}(\boldsymbol{z})\right\| \cdot \|\boldsymbol{h}\| \le \frac{(1+\delta)}{c_3\alpha_l}\|\boldsymbol{h}\|^2.$$

We next bound the contribution of $\nabla^{extra}\ell_{tr}(\boldsymbol{z})$. Introduce $\boldsymbol{q} = [q_1, \ldots, q_m]^T$, where

$$q_i := \left(\frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - y_i}{\boldsymbol{a}_i^T\boldsymbol{z}}\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i} - \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}}\mathbf{1}_{\mathcal{E}_1^i\cap\tilde{\mathcal{E}}_2^i}\right)\mathbf{1}_{\{i\in S\}}.$$

Then $|q_i| \le 2\alpha_h\|\boldsymbol{h}\|$, and $\|\boldsymbol{q}\| \le \sqrt{sm}\cdot 2\alpha_h\|\boldsymbol{h}\|$. We thus have

$$\left\|\nabla^{extra}\ell_{tr}(\boldsymbol{z})\right\| = \frac{1}{m}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\| \le 2(1+\delta)\sqrt{s}\alpha_h\|\boldsymbol{h}\|,$$
$$\left|\left\langle \nabla^{extra}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \le \|\boldsymbol{h}\| \cdot \left\|\nabla^{extra}\ell_{tr}(\boldsymbol{z})\right\| \le 2(1+\delta)\sqrt{s}\alpha_h\|\boldsymbol{h}\|^2.$$

Putting these together, one has

$$\begin{aligned}\langle\nabla\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle &\ge \left\langle \nabla^{clean}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle - \left|\left\langle \nabla^{noise}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| - \left|\left\langle \nabla^{extra}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \\ &\ge \left(1.99 - 2(\zeta_1+\zeta_2) - \sqrt{8/\pi}\alpha_h^{-1} - \epsilon - (1+\delta)(1/(c_3\alpha_z^l) + 2\sqrt{s}\alpha_h)\right)\|\boldsymbol{h}\|^2, \quad (4.32)\end{aligned}$$

and

$$\begin{aligned}\|\nabla\ell_{tr}(\boldsymbol{z})\| &\le \left\|\nabla^{clean}\ell_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{noise}\ell_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{extra}\ell_{tr}(\boldsymbol{z})\right\| \\ &\le (1+\delta)\left(2\sqrt{1.02 + 2/\alpha_h} + 1/(c_3\alpha_z^l) + 2\sqrt{s}\alpha_h\right)\|\boldsymbol{h}\|. \quad (4.33)\end{aligned}$$

The RC is guaranteed if μ, λ, ϵ are chosen properly, c_3 is sufficiently large and s is sufficiently small.

• **Regime 2**: Once the iterate enters this regime with $\|\boldsymbol{h}\| \le \frac{c_3\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{z}\|}$, each gradient iterate may not reduce the estimation error. However, in this regime each move size $\mu\nabla\ell_{tr}(\boldsymbol{z})$

is at most $\mathcal{O}(\|\boldsymbol{w}\|_\infty/\|\boldsymbol{z}\|)$. Then the estimation error cannot increase by more than $\frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{z}\|}$ with a constant factor. Thus one has

$$\text{dist}\left(\boldsymbol{z} - \mu\nabla\ell_{tr}(\boldsymbol{z}), \boldsymbol{x}\right) \leq c_5\frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|}$$

for some constant c_5. As long as $\|\boldsymbol{w}\|_\infty/\|\boldsymbol{x}\|^2$ is sufficiently small, it is guaranteed that $c_5\frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|} \leq c_4\|\boldsymbol{x}\|$. If the iterate jumps out of *Regime 2*, it falls into *Regime 1*.

4.6.3 Proofs for Median-RWF

We first show that $\nabla\mathcal{R}_{tr}(\boldsymbol{z})$ in (4.10) satisfies the RC for the noise-free case, and then extend it to the model with only sparse outliers, thus together with Proposition 4.10 establishing the global convergence of median-RWF in both cases. At last, we prove Theorem 4.3 in the presence of both sparse outliers and dense bounded noise.

Proof of Proposition 4.1

The central step to establish the RC is to show that the sample median used in the truncation rule of median-RWF concentrates on the order of $\|\boldsymbol{z} - \boldsymbol{x}\|$ as stated in the following proposition.

Proposition 4.15. *If $m > c_0 n\log n$, then with probability at least $1 - c_1\exp(-c_2 m)$,*

$$0.5\|\boldsymbol{z} - \boldsymbol{x}\| \leq \theta_{0.49}, \theta_{1/2}, \theta_{0.51}\left(\left\{\left||\boldsymbol{a}_i^T\boldsymbol{z}| - |\boldsymbol{a}_i^T\boldsymbol{x}|\right|\right\}_{i=1}^m\right) \leq 0.8\|\boldsymbol{z} - \boldsymbol{x}\|, \tag{4.34}$$

holds for all $\boldsymbol{z}, \boldsymbol{x}$ satisfying $\|\boldsymbol{z} - \boldsymbol{x}\| < 1/11\|\boldsymbol{z}\|$.

Proof. See Section 4.6.4. □

Next we give a bound on the left hand side of RC.

Proposition 4.16 (Adapted version of Proposition 2 of [61])**.** *Consider the noise-free measurements $y_i = |\boldsymbol{a}_i^T\boldsymbol{x}|$ and any fixed constant $\epsilon > 0$. If $m > c_0 n\log n$, then with probability*

at least $1 - c_1 \exp(-c_2 m)$*,*

$$\langle \nabla \mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x} \rangle \geq \{0.88 - \zeta_1' - \zeta_2' - \epsilon\} \|\boldsymbol{z} - \boldsymbol{x}\|^2 \tag{4.35}$$

holds uniformly over all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$ *satisfying* $\frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \frac{1}{20}$*, where* $c_0, c_1, c_2 > 0$ *are some universal constants, and* ζ_1', ζ_2' *are given by*

$$\zeta_1' := 1 - \min\left\{ \mathrm{E}\left[\xi^2 \mathbf{1}_{\{\xi \geq 0.5\sqrt{1.01}\alpha_h' \frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\}} \right], \mathrm{E}\left[\mathbf{1}_{\{\xi \geq 0.5\sqrt{1.01}\alpha_h' \frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\}} \right] \right\}$$
$$\zeta_2' := \mathrm{E}\left[\xi^2 \mathbf{1}_{\{|\xi| > 0.5\sqrt{0.99}\alpha_h'\}} \right]$$

for some $\xi \sim \mathcal{N}(0,1)$ *and* $\alpha_h' = 5$*.*

Proof. See Section 4.6.4. □

Proposition 4.16 indicates that $\langle \nabla \mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x} \rangle$ is lower bounded by $\|\boldsymbol{z} - \boldsymbol{x}\|^2$ with some positive constant coefficient. In order to prove the RC, it suffices to show that $\|\nabla \mathcal{R}_{tr}(\boldsymbol{z})\|$ is upper bounded by the order of $\|\boldsymbol{z} - \boldsymbol{x}\|$ when $\boldsymbol{z}$ is within the neighborhood of true signal $\boldsymbol{x}$.

Proposition 4.17 (Lemma 7 of [61])**.** *If* $m > c_0 n$*, then there exist some constants* $c_1, c_2 > 0$ *such that with probability at least* $1 - c_1 \exp(-c_2 m)$*,*

$$\|\nabla \mathcal{R}_{tr}(\boldsymbol{z})\| \leq (1.8 + \delta)\|\boldsymbol{z} - \boldsymbol{x}\| \tag{4.36}$$

holds uniformly over all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$ *satisfying* $\|\boldsymbol{x} - \boldsymbol{z}\| \leq \frac{1}{11}\|\boldsymbol{x}\|$ *where* δ *can be arbitrarily small as long as* c_0 *sufficiently large.*

Proof. See Section 4.6.4. □

With the above two propositions, RC is guaranteed by setting $\mu < \mu_0 := \frac{2(0.88-\zeta_1'-\zeta_2'-\epsilon)}{(1.8+\delta)^2}$ and $\lambda + \mu \cdot (1.8+\delta)^2 < 2(0.88 - \zeta_1' - \zeta_2' - \epsilon)$.

Proof of Theorem 4.2

We consider the model (4.2) with only outliers, i.e., $y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2 + \eta_i$ for $i = 1, \cdots, m$. It suffices to show that $\nabla \mathcal{R}_{tr}(z)$ satisfies the RC. The critical step is to lower and upper bound the sample median of the corrupted measurements. Lemma 4.8 yields

$$\theta_{\frac{1}{2}-s}(\{||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}||\}) \le \theta_{\frac{1}{2}}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}||\}) \le \theta_{\frac{1}{2}+s}(\{||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}||\}. \quad (4.37)$$

For the simplicity of notation, we let $\boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$. Then for the instance of $s = 0.01$, Proposition 4.15 yields that if $m > c_0 n \log n$, then

$$0.5\|\boldsymbol{h}\| \le \theta_{\frac{1}{2}}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}||\}) \le 0.8\|\boldsymbol{h}\| \quad (4.38)$$

holds with probability at least $1 - 2\exp(-\Omega(m))$.

To differentiate from $\mathcal{T}^i$, we define $\tilde{\mathcal{T}}^i := \left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right| \le \alpha_h' \text{med}\left\{\left|\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}\right\}$. We then have

$$\begin{aligned}\nabla \mathcal{R}_{tr}(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^{m}\left(|\boldsymbol{a}_i^T\boldsymbol{z}| - \sqrt{y_i}\right)\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i} \\ &= \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left(|\boldsymbol{a}_i^T\boldsymbol{z}| - \sqrt{y_i}\right)\boldsymbol{a}_i\mathbf{1}_{\tilde{\mathcal{T}}^i}}_{\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})} + \underbrace{\frac{1}{m}\sum_{i\in S}\left(\left(|\boldsymbol{a}_i^T\boldsymbol{z}| - \sqrt{y_i}\right)\mathbf{1}_{\mathcal{T}^i} - \left(|\boldsymbol{a}_i^T\boldsymbol{z}| - |\boldsymbol{a}_i^T\boldsymbol{x}|\right)\mathbf{1}_{\tilde{\mathcal{T}}^i}\right)\boldsymbol{a}_i}_{\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})}.\end{aligned}$$

Under the condition (4.38), the inclusion property (i.e., $\mathcal{T}_1^i \subseteq \tilde{\mathcal{T}}^i \subseteq \mathcal{T}_2^i$) holds, and all the proof arguments for Propositions 4.16 and 4.17 are also valid to $\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})$. Thus, one has

$$\begin{aligned}&\left\langle \nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle \ge (0.88 - \zeta_1' - \zeta_2' - \epsilon)\|\boldsymbol{h}\|^2 \\ &\left\|\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})\right\| \le (1.8 + \delta)\|\boldsymbol{h}\|.\end{aligned}$$

We next bound the contribution of $\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})$. Introduce $\boldsymbol{q} = [q_1, \ldots, q_m]^T$, where

$$q_i := \left(\left(|\boldsymbol{a}_i^T\boldsymbol{z}| - \sqrt{y_i}\right)\mathbf{1}_{\mathcal{T}^i} - \left(|\boldsymbol{a}_i^T\boldsymbol{z}| - |\boldsymbol{a}_i^T\boldsymbol{x}|\right)\mathbf{1}_{\tilde{\mathcal{T}}^i}\right)\mathbf{1}_{\{i\in S\}},$$

and then $|q_i| \le 1.6\alpha_h'\|\boldsymbol{h}\|$. Thus, $\|\boldsymbol{q}\| \le \sqrt{sm}\cdot 1.6\alpha_h'\|\boldsymbol{h}\|$, and

$$\begin{aligned}
\left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\| &= \frac{1}{m}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\| \le 1.6(1+\delta)\sqrt{s}\alpha_h'\|\boldsymbol{h}\|,\\
\left|\left\langle\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| &\le \|\boldsymbol{h}\|\cdot\left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\| \le 1.6(1+\delta)\sqrt{s}\alpha_h'\|\boldsymbol{h}\|^2,
\end{aligned}$$

where $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m]^T$. Then, we have

$$\begin{aligned}
\left\langle\nabla\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle &\ge \left\langle\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle - \left|\left\langle\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right|\\
&\ge \left(0.88 - \zeta_1' - \zeta_2' - \epsilon - 1.6(1+\delta)\sqrt{s}\alpha_h'\right)\|\boldsymbol{h}\|^2,
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla\mathcal{R}_{tr}(\boldsymbol{z})\| &\le \left\|\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\|\\
&\le \left(1.8 + \delta + 1.6(1+\delta)\sqrt{s}\alpha_h'\right)\|\boldsymbol{h}\|.
\end{aligned}$$

Therefore the RC is guaranteed if μ, λ are chosen properly, δ is chosen sufficiently small and s is sufficiently small.

Proof of Theorem 4.3

We consider the model (4.2) with outliers and bounded noise. We split our analysis of the gradient loop into two regimes.

- **Regime 1**: $c_4\|\boldsymbol{z}\| \ge \|\boldsymbol{h}\| \ge c_3\sqrt{\|\boldsymbol{w}\|_\infty}$. In this regime, error contraction by each

gradient step is given by

$$\text{dist}\left(\boldsymbol{z}-\mu\nabla\mathcal{R}_{tr}(\boldsymbol{z}),\boldsymbol{x}\right)\leq(1-\rho)\text{dist}(\boldsymbol{z},\boldsymbol{x}). \tag{4.39}$$

It suffices to justify that $\nabla\mathcal{R}_{tr}(\boldsymbol{z})$ satisfies the RC. Denote $\tilde{y}_i:=(\boldsymbol{a}_i^T\boldsymbol{x})^2+w_i$. Then by Lemma 4.8, we have

$$\theta_{\frac{1}{2}-s}\left\{\left|\sqrt{\tilde{y}_i}-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}\leq\text{med}\left\{\left|\sqrt{y_i}-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}\leq\theta_{\frac{1}{2}+s}\left\{\left|\sqrt{\tilde{y}_i}-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}.$$

Moreover, by Lemma 4.7 we have

$$\begin{aligned}\left|\theta_{\frac{1}{2}+s}\left\{\left|\sqrt{\tilde{y}_i}-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}-\theta_{\frac{1}{2}+s}\left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}|-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}\right|&\leq\sqrt{\|\boldsymbol{w}\|_\infty},\\ \left|\theta_{\frac{1}{2}-s}\left\{\left|\sqrt{\tilde{y}_i}-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}-\theta_{\frac{1}{2}-s}\left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}|-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}\right|&\leq\sqrt{\|\boldsymbol{w}\|_\infty}.\end{aligned}$$

Assume that $s=0.01$. By Proposition 4.15, if c_3 is sufficiently large (i.e., $c_3>100$), we still have

$$0.5\|\boldsymbol{h}\|\leq\text{med}\left\{\left|\sqrt{y_i}-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}\leq0.8\|\boldsymbol{h}\|. \tag{4.40}$$

Furthermore, recall $\tilde{\mathcal{T}}^i:=\left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}|-|\boldsymbol{a}_i^T\boldsymbol{z}|\right|\leq\alpha_h'\text{med}\left\{\left||\boldsymbol{a}_i^T\boldsymbol{z}|-\sqrt{y_i}\right|\right\}\right\}$. Then,

$$\begin{aligned}\nabla\mathcal{R}_{tr}(\boldsymbol{z})&=\frac{1}{m}\sum_{i=1}^m\left(|\boldsymbol{a}_i^T\boldsymbol{z}|-\sqrt{y_i}\right)\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i}\\&=\underbrace{\frac{1}{m}\left(\sum_{i\notin S}\left(|\boldsymbol{a}_i^T\boldsymbol{z}|-|\boldsymbol{a}_i^T\boldsymbol{x}|\right)\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i}+\sum_{i\in S}\left(|\boldsymbol{a}_i^T\boldsymbol{z}|-|\boldsymbol{a}_i^T\boldsymbol{x}|\right)\boldsymbol{a}_i\mathbf{1}_{\tilde{\mathcal{T}}^i}\right)}_{\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})}\\&\quad\underbrace{-\frac{1}{m}\sum_{i\notin S}(\sqrt{y_i}-|\boldsymbol{a}_i^T\boldsymbol{x}|)\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i}}_{\nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z})}+\underbrace{\frac{1}{m}\sum_{i\in S}\left(\left(|\boldsymbol{a}_i^T\boldsymbol{z}|-\sqrt{y_i}\right)\mathbf{1}_{\mathcal{T}^i}-\left(|\boldsymbol{a}_i^T\boldsymbol{z}|-|\boldsymbol{a}_i^T\boldsymbol{x}|\right)\mathbf{1}_{\tilde{\mathcal{T}}^i}\right)\boldsymbol{a}_i}_{\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})}.\end{aligned}$$

For $i \notin S$, the inclusion property (i.e. $\mathcal{T}_1^i \subseteq \mathcal{T}^i \subseteq \mathcal{T}_2^i$) holds because

$$\left|\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}|\right| \in \left||\boldsymbol{a}_i^T \boldsymbol{x}| - |\boldsymbol{a}_i^T \boldsymbol{z}|\right| \pm \sqrt{|w_i|}$$

and $\sqrt{|w_i|} \le \frac{1}{c_3}\|\boldsymbol{h}\|$ for some sufficient large c_3. For $i \in S$, the inclusion $\mathcal{T}_1^i \subseteq \tilde{\mathcal{T}}^i \subseteq \mathcal{T}_2^i$ holds because of (4.40). All the proof arguments for Propositions 4.16 and 4.17 are also valid for $\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})$, and thus we have

$$\begin{aligned}
&\left\langle \nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h} \right\rangle \ge (0.88 - \zeta_1' - \zeta_2' - \epsilon)\,\|\boldsymbol{h}\|^2, \\
&\left\| \nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}) \right\| \le (1.8 + \delta)\|\boldsymbol{h}\|.
\end{aligned}$$

Next, we turn to control the contribution of the noise. Let $\tilde{w}_i = (\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{x}|)\mathbf{1}_{\mathcal{T}^i}$. Then $|\tilde{w}_i| < \sqrt{|w_i|}$ and we have

$$\|\nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z})\| = \left\|\frac{1}{m}\boldsymbol{A}^T\tilde{\boldsymbol{w}}\right\| \le \left\|\frac{1}{\sqrt{m}}\boldsymbol{A}^T\right\| \left\|\frac{\tilde{\boldsymbol{w}}}{\sqrt{m}}\right\| \le (1+\delta)\|\tilde{\boldsymbol{w}}\|_\infty \le (1+\delta)\sqrt{\|\boldsymbol{w}\|_\infty},$$

when m/n is sufficiently large. Given the regime condition $\|\boldsymbol{h}\| \ge c_3\sqrt{\|\boldsymbol{w}\|_\infty}$, we further have

$$\begin{aligned}
&\|\nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z})\| \le \frac{(1+\delta)}{c_3}\|\boldsymbol{h}\|, \\
&\left|\left\langle \nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h} \right\rangle\right| \le \left\|\nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z})\right\| \cdot \|\boldsymbol{h}\| \le \frac{(1+\delta)}{c_3}\|\boldsymbol{h}\|^2.
\end{aligned}$$

We next bound the contribution of $\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})$. Introduce $\boldsymbol{q} = [q_1, \ldots, q_m]^T$, where

$$q_i := \left((|\boldsymbol{a}_i^T \boldsymbol{z}| - \sqrt{y_i})\mathbf{1}_{\mathcal{T}^i} - (|\boldsymbol{a}_i^T \boldsymbol{z}| - |\boldsymbol{a}_i^T \boldsymbol{x}|)\mathbf{1}_{\tilde{\mathcal{T}}^i}\right)\mathbf{1}_{\{i \in S\}}.$$

Then $|q_i| \leq 1.6\alpha'_h\|\boldsymbol{h}\|$, and $\|\boldsymbol{q}\| \leq \sqrt{sm} \cdot 1.6\alpha'_h\|\boldsymbol{h}\|$. We thus have

$$\begin{aligned}
&\left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\| = \frac{1}{m}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\| \leq 1.6(1+\delta)\sqrt{s}\alpha'_h\|\boldsymbol{h}\|, \\
&\left|\left\langle\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \leq \|\boldsymbol{h}\| \cdot \left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\| \leq 1.6(1+\delta)\sqrt{s}\alpha'_h\|\boldsymbol{h}\|^2.
\end{aligned}$$

Putting these together, one has

$$\begin{aligned}
\langle\nabla\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle &\geq \left\langle\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle - \left|\left\langle\nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| - \left|\left\langle\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \\
&\geq \left(0.88 - \zeta'_1 - \zeta'_2 - \epsilon - (1+\delta)(1/c_3 - 1.6\sqrt{s}\alpha'_h)\right)\|\boldsymbol{h}\|^2,
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla\mathcal{R}_{tr}(\boldsymbol{z})\| &\leq \left\|\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\| \\
&\leq \left(1.8 + \delta + (1+\delta)\cdot(1/c_3 + 1.6\sqrt{s}\alpha'_h)\right)\|\boldsymbol{h}\|. \qquad (4.41)
\end{aligned}$$

Thus, the RC is guaranteed if μ, λ, ϵ are chosen properly, c_0, c_3 are sufficiently large and s is sufficiently small.

• **Regime 2**: Once the iterate enters this regime with $\|\boldsymbol{h}\| \leq c_3\sqrt{\|\boldsymbol{w}\|_\infty}$, each gradient iterate may not reduce the estimation error. However, in this regime each move size $\mu\nabla\mathcal{R}_{tr}(\boldsymbol{z})$ is at most $\mathcal{O}(\sqrt{\|\boldsymbol{w}\|_\infty})$. Then the estimation error cannot increase by more than $\sqrt{\|\boldsymbol{w}\|_\infty}$ with a constant factor. Thus one has

$$\text{dist}\left(\boldsymbol{z} - \mu\nabla\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{x}\right) \leq c_5\sqrt{\|\boldsymbol{w}\|_\infty} \qquad (4.42)$$

for some constant c_5. As long as $\sqrt{\|\boldsymbol{w}\|_\infty}$ is sufficiently small, it is guaranteed that $c_5\sqrt{\|\boldsymbol{w}\|_\infty} \leq c_4\|\boldsymbol{x}\|$. If the iterate jumps out of *Regime 2*, it falls into *Regime 1*.

4.6.4 Proofs of Supporting Lemmas

Proof of Lemma 4.6

For simplicity, denote $\theta_p := \theta_p(F)$ and $\hat{\theta}_p := \theta_p(\{X_i\}_{i=1}^m)$. Since F' is continuous and positive, for an ϵ, there exists a constant δ_1 such that $\mathbb{P}(X \leq \theta_p - \epsilon) = p - \delta_1$, where $\delta_1 \in (\epsilon l, \epsilon L)$. Then one has

$$\mathbb{P}\left(\hat{\theta}_p < \theta_p - \epsilon\right) \overset{(a)}{=} \mathbb{P}\left(\sum_{i=1}^m \mathbf{1}_{\{X_i \leq \theta_p - \epsilon\}} \geq pm\right) = \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{X_i \leq \theta_p - \epsilon\}} \geq (p - \delta_1) + \delta_1\right)$$
$$\overset{(b)}{\leq} \exp(-2m\delta_1^2) \leq \exp(-2m\epsilon^2 l^2),$$

where (a) is due to the definition of the quantile function in (4.18) and (b) is due to the fact that $\mathbf{1}_{\{X_i \leq \theta_p - \epsilon\}} \sim$ Bernoulli$(p - \delta_1)$ i.i.d., followed by the Hoeffding inequality. Similarly, one can show for some $\delta_2 \in (\epsilon l, \epsilon L)$,

$$\mathbb{P}\left(\hat{\theta}_p > \theta_p + \epsilon\right) \leq \exp(-2m\delta_2^2) \leq \exp(-2m\epsilon^2 l^2).$$

Combining these two inequalities, one has the conclusion.

Proof of Lemma 4.7

It suffices to show that

$$|X_{(k)} - Y_{(k)}| \leq \max_l |X_l - Y_l|, \quad \forall k = 1, \cdots, n. \tag{4.43}$$

Case 1: $k = n$, suppose $X_{(n)} = X_i$ and $Y_{(n)} = Y_j$, i.e., X_i is the largest among $\{X_l\}_{l=1}^n$ and Y_j is the largest among $\{Y_l\}_{l=1}^n$. Then we have either $X_j \leq X_i \leq Y_j$ or $Y_i \leq Y_j \leq X_i$. Hence,

$$|X_{(n)} - Y_{(n)}| = |X_i - Y_j| \leq \max\{|X_i - Y_i|, |X_j - Y_j|\}.$$

Case 2: $k = 1$, suppose that $X_{(1)} = X_i$ and $Y_{(1)} = Y_j$. Similarly

$$|X_{(1)} - Y_{(1)}| = |X_i - Y_j| \leq \max\{|X_i - Y_i|, |X_j - Y_j|\}.$$

Case 3: $1 < k < n$, suppose that $X_{(k)} = X_i$, $Y_{(k)} = Y_j$, and without loss of generality assume that $X_i < Y_j$ (if $X_i = Y_j$, $0 = |X_{(k)} - Y_{(k)}| \leq \max_l |X_l - Y_l|$ holds trivially). We show the conclusion by contradiction.

Assume $|X_{(k)} - Y_{(k)}| > \max_l |X_l - Y_l|$. Then one must have $Y_i < Y_j$ and $X_j > X_i$ and $i \neq j$. Moreover for any $p < k$ and $q > k$, the index of $X_{(p)}$ cannot be equal to the index of $Y_{(q)}$; otherwise the assumption is violated.

Thus, all $Y_{(q)}$ for $q > k$ must share the same index set with $X_{(p)}$ for $p > k$. However, X_j, which is larger than X_i (thus if $X_j = X_{(k')}$, then $k' > k$), shares the same index with Y_j, where $Y_j = Y_{(k)}$. This yields contradiction.

Proof of Lemma 4.8

Assume that sm is an integer. Since there are sm corrupted samples in total, one can select at least $\lceil (p - s)m \rceil$ clean samples from the left p portion of ordered contaminated samples $\{\theta_{1/m}(\{X_i\}), \theta_{2/m}(\{X_i\}), \cdots, \theta_p(\{X_i\})\}$. Thus one has the left inequality. Furthermore, one can also select out at least $\lceil (1 - p - s)m \rceil$ clean samples from the right $1 - p$ portion of ordered contaminated samples $\{\theta_p(\{X_i\}), \cdots, \theta_1(\{X_i\})\}$. One has the right inequality.

Proof of Lemma 4.9

First we introduce some general facts for the distribution of the product of two correlated standard Gaussian random variables [143]. Let $u \sim \mathcal{N}(0, 1)$, $v \sim \mathcal{N}(0, 1)$, and their correlation coefficient be $\rho \in [-1, 1]$. Then the density of uv is given by

$$\phi_\rho(x) = \frac{1}{\pi\sqrt{1 - \rho^2}} \exp\left(\frac{\rho x}{1 - \rho^2}\right) K_0\left(\frac{|x|}{1 - \rho^2}\right), \quad x \neq 0,$$

where $K_0(\cdot)$ is the modified Bessel function of the second kind. Thus the density of $r = |uv|$ is

$$\psi_\rho(x) = \frac{1}{\pi\sqrt{1-\rho^2}}\left[\exp\left(\frac{\rho x}{1-\rho^2}\right) + \exp\left(-\frac{\rho x}{1-\rho^2}\right)\right] K_0\left(\frac{|x|}{1-\rho^2}\right), \; x > 0, \quad (4.44)$$

for $|\rho| < 1$. If $|\rho| = 1$, r becomes a χ_1^2 random variable, with the density

$$\psi_{|\rho|=1}(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} \exp(-x/2), \quad x > 0.$$

It can be seen from (4.44) that the density of r only relates to the correlation coefficient $\rho \in [-1, 1]$.

Let $\theta_{1/2}(\psi_\rho)$ be the $1/2$ quantile (median) of the distribution $\psi_\rho(x)$, and $\psi_\rho(\theta_{1/2})$ be the value of the function ψ_ρ at the point $\theta_{1/2}(\psi_\rho)$. Although it is difficult to derive the analytical expressions of $\theta_{1/2}(\psi_\rho)$ and $\psi_\rho(\theta_{1/2})$ due to the complicated form of ψ_ρ in (4.44), due to the continuity of $\psi_\rho(x)$ and $\theta_{1/2}(\psi_\rho)$, we can calculate them numerically, as illustrated in Figure 4.5. From the numerical calculation, one can see that both $\psi_\rho(\theta_{1/2})$ and $\theta_{1/2}(\psi_\rho)$

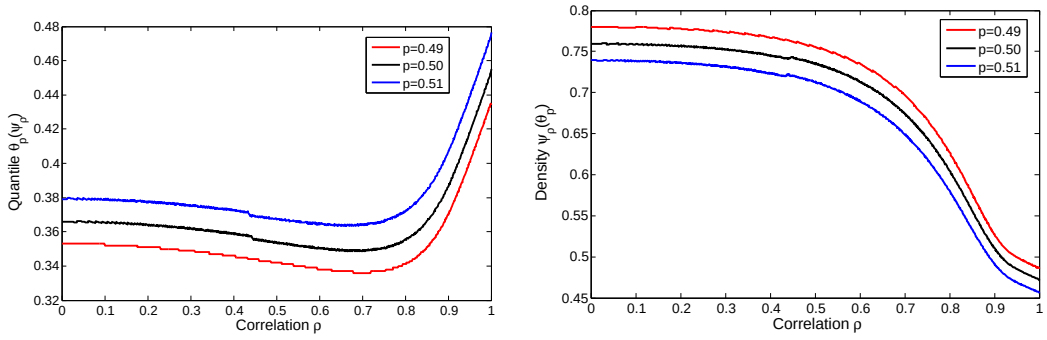


Fig. 4.5: Quantiles and density at quantiles of $\psi_\rho(x)$ across ρ.

are bounded from below and above for all $\rho \in [0, 1]$ ($\psi_\rho(\cdot)$ is symmetric over ρ, hence it is sufficient to consider $\rho \in [0, 1]$), satisfying

$$0.348 < \theta_{1/2}(\psi_\rho) < 0.455, \quad 0.47 < \psi_\rho(\theta_{1/2}) < 0.76. \quad (4.45)$$

Proof of Proposition 4.10

Denote $\tilde{y}_i := |\boldsymbol{a}_i^T \boldsymbol{x}|^2 + w_i$ for convenience. We first bound the concentration of med$(\{y_i\})$, also denoted by $\theta_{\frac{1}{2}}(\{y_i\})$. Lemma 4.8 yields

$$\theta_{\frac{1}{2}-s}(\{\tilde{y}_i\}) < \theta_{\frac{1}{2}}(\{y_i\}) < \theta_{\frac{1}{2}+s}(\{\tilde{y}_i\}). \tag{4.46}$$

Moreover, Lemma 4.7 indicates that

$$\theta_{\frac{1}{2}-s}(\{\tilde{y}_i\}) \geq \theta_{\frac{1}{2}-s}(\{|\boldsymbol{a}_i^T \boldsymbol{x}|^2\}) - \|\boldsymbol{w}\|_\infty, \tag{4.47}$$

$$\theta_{\frac{1}{2}+s}(\{\tilde{y}_i\}) \leq \theta_{\frac{1}{2}+s}(\{|\boldsymbol{a}_i^T \boldsymbol{x}|^2\}) + \|\boldsymbol{w}\|_\infty. \tag{4.48}$$

Observe that $\boldsymbol{a}_i^T \boldsymbol{x} = \tilde{a}_{i1}^2 \|\boldsymbol{x}\|^2$, where $\tilde{a}_{i1} = \boldsymbol{a}_i^T \boldsymbol{x}/\|\boldsymbol{x}\|$ is a standard Gaussian random variable. Thus $|\tilde{a}_{i1}|^2$ is a χ_1^2 random variable, whose cumulative distribution function is denoted as $K(x)$. Moreover by Lemma 4.6, for a small ϵ, one has $\left|\theta_{\frac{1}{2}-s}(\{|\tilde{a}_{i1}|^2\}) - \theta_{\frac{1}{2}-s}(K)\right| < \epsilon$ and $\left|\theta_{\frac{1}{2}+s}(\{|\tilde{a}_{i1}|^2\}) - \theta_{\frac{1}{2}+s}(K)\right| < \epsilon$ with probability $1 - 2\exp(-cm\epsilon^2)$ and c is a constant around 2×0.47^2 (see Figure 4.5). We note that $\theta_{\frac{1}{2}}(K) = 0.455$ and both $\theta_{\frac{1}{2}-s}(K)$ and $\theta_{\frac{1}{2}+s}(K)$ can be arbitrarily close to $\theta_{\frac{1}{2}}(K)$ simultaneously as long as s is small enough (independent of n). Thus, one has

$$\left(\theta_{\frac{1}{2}-s}(K) - \epsilon - c\right) \|\boldsymbol{x}\|^2 < \theta_{\frac{1}{2}}(\{y_i\}) < \left(\theta_{\frac{1}{2}+s}(K) + \epsilon + c\right) \|\boldsymbol{x}\|^2, \tag{4.49}$$

with probability at least $1 - \exp(-cm\epsilon^2)$. For the sake of simplicity, we introduce two new notations $\zeta_s := \theta_{\frac{1}{2}-s}(K)$ and $\zeta^s := \theta_{\frac{1}{2}+s}(K)$. Specifically for the instance of $s = 0.01$, one has $\zeta_s = 0.434$ and $\zeta^s = 0.477$. It is easy to see that $\zeta^s - \zeta_s$ can be arbitrarily small if s is small enough.

We next estimate the direction of $\boldsymbol{x}$, assuming $\|\boldsymbol{x}\| = 1$. On the event that (4.49) holds,

the truncation function has the following bounds,

$$\mathbf{1}_{\{y_i\le\alpha_y^2\theta_{1/2}(\{y_i\})/0.455\}} \le \mathbf{1}_{\left\{y_i\le\alpha_y^2(\zeta^s+\epsilon)/0.455\right\}} \le \mathbf{1}_{\left\{(\boldsymbol{a}_i^T\boldsymbol{x})^2\le\alpha_y^2(\zeta^s+\epsilon+c)/0.455\right\}}$$
$$\mathbf{1}_{\{y_i\le\alpha_y^2\theta_{1/2}(\{y_i\})/0.455\}} \ge \mathbf{1}_{\left\{y_i\le\alpha_y^2(\zeta_s-\epsilon)/0.455\right\}} \ge \mathbf{1}_{\left\{(\boldsymbol{a}_i^T\boldsymbol{x})^2\le\alpha_y^2(\zeta_s-\epsilon-c)/0.455\right\}}.$$

On the other hand, denote the support of the outliers as S, and we have

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T\tilde{y}_i\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})^2\le\alpha_y^2\theta_{1/2}(\{y_i\})/0.455\}} + \frac{1}{m}\sum_{i\in S}\boldsymbol{a}_i\boldsymbol{a}_i^T y_i\mathbf{1}_{\{y_i\le\alpha_y^2\theta_{1/2}(\{y_i\})/0.455\}}.$$

Consequently, one can bound $\boldsymbol{Y}$ as

$$\begin{aligned}\boldsymbol{Y}_1 &:= \frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T(\boldsymbol{a}_i^T\boldsymbol{x})^2\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})^2\le\alpha_y^2(\zeta_s-\epsilon-c)/0.455\}} - c\cdot\frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T \preceq \boldsymbol{Y}\\ &\preceq \frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T(\boldsymbol{a}_i^T\boldsymbol{x})^2\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})^2\le\alpha_y^2(\zeta^s+\epsilon+c)/0.455\}} + c\cdot\frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T\\ &\quad + \frac{1}{m}\sum_{i\in S}\boldsymbol{a}_i\boldsymbol{a}_i^T\alpha_y^2(\zeta^s+\epsilon+c)/0.455 =: \boldsymbol{Y}_2,\end{aligned}$$

where we have

$$\begin{aligned}\mathrm{E}[\boldsymbol{Y}_1] &= (1-s)(\beta_1\boldsymbol{x}\boldsymbol{x}^T + \beta_2\boldsymbol{I} - c\boldsymbol{I}),\\ \mathrm{E}[\boldsymbol{Y}_2] &= (1-s)(\beta_3\boldsymbol{x}\boldsymbol{x}^T + \beta_4\boldsymbol{I} + c\boldsymbol{I}) + s\alpha_y^2\frac{(\zeta^s+\epsilon)}{0.455}\boldsymbol{I},\end{aligned}$$

with

$$\begin{aligned}\beta_1 &:= \mathrm{E}\left[\xi^4\mathbf{1}_{\left\{|\xi|\le\alpha_y\sqrt{(\zeta_s-\epsilon-c)/0.455}\right\}}\right] - \mathrm{E}\left[\xi^2\mathbf{1}_{\left\{|\xi|\le\alpha_y\sqrt{(\zeta_s-\epsilon-c)/0.455}\right\}}\right]\\ \beta_2 &:= \mathrm{E}\left[\xi^2\mathbf{1}_{\left\{|\xi|\le\alpha_y\sqrt{(\zeta_s-\epsilon-c)/0.455}\right\}}\right]\\ \beta_3 &:= \mathrm{E}\left[\xi^4\mathbf{1}_{\left\{|\xi|\le\alpha_y\sqrt{(\zeta^s+\epsilon+c)/0.455}\right\}}\right] - \mathrm{E}\left[\xi^2\mathbf{1}_{\left\{|\xi|\le\alpha_y\sqrt{(\zeta^s+\epsilon+c)/0.455}\right\}}\right]\\ \beta_4 &:= \mathrm{E}\left[\xi^2\mathbf{1}_{\left\{|\xi|\le\alpha_y\sqrt{(\zeta^s+\epsilon+c)/0.455}\right\}}\right]\end{aligned}$$

where $\xi \sim \mathcal{N}(0,1)$.

Applying standard results on random matrices with non-isotropic sub-Gaussian rows [142, equation (5.26)] and noticing that $\boldsymbol{a}_i\boldsymbol{a}_i^T(\boldsymbol{a}_i^T\boldsymbol{x})^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|\le c\}}$ can be rewritten as $\boldsymbol{b}_i\boldsymbol{b}_i^T$ where $\boldsymbol{b}_i := \boldsymbol{a}_i(\boldsymbol{a}_i^T\boldsymbol{x})\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|\le c\}}$ is sub-Gaussian, one can obtain

$$\|\boldsymbol{Y}_1 - \mathrm{E}[\boldsymbol{Y}_1]\| \le \delta, \quad \|\boldsymbol{Y}_2 - \mathrm{E}[\boldsymbol{Y}_2]\| \le \delta \tag{4.50}$$

with probability $1 - \exp(-\Omega(m))$, provided that m/n exceeds some large constant. Furthermore, when ϵ, c and s are sufficiently small, one further has $\|\mathrm{E}[\boldsymbol{Y}_1] - \mathrm{E}[\boldsymbol{Y}_2]\| \le \delta$. Putting these together, one has

$$\|\boldsymbol{Y} - (1-s)(\beta_1\boldsymbol{x}\boldsymbol{x}^T + \beta_2\boldsymbol{I} - c\boldsymbol{I})\| \le 3\delta. \tag{4.51}$$

Let $\tilde{\boldsymbol{z}}^{(0)}$ be the normalized leading eigenvector of $\boldsymbol{Y}$. Repeating the same argument as in [1, Section 7.8] and taking δ, ϵ to be sufficiently small, one has

$$\mathrm{dist}(\tilde{\boldsymbol{z}}^{(0)}, \boldsymbol{x}) \le \tilde{\delta}, \tag{4.52}$$

for a given $\tilde{\delta} > 0$, as long as m/n exceeds some large constant.

Furthermore let $\boldsymbol{z}^{(0)} = \sqrt{\mathrm{med}\{y_i\}/0.455}\,\tilde{\boldsymbol{z}}^{(0)}$ to handle cases $\|\boldsymbol{x}\| \ne 1$. By the bound (4.49), one has

$$\begin{aligned}\left|\frac{\mathrm{med}(\{y_i\})}{0.455} - \|\boldsymbol{x}\|^2\right| &\le \max\left\{\left|\frac{\zeta_s - \epsilon - c}{0.455} - 1\right|, \left|\frac{\zeta^s + \epsilon + c}{0.455} - 1\right|\right\}\|\boldsymbol{x}\|^2 \\ &\le \frac{\zeta^s - \zeta_s + 2\epsilon + 2c}{0.455}\|\boldsymbol{x}\|^2. \end{aligned}\tag{4.53}$$

Thus

$$\mathrm{dist}(\boldsymbol{z}^{(0)}, \boldsymbol{x}) \le \frac{\zeta^s - \zeta_s + 2\epsilon + 2c}{0.455}\|\boldsymbol{x}\| + \tilde{\delta}\|\boldsymbol{x}\| \le \frac{1}{11}\|\boldsymbol{x}\|$$

as long as s and c are small enough constants. □

Proof of Proposition 4.12

We show that the sample median used in the truncation rule concentrates at the level $\|\boldsymbol{z}-\boldsymbol{x}\|\|\boldsymbol{z}\|$. Along the way, we also establish that the sample quantiles around the median are also concentrated at the level $\|\boldsymbol{z}-\boldsymbol{x}\|\|\boldsymbol{z}\|$.

We first show that for a fixed pair $\boldsymbol{z}$ and $\boldsymbol{x}$, (4.23) holds with high probability. For simplicity of notation, we let $\boldsymbol{h} := \boldsymbol{z}-\boldsymbol{x}$. Let $r_i = |(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2|$. Then r_i's are i.i.d. copies of a random variable r, where $r = |(\boldsymbol{a}^T\boldsymbol{x})^2 - (\boldsymbol{a}^T\boldsymbol{z})^2|$ with the entries of $\boldsymbol{a}$ composed of i.i.d. standard Gaussian random variables. Note that the distribution of r is fixed once given $\boldsymbol{h}$ and $\boldsymbol{z}$. Let $\boldsymbol{x}(1)$ denote the first element of a generic vector $\boldsymbol{x}$, and $\boldsymbol{x}_{-1}$ denote the remaining vector of $\boldsymbol{x}$ after eliminating the first element. Let $\boldsymbol{U}_h$ be an orthonormal matrix with first row being $\boldsymbol{h}^T/\|\boldsymbol{h}\|$, $\tilde{\boldsymbol{a}} = \boldsymbol{U}_h\boldsymbol{a}$, and $\tilde{\boldsymbol{z}} = \boldsymbol{U}_h\boldsymbol{z}$. Similarly, define $\boldsymbol{U}_{\tilde{z}_{-1}}$ and let $\tilde{\boldsymbol{b}} = \boldsymbol{U}_{\tilde{z}_{-1}}\tilde{\boldsymbol{a}}_{-1}$. Then $\tilde{\boldsymbol{a}}(1)$ and $\tilde{\boldsymbol{b}}(1)$ are independent standard normal random variables. We further express r as follows.

$$
\begin{aligned}
r &= |(\boldsymbol{a}^T\boldsymbol{z})^2 - (\boldsymbol{a}^T\boldsymbol{x})^2| \\
&= |(2\boldsymbol{a}^T\boldsymbol{z} - \boldsymbol{a}^T\boldsymbol{h})(\boldsymbol{a}^T\boldsymbol{h})| \\
&= |(2\tilde{\boldsymbol{a}}^T\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{a}}(1)\|\boldsymbol{h}\|)(\tilde{\boldsymbol{a}}(1)\|\boldsymbol{h}\|)| \\
&= |(2\boldsymbol{h}^T\boldsymbol{z} - \|\boldsymbol{h}\|^2)\tilde{\boldsymbol{a}}(1)^2 + 2(\tilde{\boldsymbol{a}}_{-1}^T\tilde{\boldsymbol{z}}_{-1})(\tilde{\boldsymbol{a}}(1)\|\boldsymbol{h}\|)| \\
&= |(2\boldsymbol{h}^T\boldsymbol{z} - \|\boldsymbol{h}\|^2)\tilde{\boldsymbol{a}}(1)^2 + 2\tilde{\boldsymbol{b}}(1)\|\tilde{\boldsymbol{z}}_{-1}\|\tilde{\boldsymbol{a}}(1)\|\boldsymbol{h}\|| \\
&= |(2\boldsymbol{h}^T\boldsymbol{z} - \|\boldsymbol{h}\|^2)\tilde{\boldsymbol{a}}(1)^2 + 2\sqrt{\|\boldsymbol{z}\|^2 - \tilde{\boldsymbol{z}}(1)^2}\tilde{\boldsymbol{a}}(1)\tilde{\boldsymbol{b}}(1)\|\boldsymbol{h}\|| \\
&= \left|\left(2\frac{\boldsymbol{h}^T\boldsymbol{z}}{\|\boldsymbol{h}\|\|\boldsymbol{z}\|} - \frac{\|\boldsymbol{h}\|}{\|\boldsymbol{z}\|}\right)\tilde{\boldsymbol{a}}(1)^2 + 2\sqrt{1-\left(\frac{\boldsymbol{h}^T\boldsymbol{z}}{\|\boldsymbol{h}\|\|\boldsymbol{z}\|}\right)^2}\tilde{\boldsymbol{a}}(1)\tilde{\boldsymbol{b}}(1)\right| \cdot \|\boldsymbol{h}\|\|\boldsymbol{z}\| \\
&=: \left|(2\cos(\omega) - t)\tilde{\boldsymbol{a}}(1)^2 + 2\sqrt{1-\cos^2(\omega)}\tilde{\boldsymbol{a}}(1)\tilde{\boldsymbol{b}}(1)\right| \cdot \|\boldsymbol{h}\|\|\boldsymbol{z}\| \\
&=: |u\tilde{v}| \cdot \|\boldsymbol{h}\|\|\boldsymbol{z}\|
\end{aligned}
$$

where ω is the angle between $\boldsymbol{h}$ and $\boldsymbol{z}$, and $t = \|\boldsymbol{h}\|/\|\boldsymbol{z}\| < 1/11$. Consequently, $u = \tilde{\boldsymbol{a}}(1) \sim \mathcal{N}(0,1)$ and $\tilde{v} = (2\cos(\omega) - t)\tilde{\boldsymbol{a}}(1) + 2|\sin(\omega)|\tilde{\boldsymbol{b}}(1)$ is also a Gaussian random variable with variance $3.6 < \mathsf{Var}(\tilde{v}) < 4$ under the assumption $t < 1/11$.

Let $v = \tilde{v}/\sqrt{\mathsf{Var}(\tilde{v})}$, and then $v \sim \mathcal{N}(0,1)$. Furthermore, let $r' = |uv|$. Denote the density function of r' as $\psi_\rho(\cdot)$ and the $1/2$-quantile point of r' as $\theta_{1/2}(\psi_\rho)$. By Lemma 4.9, we have

$$0.47 < \psi_\rho(\theta_{1/2}) < 0.76, \quad 0.348 < \theta_{1/2}(\psi_\rho) < 0.455. \tag{4.54}$$

By Lemma 4.6, we have with probability at least $1 - 2\exp(-cm\epsilon^2)$ (here c is around 2×0.47^2),

$$0.348 - \epsilon < \mathsf{med}(\{r'_i\}_{i=1}^m) < 0.455 + \epsilon. \tag{4.55}$$

The same arguments carry over to other quantiles $\theta_{0.49}(\{r'_i\})$ and $\theta_{0.51}(\{r'_i\})$. From Figure. 4.5, we observe that for $\rho \in [0,1]$

$$0.45 < \psi_\rho(\theta_{0.49}), \psi_\rho(\theta_{0.51}) < 0.78, \quad 0.336 < \theta_{0.49}(\psi_\rho), \theta_{0.51}(\psi_\rho) < 0.477 \tag{4.56}$$

and then we have with probability at least $1 - 2\exp(-cm\epsilon^2)$ (here c is around 2×0.45^2),

$$0.336 - \epsilon < \theta_{0.49}(\{r'_m\}), \theta_{0.51}(\{r'_m\}) < 0.477 + \epsilon. \tag{4.57}$$

Hence, by multiplying by $\sqrt{\mathsf{Var}(\tilde{v})}$, we have with probability $1-2\exp(-cm\epsilon^2)$,

$$(0.65-\epsilon)\|\boldsymbol{z}-\boldsymbol{x}\|\|\boldsymbol{z}\| \le \text{med}\left(\left\{|(\boldsymbol{a}_i^T\boldsymbol{z})^2-(\boldsymbol{a}_i^T\boldsymbol{x})^2|\right\}\right) \le (0.91+\epsilon)\|\boldsymbol{z}-\boldsymbol{x}\|\|\boldsymbol{z}\|, \tag{4.58}$$

$$(0.63-\epsilon)\|\boldsymbol{z}-\boldsymbol{x}\|\|\boldsymbol{z}\| \le \theta_{0.49},\theta_{0.51}\left(\left\{|(\boldsymbol{a}_i^T\boldsymbol{z})^2-(\boldsymbol{a}_i^T\boldsymbol{x})^2|\right\}\right) \le (0.95+\epsilon)\|\boldsymbol{z}-\boldsymbol{x}\|\|\boldsymbol{z}\|. \tag{4.59}$$

We note that, to keep notation simple, c and ϵ may vary line by line within constant factors.

Up to now, we prove that for any fixed $\boldsymbol{z}$ and $\boldsymbol{x}$, the median or neighboring quantiles of $\left\{|(\boldsymbol{a}_i^T\boldsymbol{z})^2-(\boldsymbol{a}_i^T\boldsymbol{x})^2|\right\}$ are upper and lower bounded by $\|\boldsymbol{z}-\boldsymbol{x}\|\|\boldsymbol{z}\|$ times constant factors. To prove (4.23) for all $\boldsymbol{z}$ and $\boldsymbol{x}$ with $\|\boldsymbol{z}-\boldsymbol{x}\| \le \frac{1}{11}\|\boldsymbol{z}\|$, we use the net covering argument. Still we argue for median first and the same arguments carry over to other quantiles.

To proceed, we restate (4.58) as

$$(0.65-\epsilon) \le \text{med}\left(\left\{\left|\left(\frac{2(\boldsymbol{a}_i^T\boldsymbol{z})}{\|\boldsymbol{z}\|}-\frac{\boldsymbol{a}_i^T\boldsymbol{h}}{\|\boldsymbol{h}\|}\frac{\|\boldsymbol{h}\|}{\|\boldsymbol{z}\|}\right)\frac{\boldsymbol{a}_i^T\boldsymbol{h}}{\|\boldsymbol{h}\|}\right|\right\}\right) \le (0.91+\epsilon) \tag{4.60}$$

holds with probability at least $1-2\exp(-cm\epsilon^2)$ for a given pair $\boldsymbol{h},\boldsymbol{z}$ satisfying $\|\boldsymbol{h}\|/\|\boldsymbol{z}\| \le 1/11$.

Let $\tau=\epsilon/(6n+6m)$, let $\mathcal{S}_\tau$ be a τ-net covering the unit sphere, $\mathcal{L}_\tau$ be a τ-net covering a line with length $1/11$, and set

$$\mathcal{N}_\tau = \{(\boldsymbol{z}_0,\boldsymbol{h}_0,t_0) : (\boldsymbol{z}_0,\boldsymbol{h}_0,t_0) \in \mathcal{S}_\tau \times \mathcal{S}_\tau \times \mathcal{L}_\tau\}. \tag{4.61}$$

One has cardinality bound (i.e., the upper bound on the covering number) $|\mathcal{N}_\tau| \le (1+2/\tau)^{2n}/(11\tau) < (1+2/\tau)^{2n+1}$. Taking the union bound, we have

$$(0.65-\epsilon) \le \text{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}_0)-(\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0||\boldsymbol{a}_i^T\boldsymbol{h}_0|\right\}\right) \le (0.91+\epsilon), \quad \forall(\boldsymbol{z}_0,\boldsymbol{h}_0,t_0)\in\mathcal{N}_\epsilon \tag{4.62}$$

with probability at least $1-(1+2/\tau)^{2n+1}\exp(-cm\epsilon^2)$.

We next argue that (4.62) holds with probability $1-c_1\exp(-c_2m\epsilon^2)$ for some constants c_1, c_2 as long as $m \geq c_0(\epsilon^{-2}\log\epsilon^{-1})n\log n$ for sufficiently large constant c_0. To prove this claim, we first observe

$$(1+2/\tau)^{2n+1} \asymp \exp(2n(\log(n+m)+\log 12+\log(1/\epsilon))) \asymp \exp(2n(\log m)).$$

We note that once ϵ is chosen, it is fixed in the whole proof and does not scale with m or n. For simplicity, assume that $\epsilon < 1/e$. Fix some positive constant $c' < c - c_2$. It then suffices to show that there exists a large constant c_0 such that if $m \geq c_0(\epsilon^{-2}\log\epsilon^{-1})n\log n$, then

$$2n\log m < c'm\epsilon^2. \tag{4.63}$$

For any fixed n, if (4.63) holds for some m and $m > (2/c')\epsilon^{-2}n$, then (4.63) always holds for larger m, because

$$\begin{aligned} 2n\log(m+1) &= 2n\log m + 2n(\log(m+1)-\log m) = 2n\log m + \frac{2n}{m}\log(1+\frac{1}{m})^m \\ &\leq 2n\log m + \frac{2n}{m} \leq c'm\epsilon^2 + c'\epsilon^2 = c'(m+1)\epsilon^2. \end{aligned}$$

Next, we can always find a constant c_0 such that (4.63) holds for $m = c_0(\epsilon^{-2}\log\epsilon^{-1})n\log n$ for any n. Such c_0 can be easily found for large n. For example, $c_0 = 4/c'$ is a valid option if

$$(4/c')(\epsilon^{-2}\log\epsilon^{-1})n\log n < n^2. \tag{4.64}$$

Moreover, since the number of n that violates (4.64) is finite, the maximum over all such c_0 serves the purpose. Next, one needs to bound

$$\left|\text{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}_0)-(\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0||\boldsymbol{a}_i^T\boldsymbol{h}_0|\right\}\right) - \text{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z})-(\boldsymbol{a}_i^T\boldsymbol{h})t||\boldsymbol{a}_i^T\boldsymbol{h}|\right\}\right)\right|$$

for any $\|\boldsymbol{z}-\boldsymbol{z}_0\|<\tau, \|\boldsymbol{z}-\boldsymbol{z}_0\|<\tau$ and $\|t-t_0\|<\tau$.

By Lemma 4.7 and the inequality $\Big||x|-|y|\Big| \le |x-y|$, we have

$$
\begin{aligned}
&\left|\text{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}_0)-(\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0||\boldsymbol{a}_i^T\boldsymbol{h}_0|\right\}\right)-\text{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z})-(\boldsymbol{a}_i^T\boldsymbol{h})t||\boldsymbol{a}_i^T\boldsymbol{h}|\right\}\right)\right| \\
&\le \max_{i\in[m]}\left|\left(2(\boldsymbol{a}_i^T\boldsymbol{z}_0)-(\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0\right)(\boldsymbol{a}_i^T\boldsymbol{h}_0)-\left(2(\boldsymbol{a}_i^T\boldsymbol{z})-(\boldsymbol{a}_i^T\boldsymbol{h})t\right)(\boldsymbol{a}_i^T\boldsymbol{h})\right| \\
&\le \max_{i\in[m]}\left|\left(2(\boldsymbol{a}_i^T\boldsymbol{z}_0)-(\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0\right)(\boldsymbol{a}_i^T\boldsymbol{h}_0)-\left(2(\boldsymbol{a}_i^T\boldsymbol{z})-(\boldsymbol{a}_i^T\boldsymbol{h})t\right)(\boldsymbol{a}_i^T\boldsymbol{h}_0)\right| \\
&\quad+\max_{i\in[m]}\left|\left(2(\boldsymbol{a}_i^T\boldsymbol{z})-(\boldsymbol{a}_i^T\boldsymbol{h})t\right)(\boldsymbol{a}_i^T\boldsymbol{h}_0)-\left(2(\boldsymbol{a}_i^T\boldsymbol{z})-(\boldsymbol{a}_i^T\boldsymbol{h})t\right)(\boldsymbol{a}_i^T\boldsymbol{h})\right| \\
&\le \max_{i\in[m]}\left(\left|2\boldsymbol{a}_i^T(\boldsymbol{z}_0-\boldsymbol{z})\right|+\left|(\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0-(\boldsymbol{a}_i^T\boldsymbol{h})t\right|\right)\left|\boldsymbol{a}_i^T\boldsymbol{h}_0\right| \\
&\quad+\max_{i\in[m]}\left|2(\boldsymbol{a}_i^T\boldsymbol{z})-(\boldsymbol{a}_i^T\boldsymbol{h})t\right||\boldsymbol{a}_i^T(\boldsymbol{h}_0-\boldsymbol{h})| \\
&\le \max_{i\in[m]}\|\boldsymbol{a}_i\|^2(3+t)\tau+\max_{i\in[m]}\|\boldsymbol{a}_i\|^2(2+t)\tau \le \max_{i\in[m]}\|\boldsymbol{a}_i\|^2(5+2t)\tau
\end{aligned}
$$

On the event $E_1 := \left\{\max_{i\in[m]}\|\boldsymbol{a}_i\|^2 \le m+n\right\}$, one can show that

$$
\begin{aligned}
&\left|\text{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}_0)-(\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0||\boldsymbol{a}_i^T\boldsymbol{h}_0|\right\}\right)-\text{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z})-(\boldsymbol{a}_i^T\boldsymbol{h})t||\boldsymbol{a}_i^T\boldsymbol{h}|\right\}\right)\right| \\
&\quad<6(m+n)\tau<\epsilon.
\end{aligned}
$$

We claim that E_1 holds with probability at least $1-m\exp(-m/8)$ if $m>n$. This can be argued as follows. Note that $\|\boldsymbol{a}_i\|^2=\sum_{j=1}^n \boldsymbol{a}_i(j)^2$, where $\boldsymbol{a}_i(j)$ is the j-th element of $\boldsymbol{a}_i$. Hence, $\|\boldsymbol{a}_i\|^2$ is a sum of n i.i.d. χ_1^2 random variables. Applying the Bernstein-type inequality [142, Corollary 5.17] and observing that the sub-exponential norm of χ_1^2 is smaller than 2, we have

$$
\mathbb{P}\left\{\|\boldsymbol{a}_i\|^2 \ge m+n\right\} \le \exp(-m/8). \tag{4.65}
$$

Then a union bound concludes the claim.

Further note that (4.62) holds on an event E_2, which has probability $1-c_1\exp(-c_2m\epsilon^2)$

as long as $m \geq c_0(\epsilon^{-2}\log\frac{1}{\epsilon})n\log n$. On the intersection of E_1 and E_2, inequality for $\theta_{\frac{1}{2}}$ (i.e., median) in (4.23) holds. Such net covering arguments can also carry over to show that inequalities of $\theta_{0.49}$ and $\theta_{0.51}$ in (4.23) also hold for all $\boldsymbol{x}$ and $\boldsymbol{z}$ obeying $\|\boldsymbol{x}-\boldsymbol{z}\| \leq \frac{1}{11}\|\boldsymbol{z}\|$.

Proof of Proposition 4.13

The proof adapts that of [61, Proposition 2]. We outline the main steps for completeness. Observe that for the noise-free case, $y_i = (\boldsymbol{a}_i^T\boldsymbol{x})^2$. We obtain

$$\begin{aligned}\nabla\ell_{tr}(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^{m}\frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2-(\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}}\boldsymbol{a}_i\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i}\\ &= \frac{1}{m}\sum_{i=1}^{m}2(\boldsymbol{a}_i^T\boldsymbol{h})\boldsymbol{a}_i\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i}-\frac{1}{m}\sum_{i=1}^{m}\frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\boldsymbol{a}_i^T\boldsymbol{z}}\boldsymbol{a}_i\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i}. \qquad (4.66)\end{aligned}$$

One expects the contribution of the second term in (4.66) to be small as $\|\boldsymbol{h}\|/\|\boldsymbol{z}\|$ decreases.

For each i, we introduce two new events

$$\begin{aligned}\mathcal{E}_3^i &:= \{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \leq 0.6\alpha_h\|\boldsymbol{h}\|\cdot|\boldsymbol{a}_i^T\boldsymbol{z}|\},\\ \mathcal{E}_4^i &:= \{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2-(\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \leq 1.0\alpha_h\|\boldsymbol{h}\|\cdot|\boldsymbol{a}_i^T\boldsymbol{z}|\}.\end{aligned}$$

One the event that Proposition 4.12 holds, the following inclusion property

$$\mathcal{E}_3^i \subseteq \mathcal{E}_2^i \subseteq \mathcal{E}_4^i \qquad (4.67)$$

is true for all i, where $\mathcal{E}_2^i$ is defined in Algorithm 3. It is easier to work with these new events because $\mathcal{E}_3^i$'s (resp. $\mathcal{E}_4^i$'s) are statistically independent across i for any fixed $\boldsymbol{x}$ and $\boldsymbol{z}$. To further decouple the quadratic inequalities in $\mathcal{E}_3^i$ and $\mathcal{E}_4^i$ into linear inequalities, we introduce two more events and state their properties in the following lemma.

Lemma 4.18 (Lemma 3 in [61]). *For any $\gamma > 0$, define*

$$\mathcal{D}_\gamma^i := \{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \le \gamma\|\boldsymbol{h}\|\|\boldsymbol{a}_i^T\boldsymbol{z}|\}, \tag{4.68}$$

$$\mathcal{D}_\gamma^{i,1} := \left\{\frac{|\boldsymbol{a}_i^T\boldsymbol{h}|}{\|\boldsymbol{h}\|} \le \gamma\right\}, \tag{4.69}$$

$$\mathcal{D}_\gamma^{i,2} := \left\{\left|\frac{\boldsymbol{a}_i^T\boldsymbol{h}}{\|\boldsymbol{h}\|} - \frac{2\boldsymbol{a}_i^T\boldsymbol{z}}{\|\boldsymbol{h}\|}\right| \le \gamma\right\}. \tag{4.70}$$

On the event $\mathcal{E}_1^i$ defined in Algorithm 3, the quadratic inequality specifying $\mathcal{D}_\gamma^i$ implicates that $\boldsymbol{a}_i^T\boldsymbol{h}$ belongs to two intervals centered around 0 and $2\boldsymbol{a}_i^T\boldsymbol{z}$, respectively, i.e., $\mathcal{D}_\gamma^{i,1}$ and $\mathcal{D}_\gamma^{i,2}$. The following inclusion property holds

$$\left(\mathcal{D}^{i,1}_{\frac{\gamma}{1+\sqrt{2}}} \cap \mathcal{E}_1^i\right) \cup \left(\mathcal{D}^{i,2}_{\frac{\gamma}{1+\sqrt{2}}} \cap \mathcal{E}_1^i\right) \subseteq \mathcal{D}_\gamma^i \cap \mathcal{E}_1^i \subseteq \left(\mathcal{D}_\gamma^{i,1} \cap \mathcal{E}_1^i\right) \cup \left(\mathcal{D}_\gamma^{i,2} \cap \mathcal{E}_1^i\right). \tag{4.71}$$

Specifically, following the two inclusion properties (4.67) and (4.71), we have

$$\mathcal{D}_{\gamma_3}^{i,1} \cap \mathcal{E}_{1,\gamma_3}^i \subseteq \mathcal{E}_3^i \cap \mathcal{E}_1^i \subseteq \mathcal{E}_2^i \cap \mathcal{E}_1^i \subseteq \mathcal{E}_4^i \cap \mathcal{E}_1^i \subseteq (\mathcal{D}_{\gamma_4}^{i,1} \cup \mathcal{D}_{\gamma_4}^{i,2}) \cap \mathcal{E}_1^i \tag{4.72}$$

where the parameters γ_3, γ_4 are given by

$$\gamma_3 := 0.248\alpha_h, \quad \text{and} \quad \gamma_4 := \alpha_h.$$

Further using the identity (4.66), we have the following lower bound

$$\langle \nabla \ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle \ge \frac{2}{m}\sum_{i=1}^m (\boldsymbol{a}_i^T\boldsymbol{h})^2 \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{D}_{\gamma_3}^{i,1}} - \frac{1}{m}\sum_{i=1}^m \frac{|\boldsymbol{a}_i^T\boldsymbol{h}|^3}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\mathbf{1}_{\mathcal{D}_{\gamma_4}^{i,1}\cap\mathcal{E}_1^i} - \frac{1}{m}\sum_{i=1}^m \frac{|\boldsymbol{a}_i^T\boldsymbol{h}|^3}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\mathbf{1}_{\mathcal{D}_{\gamma_4}^{i,2}\cap\mathcal{E}_1^i}. \tag{4.73}$$

The three terms in (4.73) can be bounded following Lemmas 4, 5, and 6 in [61], which concludes the proof.

Proof of Proposition 4.15

Observe that

$$||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|| = \begin{cases} |\boldsymbol{a}_i^T\boldsymbol{h}|, & \text{if } (\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}) \geq 0; \\ |2\boldsymbol{a}_i^T\boldsymbol{x} + \boldsymbol{a}_i^T\boldsymbol{h}|, & \text{if } (\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}) < 0. \end{cases}$$

The following lemma states that $\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}) < 0\}$ are rare events when $\|\boldsymbol{x} - \boldsymbol{z}\|$ is small. Hence, med $\left(\left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}_{i=1}^m\right)$ can be viewed as med$(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}_{i=1}^m)$ with a small perturbation.

Lemma 4.19. *If $m > c_0 n$, then with probability at least $1 - c_1\exp(-c_2 m)$,*

$$\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<0\}} < 0.05 \tag{4.74}$$

holds for all $\boldsymbol{z}, \boldsymbol{x}$ satisfying $\|\boldsymbol{z} - \boldsymbol{x}\| < \frac{1}{11}\|\boldsymbol{x}\|$.

Proof. See Section 4.6.4. □

By Lemma 4.8 and Lemma 4.19, we have

$$\theta_{p-0.05}\left(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}\right) \leq \theta_p\left(\left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}\right) \leq \theta_{p+0.05}\left(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}\right) \tag{4.75}$$

for all $\boldsymbol{x}$ and $\boldsymbol{z}$ satisfying $\|\boldsymbol{x} - \boldsymbol{z}\| \leq \frac{1}{11}\|\boldsymbol{z}\|$ with high probability.

For the model (4.1) with a fraction s of outliers, due to Lemma 4.8, we have that

$$\theta_{\frac{1}{2}-s}(\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\}) \leq \theta_{\frac{1}{2}}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}||\}) \leq \theta_{\frac{1}{2}+s}(\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\}). \tag{4.76}$$

Combining with (4.75), we obtain that

$$\theta_{0.45-s}(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}) \leq \theta_{\frac{1}{2}}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}||\}) \leq \theta_{0.55+s}(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}). \tag{4.77}$$

Next it suffices to show that $\theta_{0.45-s}, \theta_{0.55+s}(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\})$ are on the order of $\|\boldsymbol{h}\|$ for small s.

Let $\tilde{a}_i = |\boldsymbol{a}_i^T \boldsymbol{h}|/\|\boldsymbol{h}\|$. Then $\tilde{a}_i$'s are i.i.d. copies of a *folded standard Gaussian* random variable (i.e., $|\xi|$ where $\xi \sim \mathcal{N}(0,1)$). We use $\phi(\cdot)$ to denote the density of folded standard Gaussian distribution.

For $s = 0.01$, we calculate that

$$\phi(\theta_{0.44}) = 0.67, \quad \phi(\theta_{0.45}) = 0.67, \quad \phi(\theta_{0.55}) = 0.60, \quad \phi(\theta_{0.56}) = 0.59 \tag{4.78}$$

$$\theta_{0.44}(\phi) = 0.58, \quad \theta_{0.45}(\phi) = 0.6, \quad \theta_{0.55}(\phi) = 0.76, \quad \theta_{0.56}(\phi) = 0.78. \tag{4.79}$$

By Lemma 4.6, the sample quantiles concentrate on population quantiles. Thus, for any fixed pair $(\boldsymbol{x}, \boldsymbol{z})$,

$$(0.6 - \epsilon)\|\boldsymbol{h}\| \le \theta_{1/2}(\{\big||\boldsymbol{a}_i^T \boldsymbol{x}| - |\boldsymbol{a}_i^T \boldsymbol{z}|\big|\}_{i=1}^m) \le (0.76 + \epsilon)\|\boldsymbol{h}\|, \tag{4.80}$$

holds with probability at least $1 - 2\exp(-cm\epsilon^{-2})$.

Following the argument of net covering similarly to that in Section 4.6.4, the proposition is proved.

Proof of Proposition 4.16

The proof adapts the proof of Proposition 2 in [61]. We outline the main steps for completeness. Observe that for the noise-free case, $y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|$. We obtain

$$\begin{aligned} \nabla \mathcal{R}_{tr}(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^m \left((\boldsymbol{a}_i^T \boldsymbol{z}) - |\boldsymbol{a}_i^T \boldsymbol{x}| \cdot \frac{\boldsymbol{a}_i^T \boldsymbol{z}}{|\boldsymbol{a}_i^T \boldsymbol{z}|} \right) \boldsymbol{a}_i \mathbf{1}_{\mathcal{T}^i} \\ &= \frac{1}{m}\sum_{i \notin \mathcal{B}} (\boldsymbol{a}_i^T \boldsymbol{h}) \boldsymbol{a}_i \mathbf{1}_{\mathcal{T}^i} + \frac{1}{m}\sum_{i \in \mathcal{B}} (\boldsymbol{a}_i^T \boldsymbol{z} + \boldsymbol{a}_i^T \boldsymbol{x}) \boldsymbol{a}_i \mathbf{1}_{\mathcal{T}^i}, \end{aligned} \tag{4.81}$$

where $\mathcal{B} := \{i : (\boldsymbol{a}_i^T \boldsymbol{x})(\boldsymbol{a}_i^T \boldsymbol{z}) < 0\}$. If $\|\boldsymbol{h}\|/\|\boldsymbol{x}\|$ is small enough, the cardinality of $\mathcal{B}$ is small and thus one expects the contribution of the second term in (4.81) to be negligible.

We note that events $\mathcal{T}^i$ are not statistically independent. To remove such dependency,

we introduce two new series of events

$$\mathcal{T}_1^i := \{\big||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\big| \le 0.5\alpha_h'\|\boldsymbol{h}\|\}, \tag{4.82}$$

$$\mathcal{T}_2^i := \{\big||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\big| \le 0.8\alpha_h'\|\boldsymbol{h}\|\}. \tag{4.83}$$

Due to Proposition 4.15, the following inclusion property

$$\mathcal{T}_1^i \subseteq \mathcal{T}^i \subseteq \mathcal{T}_2^i \tag{4.84}$$

holds for all i, where $\mathcal{T}^i$ is defined in Algorithm 4. It is easier to work with these new events because $\mathcal{T}_1^i$'s (resp. $\mathcal{T}_2^i$'s) are statistically independent for any fixed $\boldsymbol{x}$ and $\boldsymbol{z}$. Because of the inclusion property (4.84), we have

$$\langle \nabla\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle \ge \frac{1}{m}\sum_{i\notin\mathcal{B}}(\boldsymbol{a}_i^T\boldsymbol{h})^2\mathbf{1}_{\mathcal{T}_1^i} - \frac{1}{m}\sum_{i\in\mathcal{B}}|\boldsymbol{a}_i^T\boldsymbol{z} + \boldsymbol{a}_i^T\boldsymbol{x}|\cdot|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\mathcal{T}_2^i}. \tag{4.85}$$

Under the condition $i \notin \mathcal{B}$, we have $\mathcal{T}_1^i = \{\big|\boldsymbol{a}_i^T\boldsymbol{h}\big| \le 0.5\alpha_h'\|\boldsymbol{h}\|\}$. Under the condition $i \in \mathcal{B}$, we have $\mathcal{T}_2^i = \{\big|\boldsymbol{a}_i^T\boldsymbol{x} + \boldsymbol{a}_i^T\boldsymbol{z}\big| \le 0.8\alpha_h'\|\boldsymbol{h}\|\}$. For convenience, we introduce two parameters $\gamma_1 = 0.5\alpha_h'$ and $\gamma_2 = 0.8\alpha_h'$.

We next bound the two terms in (4.85) respectively. For the first term, because of the inclusion $\mathcal{B} \subseteq \{i : |\boldsymbol{a}_i^T\boldsymbol{x}| < |\boldsymbol{a}_i^T\boldsymbol{h}|\}$, we have

$$\begin{aligned}
\frac{1}{m}\sum_{i\notin\mathcal{B}}(\boldsymbol{a}_i^T\boldsymbol{h})^2\mathbf{1}_{\mathcal{T}_1^i} &= \frac{1}{m}\sum_{i\notin\mathcal{B}}(\boldsymbol{a}_i^T\boldsymbol{h})^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\le\gamma_1\|\boldsymbol{h}\|\}} \\
&\ge \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\le\gamma_1\|\boldsymbol{h}\|\}}\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|\ge|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \\
&\ge \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\le\gamma_1\|\boldsymbol{h}\|\}}\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|\ge\gamma_1\|\boldsymbol{h}\|\}}.
\end{aligned}$$

A simpler version of Lemma 4 in [61] gives that if $m > c_0 n$, with probability at least

$1 - c_1 \exp(-c_2 m\epsilon^2)$

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\le\gamma_1\|\boldsymbol{h}\|\}}\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|\ge\gamma_1\|\boldsymbol{h}\|\}} \ge (1-\zeta_1'-\zeta_2'-\epsilon)\|\boldsymbol{h}\|^2 \tag{4.86}$$

holds for all $\boldsymbol{h}\in\mathbb{R}^n$, where $\zeta_1' := 1 - \min\left\{\mathrm{E}\left[\xi^2\mathbf{1}_{\{\xi\ge\sqrt{1.01}\gamma_1\frac{\|\boldsymbol{h}\|}{\|\boldsymbol{x}\|}\}}\right], \mathrm{E}\left[\mathbf{1}_{\{\xi\ge\sqrt{1.01}\gamma_1\frac{\|\boldsymbol{h}\|}{\|\boldsymbol{x}\|}\}}\right]\right\}$ and $\zeta_2' := \mathrm{E}\left[\xi^2\mathbf{1}_{\{|\xi|>\sqrt{0.99}\gamma_1\}}\right]$ for $\xi\sim\mathcal{N}(0,1)$.

For the second term, we have

$$\frac{1}{m}\sum_{i\in\mathcal{B}}|\boldsymbol{a}_i^T\boldsymbol{z}+\boldsymbol{a}_i^T\boldsymbol{x}||\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\mathcal{T}_2^i} \le \gamma_2\|\boldsymbol{h}\|\frac{1}{m}\sum_{i\in\mathcal{B}}|\boldsymbol{a}_i^T\boldsymbol{h}| \le \gamma_2\|\boldsymbol{h}\|\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}}, \tag{4.87}$$

where the second inequality is due to the inclusion property $\mathcal{B}\subseteq\{i:|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}$.

Lemma 4.20. *For any $\epsilon>0$, if $m > c_0 n\epsilon^{-2}\log\epsilon^{-1}$, then with probability at least $1 - C\exp(-c_1\epsilon^2 m)$,*

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\cdot\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \le (0.12+\epsilon)\|\boldsymbol{h}\| \tag{4.88}$$

holds for all non-zero vectors $\boldsymbol{x},\boldsymbol{h}\in\mathbb{R}^n$ satisfying $\|\boldsymbol{h}\|\le\frac{1}{20}\|\boldsymbol{x}\|$. Here, $c_0, c_1, C>0$ are some universal constants.

Proof. See Section 4.6.4. □

Thus, putting together (4.86), (4.87) and Lemma 4.20 concludes the proof.

Proof of Proposition 4.17

This proof adapts the proof of Lemma 7 in [61]. Denote $v_i := \left(\boldsymbol{a}_i^T\boldsymbol{z} - |\boldsymbol{a}_i^T\boldsymbol{x}|\mathrm{sgn}(\boldsymbol{a}_i^T\boldsymbol{z})\right)\mathbf{1}_{\mathcal{T}^i}$. Then

$$\nabla\mathcal{R}_{tr}(\boldsymbol{z}) = \frac{1}{m}\boldsymbol{A}^T\boldsymbol{v},$$

where $\boldsymbol{A}$ is a matrix with each row being $\boldsymbol{a}_i^T$ and $\boldsymbol{v}$ is a $m-$dimensional vector with each entry being v_i. Thus, for sufficiently large m/n, we have

$$\|\nabla \mathcal{R}_{tr}(\boldsymbol{z})\| = \left\| \frac{1}{m} \boldsymbol{A}^T \boldsymbol{v} \right\| \leq \frac{1}{m} \|\boldsymbol{A}\| \cdot \|\boldsymbol{v}\| \leq (1+\delta) \frac{\|\boldsymbol{v}\|}{\sqrt{m}}$$

where the last inequality is due to the spectral norm bound $\|\boldsymbol{A}\| \leq \sqrt{m}(1+\delta)$ following from [142, Theorem 5.32].

We next bound $\|\boldsymbol{v}\|$. Let $\boldsymbol{v} = \boldsymbol{v}^{(1)} + \boldsymbol{v}^{(2)}$, where $v_i^{(1)} = \boldsymbol{a}_i^T \boldsymbol{h} \mathbf{1}_{\mathcal{T}^i \setminus \mathcal{B}^i}$ and $v_i^{(2)} = (\boldsymbol{a}_i^T \boldsymbol{x} + \boldsymbol{a}_i^T \boldsymbol{z}) \mathbf{1}_{\mathcal{T}^i \cap \mathcal{B}^i}$, where $B^i := \{(\boldsymbol{a}_i^T \boldsymbol{x})(\boldsymbol{a}_i^T \boldsymbol{z}) < 0\}$. By triangle inequality, we have $\|\boldsymbol{v}\| \leq \|\boldsymbol{v}^{(1)}\| + \|\boldsymbol{v}^{(2)}\|$. Furthermore, given $m > c_0 n$, by [45, Lemma 3.1] with probability $1 - \exp(-cm)$, we have

$$\frac{1}{m} \|\boldsymbol{v}^{(1)}\|^2 = \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{a}_i^T \boldsymbol{h})^2 \leq (1+\delta) \|\boldsymbol{h}\|^2.$$

By Lemma 4.19, we have with probability $1 - C\exp(-c_1 m)$

$$\frac{1}{m} \|\boldsymbol{v}^{(2)}\|^2 \leq (0.8 \alpha_h' \|\boldsymbol{h}\|)^2 \cdot \left(\frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{(\boldsymbol{a}_i^T \boldsymbol{x})(\boldsymbol{a}_i^T \boldsymbol{z}) < 0\}} \right) \leq 0.8 \|\boldsymbol{h}\|^2$$

holds, where the last inequality is due to Lemma 4.19. Hence,

$$\frac{\|\boldsymbol{v}\|}{\sqrt{m}} \leq \left(\sqrt{1+\delta} + \sqrt{0.8} \right) \|\boldsymbol{h}\|.$$

This concludes the proof.

Proof of Lemma 4.19

Denote correlation $\rho := \frac{\boldsymbol{z}^T \boldsymbol{x}}{\|\boldsymbol{z}\| \|\boldsymbol{x}\|}$. Under the condition $\|\boldsymbol{z} - \boldsymbol{x}\| \leq \frac{1}{11} \|\boldsymbol{x}\|$, simple calculation yields $0.995 < \rho \leq 1$. It suffices to show that the result holds with high probability for all $\boldsymbol{x}$ and $\boldsymbol{z}$ satisfying $\rho > 0.995$. Since now the claim is invariant with the norms of $\boldsymbol{x}$ and $\boldsymbol{z}$, we assume that both $\boldsymbol{x}$ and $\boldsymbol{z}$ have unit length without loss of generality.

We first establish the result for any fixed $\boldsymbol{x}$ and $\boldsymbol{z}$ and then develop a uniform bound by covering net argument in the end. We introduce a Lipschitz function to approximate the indicator function. Define

$$\chi(t) := \begin{cases} 1, & \text{if } t < 0; \\ -\frac{1}{\delta} \cdot t + 1, & \text{if } 0 \leq t \leq \delta; \\ 0, & \text{else;} \end{cases}$$

and then $\chi(t)$ is a Lipschitz function with Lipschitz constant $\frac{1}{\delta}$. In the following proof, we set $\delta = 0.001$. We further have

$$\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<0\}} \leq \chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) \leq \mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}. \tag{4.89}$$

For convenience, we denote $b_i := \boldsymbol{a}_i^T\boldsymbol{x}$ and $\tilde{b}_i := \boldsymbol{a}_i^T\boldsymbol{z}$. Then $(b_i, \tilde{b}_i)$ takes the jointly Gaussian distribution with mean $\mu = (0,0)^T$ and correlation ρ (b_i and $\tilde{b}_i$ have unit variance). We next estimate the expectation of $\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}$ as follows.

$$\mathrm{E}[\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}] = \mathbb{P}\left\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}) < \delta\right\} = \iint_{\tau_1\cdot\tau_2<\delta} f(\tau_1,\tau_2)d\tau_1 d\tau_2, \tag{4.90}$$

where $f(\tau_1, \tau_2)$ is the density of the jointly Gaussian random variables $(b_i, \tilde{b}_i)$. Note that $\mathrm{E}[\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}]$ is decreasing on ρ and for the case $\rho = 0.995$ we get $\mathrm{E}[\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}] = 0.045$ numerically. This implies that

$$\mathrm{E}[\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right)] \leq 0.045$$

for $\delta = 0.001$. Furthermore, $\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right)$ for all i are bounded and hence sub-

Gaussian. By Hoeffding type inequality for sub-Gaussian tail [142], we have

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m}\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) > (0.045+\epsilon)\right] < \exp(-cm\epsilon^2), \tag{4.91}$$

for some universal constant c, as long as $\rho \geq 0.995$.

We have proved so far that the claim holds for fixed $\boldsymbol{x}$ and $\boldsymbol{z}$. We next obtain a uniform bound over all $\boldsymbol{x}$ and $\boldsymbol{z}$ with unit length. Let $\mathcal{N}_\epsilon'$ be an ϵ-net covering the unit sphere in $\mathbb{R}^n$ and set

$$\mathcal{N}_\epsilon = \{(\boldsymbol{x}_0, \boldsymbol{z}_0) : (\boldsymbol{x}_0, \boldsymbol{z}_0) \in \mathcal{N}_\epsilon' \times \mathcal{N}_\epsilon'\}. \tag{4.92}$$

One has cardinality bound (i.e., the upper bound on the covering number) $|\mathcal{N}_\epsilon| \leq (1+2/\epsilon)^{2n}$. Then for any pair $(\boldsymbol{x}, \boldsymbol{z})$ with $\|\boldsymbol{x}\| = \|\boldsymbol{z}\| = 1$, there exists a pair $(\boldsymbol{x}_0, \boldsymbol{z}_0) \in \mathcal{N}_\epsilon$ such that $\|\boldsymbol{x}-\boldsymbol{x}_0\| \leq \epsilon$ and $\|\boldsymbol{z}-\boldsymbol{z}_0\| \leq \epsilon$. Taking the union bound for all the points on the net, we claim that

$$\frac{1}{m}\sum_{i=1}^{m}\chi\left((\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right) \leq 0.045+\epsilon, \quad \forall(\boldsymbol{x}_0, \boldsymbol{z}_0) \in \mathcal{N}_\epsilon \tag{4.93}$$

holds with probability at least $1-(1+2/\epsilon)^{2n}\exp(-cm\epsilon^2)$.

Since $\chi(t)$ is Lipschitz with constant $1/\delta$, we have

$$\left|\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) - \chi\left((\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right)\right| \leq \frac{1}{\delta}\left|(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right|. \tag{4.94}$$

Moreover, by [61, Lemma 1],

$$\frac{1}{m}\|\mathcal{A}(\boldsymbol{M})\|_1 \leq c_2\|\boldsymbol{M}\|_F, \qquad \text{for all symmetric rank-2 matrices } \boldsymbol{M} \in \mathbb{R}^{n\times n}, \tag{4.95}$$

holds with probability at least $1-C\exp(-c_1 m)$ as long as $m > c_0 n$ for some constants

$C, c_0, c_1, c_2 > 0$. Consequently, on the event that (4.95) holds, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) - \frac{1}{m}\sum_{i=1}^{m}\chi\left((\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right)\right| \\
&\le \frac{1}{m}\sum_{i=1}^{m}\left|\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) - \chi\left((\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right)\right| \\
&\le \frac{1}{\delta}\cdot\frac{1}{m}\|\mathcal{A}(\boldsymbol{x}\boldsymbol{z}^T - \boldsymbol{x}_0\boldsymbol{z}_0^T)\|_1 \qquad \text{due to (4.94)} \\
&\le \frac{1}{\delta}\cdot c_2\|\boldsymbol{x}\boldsymbol{z}^T - \boldsymbol{x}_0\boldsymbol{z}_0^T\|_F \qquad \text{due to (4.95)} \\
&\le \frac{1}{\delta}\cdot c_2(\|\boldsymbol{x}-\boldsymbol{x}_0\|\cdot\|\boldsymbol{z}\| + \|\boldsymbol{z}-\boldsymbol{z}_0\|\cdot\|\boldsymbol{x}_0\|) \le 2c_3\epsilon/\delta.
\end{aligned}$$

On the intersection of events that (4.93) and (4.95) hold, we have

$$\frac{1}{m}\sum_{i=1}^{m}\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) \le (0.045 + \epsilon + 2c_3\epsilon/\delta)\,, \tag{4.96}$$

for all $\boldsymbol{x}$ and $\boldsymbol{z}$ with unit length and $\rho \ge 0.995$. Since ϵ can be arbitrarily small, the proof is completed.

Proof of Lemma 4.20

We first observe that for any γ,

$$\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \le \mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma\|\boldsymbol{x}\|\}} + \mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\ge\gamma\|\boldsymbol{x}\|\}} \le \mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma\|\boldsymbol{x}\|\}} + \mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\ge 20\gamma\|\boldsymbol{h}\|\}} \tag{4.97}$$

where the last inequality is due to the assumption $\frac{\|\boldsymbol{h}\|}{\|\boldsymbol{x}\|} \le \frac{1}{20}$.

To establish the lemma, we set $\gamma = 0.15$ and denote $\gamma' := 20\gamma = 3$. We next respectively show that

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma\|\boldsymbol{x}\|\}} \le (0.11+\epsilon)\|\boldsymbol{h}\| \tag{4.98}$$

for all $\boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^n$, and

$$\frac{1}{m}\sum_{i=1}^{m} |\boldsymbol{a}_i^T \boldsymbol{h}| \mathbf{1}_{\{|\boldsymbol{a}_i^T \boldsymbol{h}| > \gamma' \|\boldsymbol{h}\|\}} \leq (0.01 + \epsilon)\|\boldsymbol{h}\| \tag{4.99}$$

for all $\boldsymbol{h} \in \mathbb{R}^n$.

We first prove (4.98). Without loss of generality, we assume that $\boldsymbol{h}$ and $\boldsymbol{x}$ have unit length. We introduce a Lipschitz function to approximate the indicator functions, which is defined as

$$\chi_x(t) := \begin{cases} 1, & \text{if } |t| < \gamma; \\ \frac{1}{\delta}(\gamma - |t|) + 1, & \text{if } \gamma \leq |t| \leq \gamma + \delta; \\ 0, & \text{else.} \end{cases}$$

Then $\chi_x(t)$ is a Lipschitz function with constant $\frac{1}{\delta}$. We further have

$$\mathbf{1}_{\{|\boldsymbol{a}_i^T \boldsymbol{x}| < \gamma\}} \leq \chi_x(\boldsymbol{a}_i^T \boldsymbol{x}) \leq \mathbf{1}_{\{|\boldsymbol{a}_i^T \boldsymbol{x}| < \gamma + \delta\}}. \tag{4.100}$$

We first prove bounds for any fixed pair $\boldsymbol{h}, \boldsymbol{x}$, and then develop a uniform bound later on.

We next estimate the expectation of $|\boldsymbol{a}_i^T \boldsymbol{h}| \mathbf{1}_{\{|\boldsymbol{a}_i^T \boldsymbol{x}| < \gamma + \delta\}}$,

$$\mathrm{E}[|\boldsymbol{a}_i^T \boldsymbol{h}| \mathbf{1}_{\{|\boldsymbol{a}_i^T \boldsymbol{x}| < \gamma + \delta\}}] = \iint_{-\infty}^{\infty} |\tau_1| \mathbf{1}_{\{|\tau_2| < \gamma + \delta\}} \cdot f(\tau_1, \tau_2) d\tau_1 d\tau_2, \tag{4.101}$$

where $f(\tau_1, \tau_2)$ is the density of two jointly Gaussian random variables with correlation

$\rho = \frac{\boldsymbol{h}^T\boldsymbol{x}}{\|\boldsymbol{h}\|\|\boldsymbol{x}\|} \neq \pm 1$. We then continue to derive

$$\begin{aligned}
&\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}] \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} |\tau_1| \exp\left(-\frac{\tau_1^2}{2}\right) \cdot \int_{-(\gamma+\delta)}^{\gamma+\delta} \exp\left(-\frac{(\tau_2-\rho\tau_1)^2}{2(1-\rho^2)}\right) d\tau_2 d\tau_1 \qquad (4.102) \\
&= \frac{1}{\sqrt{2}\pi} \int_{-\infty}^{\infty} |\tau_1| \exp\left(-\frac{\tau_1^2}{2}\right) \cdot \int_{\frac{-\gamma-\delta-\rho\tau_1}{\sqrt{2(1-\rho^2)}}}^{\frac{\gamma+\delta-\rho\tau_1}{\sqrt{2(1-\rho^2)}}} \exp\left(-\tau^2\right) d\tau d\tau_1 \qquad \text{by changing variables} \\
&= \frac{1}{\sqrt{8}\pi} \int_{-\infty}^{\infty} |\tau_1| \exp\left(-\frac{\tau_1^2}{2}\right) \cdot \left(\mathrm{erf}\left(\frac{\gamma+\delta-\rho\tau_1}{\sqrt{2(1-\rho^2)}}\right) - \mathrm{erf}\left(\frac{-\gamma-\delta-\rho\tau_1}{\sqrt{2(1-\rho^2)}}\right)\right) d\tau_1 \qquad (4.103)
\end{aligned}$$

For $|\rho| < 1$, $\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]$ is a continuous function of ρ. The last integral (4.103) can be calculated numerically. Figure 4.6 plots $\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]$ for $\gamma = 0.15$ and $\delta = 0.01$ over $\rho \in (-1, 1)$. Furthermore, (4.102) indicates that $\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]$ is monotonically increasing with both θ and δ. Thus, we obtain a universal bound

$$\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}] \leq 0.11\|\boldsymbol{h}\| \quad \text{for } \gamma < 0.15 \text{ and } \delta = 0.01, \qquad (4.104)$$

which further implies $\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x(\boldsymbol{a}_i^T\boldsymbol{x})] \leq 0.11\|\boldsymbol{h}\|$ for $\gamma < 0.15$ and $\delta = 0.01$.

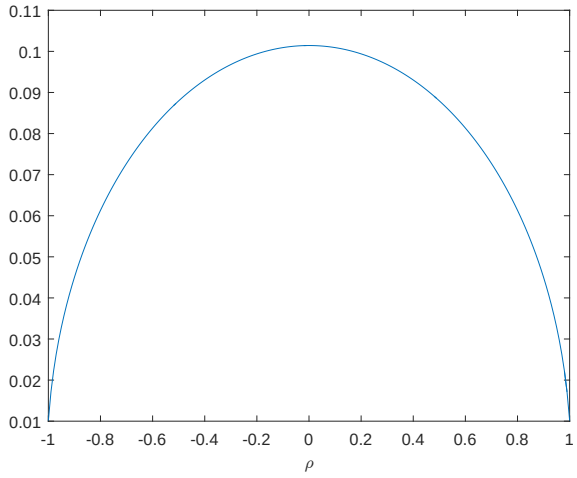


Fig. 4.6: $\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]$ with respect to ρ

Furthermore, $|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x(\boldsymbol{a}_i^T\boldsymbol{x})$'s are sub-Gaussian with sub-Gaussian norm $\mathcal{O}(\|\boldsymbol{h}\|)$. By the Hoeffding type of sub-Gaussian tail bound [142], we have

$$\mathcal{P}\left[\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x(\boldsymbol{a}_i^T\boldsymbol{x}) > (0.11+\epsilon)\|\boldsymbol{h}\|\right] < \exp(-cm\epsilon^2), \tag{4.105}$$

for some universal constant c.

We have proved so far that the claim holds for a fixed pair $\boldsymbol{h}, \boldsymbol{x}$. We next obtain a uniform bound over all $\boldsymbol{x}$ and $\boldsymbol{h}$ with unit length. Let $\mathcal{N}_\epsilon'$ be a ϵ-net covering the unit sphere in $\mathbb{R}^n$ and set

$$\mathcal{N}_\epsilon = \{(\boldsymbol{x}_0, \boldsymbol{h}_0) : (\boldsymbol{x}_0, \boldsymbol{h}_0) \in \mathcal{N}_\epsilon' \times \mathcal{N}_\epsilon'\}.$$

One has cardinality bound (i.e., the upper bound on the covering number) $|\mathcal{N}_\epsilon| \leq (1 + 2/\epsilon)^{2n}$. Then for any pair $(\boldsymbol{x}, \boldsymbol{h})$ with $\|\boldsymbol{x}\| = \|\boldsymbol{h}\| = 1$, there exists a pair $(\boldsymbol{x}_0, \boldsymbol{h}_0) \in \mathcal{N}_\epsilon$ such that $\|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \epsilon$ and $\|\boldsymbol{h} - \boldsymbol{h}_0\| \leq \epsilon$. Taking the union bound for all the points on the net, one can show

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}_0|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right) \leq 0.11+\epsilon, \quad \forall(\boldsymbol{x}_0, \boldsymbol{h}_0) \in \mathcal{N}_\epsilon \tag{4.106}$$

holds with probability at least $1 - (1 + 2/\epsilon)^{2n}\exp(-cm\epsilon^2)$.

Since $\chi_x(t)$ is Lipschitz with constant $1/\delta$, we have the following bound

$$\left|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}\right) - \chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)\right| \leq \frac{1}{\delta}\left|\boldsymbol{a}_i^T(\boldsymbol{x} - \boldsymbol{x}_0)\right|. \tag{4.107}$$

Consequently, on the event that (4.95) holds, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)-\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}_0|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)\right| \\
&\le \frac{1}{m}\sum_{i=1}^{m}\left||\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)-|\boldsymbol{a}_i^T\boldsymbol{h}_0|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)\right| \\
&\le \frac{1}{m}\sum_{i=1}^{m}\left|\boldsymbol{a}_i^T(\boldsymbol{h}-\boldsymbol{h}_0)\right|+\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\delta}\left|\boldsymbol{a}_i^T\boldsymbol{h}_0\right|\cdot\left|\boldsymbol{a}_i^T\boldsymbol{x}-\boldsymbol{a}_i^T\boldsymbol{x}_0\right| \qquad \text{due to (4.107)} \\
&\le c_2'\|\boldsymbol{h}-\boldsymbol{h}_0\|+\frac{1}{\delta}\cdot c_2\|\boldsymbol{h}_0(\boldsymbol{x}-\boldsymbol{x}_0)^T\|_F \qquad \text{due to (4.95)} \\
&\le c_3\epsilon/\delta.
\end{aligned}$$

On the intersection of events that (4.106) and (4.95) hold, we have

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)\le\left(0.11+\epsilon+2c_3\epsilon/\delta\right), \tag{4.108}$$

for all $\boldsymbol{x}$ and $\boldsymbol{h}$ with unit length.

We next prove (4.99). Without loss of generality, we assume that $\boldsymbol{h}$ has unit length. We introduce a Lipschitz function to approximate the indicator functions, which is defined as

$$\chi_h(t):=\begin{cases}|t|, & \text{if } |t|>\gamma';\\ \frac{1}{\delta}(|t|-\gamma')+\gamma', & \text{if } \gamma'(1-\delta)\le|t|\le\gamma';\\ 0, & \text{else.}\end{cases}$$

Then, $\chi_h(t)$ is a Lipschitz function with constant $\frac{1}{\delta}$. We further have

$$|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'\|\boldsymbol{h}\|\}}\le\chi_h(\boldsymbol{a}_i^T\boldsymbol{h})\le|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'(1-\delta)\|\boldsymbol{h}\|\}}. \tag{4.109}$$

We first prove bounds for any fixed $\boldsymbol{h}$, and then develop a uniform bound later on.

We next estimate the expectation of $|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'(1-\delta)\|\boldsymbol{h}\|\}}$ as follows:

$$\begin{aligned}
\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'(1-\delta)\|\boldsymbol{h}\|\}}] &= \int_{-\infty}^{\infty} |\tau|\mathbf{1}_{\{|\tau|>\gamma'(1-\delta)\}} \cdot f(\tau)d\tau, \\
&= 2 \cdot \frac{1}{\sqrt{2\pi}} \int_{\gamma'(1-\delta)}^{\infty} \tau \exp\left(-\frac{\tau^2}{2}\right) d\tau \\
&= \sqrt{\frac{2}{\pi}} \exp(-\gamma'^2(1-\delta)^2/2) < 0.01 \qquad \text{for } \gamma' = 3, \delta = 0.01,
\end{aligned} \tag{4.110}$$

where $f(\tau)$ is the density of the standard Gaussian. We note that $\mathrm{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'(1-\delta)\|\boldsymbol{h}\|\}}]$ is monotonically increasing with δ and decreasing with γ'. Furthermore, $\mathrm{E}[\chi_h(\boldsymbol{a}_i^T\boldsymbol{h})] \leq 0.01\|\boldsymbol{h}\|$ for $\gamma' \geq 3$ and $\delta \leq 0.01$.

Moreover, $\chi_h(\boldsymbol{a}_i^T\boldsymbol{h})$ for all i are sub-Gaussian with sub-Gaussian norm $\mathcal{O}(\|\boldsymbol{h}\|)$. By the Hoeffding type sub-Gaussian tail bound [142], we have

$$\mathcal{P}\left[\frac{1}{m}\sum_{i=1}^{m} \chi_h(\boldsymbol{a}_i^T\boldsymbol{h}) > (0.01+\epsilon)\|\boldsymbol{h}\|\right] < \exp(-cm\epsilon^2), \tag{4.111}$$

for some universal constant c.

We have proved so far that the claim holds for a fixed $\boldsymbol{h}$. We next obtain a uniform bound over all $\boldsymbol{h}$ with unit length. Let $\mathcal{N}_\epsilon$ be an ϵ-net covering the unit sphere in $\mathbb{R}^n$. One has cardinality bound (i.e., the upper bound on the covering number) $|\mathcal{N}_\epsilon| \leq (1+2/\epsilon)^n$. Then for any $\boldsymbol{h}$ with unit length, there exists a $\boldsymbol{h}_0 \in \mathcal{N}_\epsilon$ such that $\|\boldsymbol{h}-\boldsymbol{h}_0\| \leq \epsilon$. Taking the union bound for all the points on the net, one can show

$$\frac{1}{m}\sum_{i=1}^{m} \chi_h(\boldsymbol{a}_i^T\boldsymbol{h}_0) \leq 0.01+\epsilon, \quad \forall \boldsymbol{h}_0 \in \mathcal{N}_\epsilon \tag{4.112}$$

holds with probability at least $1-(1+2/\epsilon)^n \exp(-cm\epsilon^2)$.

Consequently, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}\chi_h(\boldsymbol{a}_i^T\boldsymbol{h}) - \frac{1}{m}\sum_{i=1}^{m}\chi_h(\boldsymbol{a}_i^T\boldsymbol{h}_0)\right| \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\left|\chi_h(\boldsymbol{a}_i^T\boldsymbol{h}) - \chi_h(\boldsymbol{a}_i^T\boldsymbol{h}_0)\right| \\
&\leq \frac{1}{\delta}\cdot\frac{1}{m}\sum_{i=1}^{m}\left|\boldsymbol{a}_i^T(\boldsymbol{h}-\boldsymbol{h}_0)\right| \\
&\leq \frac{1}{\delta}c_2'\|\boldsymbol{h}-\boldsymbol{h}_0\| \leq c_3\epsilon/\delta,
\end{aligned}$$

where the second inequality is because $\chi_h(t)$ is Lipschitz continuous with constant $1/\delta$.

On the intersection of events that (4.112) and (4.95) hold, we have

$$\frac{1}{m}\sum_{i=1}^{m}\chi_h(\boldsymbol{a}_i^T\boldsymbol{h}) \leq \left(0.01+\epsilon+c_3\epsilon/\delta\right), \tag{4.113}$$

for all $\boldsymbol{h}$ with unit length.

Putting together (4.108) and (4.113), and since ϵ can be arbitrarily small, the proof is completed.

CHAPTER 5
FUTURE WORK

We study the robust PCA problem and the phase retrieval problem in this dissertation. For the robust PCA problem we provide a more refined analysis of PCP via local coherence. For the phase retrieval problem, we first propose a fast algorithm RWF that solves phase retrieval by minimizing a *nonconvex and nonsmooth* quadratic loss function. This loss function sacrifices the smoothness but enjoys advantages in statistical and computational efficiency. We further propose a median truncation approach to modify existing TWF and RWF algorithms to resist the outliers that are often encountered in the phase retrieval measurements. Our median truncation modified algorithms do not require prior knowledge of outliers and perform well in an oblivious manner.

These works motivate us to further explore several directions in the future. We describe two interesting problems that are under consideration.

5.1 Structural Phase Retrieval

In practice, signals often have specific structures e.g., non-negativity, sparsity, etc. Motivated by the idea of compressive sensing, it is anticipated that with the prior structural knowledge, the number of measurements that is required to guarantee exact recovery can

be substantially reduced.

Specifically, [79] proposed the thresholded Wirtinger flow algorithm to recover a sparse underlying signal, which showed that the true signal $\boldsymbol{x}$ can be recovered with high probability and a linear convergence rate, as long as the number of samples is $\mathcal{O}(k^2 \log n)$, where k is cardinality of the support set of $\boldsymbol{x}$. One interesting future direction is to modify the RWF algorithm to exploit such structures to assist the recovery. The lower-order loss function may offer simplicity and improvement to the proof of the performance guarantee in such cases.

The heuristic algorithm can be designed as in Algorithm 5. We introduce an amending function $H(\cdot)$ which can admit the prior knowledge of the signal, e.g., sparsity level and non-negativity.

Algorithm 5 Reshaped Wirtinger Flow with Prior

Input: $\boldsymbol{y} = \{y_i\}_{i=1}^m$, $\{\boldsymbol{a}_i\}_{i=1}^m$, amending function H, gradient stepsize μ;

Initialization: Same as in RWF (see Algorithm 1).
Gradient loop: for $t = 0 : T-1$ do

$$\begin{aligned} &\text{Compute the amending parameter } \tau^{(t)}, \\ &\text{Update} \qquad \boldsymbol{z}^{(t+1)} = H_{\tau^{(t)}}\left(\boldsymbol{z}^{(t)} - \mu \cdot \nabla \ell(\boldsymbol{z}^{(t)})\right). \end{aligned} \tag{5.1}$$

Output $\boldsymbol{z}^{(T)}$.

We anticipate such a modified algorithm can improve the sample complexity achieved in [79] and expedite the convergence due to the lower-order loss function.

5.2 Fast and Robust Low-rank Matrix Recovery

Motivated by the success of nonconvex heuristics in phase retrieval problem, researchers have extensively studied the nonconvex approaches to recover the low-rank matrix.

Problem 5.1 (**Low-rank Matrix Recovery**). *Find $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with minimal rank such that*

$$\mathcal{A}(\boldsymbol{X}) = \boldsymbol{y} \tag{5.2}$$

where $\mathcal{A} : \mathbb{R}^{n\times p} \to \mathbb{R}^m$ is a affine transformation (known) and $\boldsymbol{y} \in \mathbb{R}^m$ is the observation vector (known).

Suppose that the underlying solution $\boldsymbol{X}$ is a rank-r positive semidefinite matrix, and can be decomposed as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$, where $\boldsymbol{U} \in \mathbb{R}^{n\times r}$. Recently, [85, 86] proposed to minimize the following loss function

$$\ell(\boldsymbol{Z}) := \frac{1}{4m}\|\mathcal{A}(\boldsymbol{Z}\boldsymbol{Z}^T) - \boldsymbol{y}\|^2, \tag{5.3}$$

where $\boldsymbol{Z} \in \mathbb{R}^{n\times r}$. The recovery of the true matrix is guaranteed by minimizing the non-convex objective (5.3), when $\mathcal{A}$ is composed of independent Gaussian matrix measurements (i.e., elements are independent Gaussian random variables). Specifically, [86] proposed an algorithm that achieves the linear convergence rate and the sample complexity $\mathcal{O}(nr^3 \log n)$.

However, it consumes huge space to store the Gaussian matrices especially when the dimension becomes large. Thus rank-1 measurement matrices are preferred in this scenario [27]. For the positive semidefinite (PSD) matrix $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^T$, the measurements become

$$y_i = \langle \boldsymbol{a}_i\boldsymbol{a}_i^T, \boldsymbol{X}\rangle = \boldsymbol{a}_i^T\boldsymbol{X}\boldsymbol{a}_i = \|\boldsymbol{U}^T\boldsymbol{a}_i\|^2, \quad i = 1, \ldots, m, \tag{5.4}$$

where $\boldsymbol{a}_i \in \mathbb{R}^n$ are the measurement vectors and m is the number of measurements.

It is interesting to extend the idea of RWF to improve the low-rank matrix recovery. Instead of minimizing (5.2), we propose to minimize a lower-order loss function

$$\mathcal{R}(\boldsymbol{Z}) := \frac{1}{2m}\sum_{i=1}^{m}(\|\boldsymbol{Z}^T\boldsymbol{a}_i\| - \sqrt{y_i})^2. \tag{5.5}$$

We anticipate that the above loss function will bring us benefit on the convergence rate and the sample complexity.

It is also interesting to consider the robust low-rank matrix recovery [146]. We expect that the median-truncation approach will provide new improvement along this line.

REFERENCES

[1] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[6] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[7] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 597–610, 2013.

[8] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National science review*, vol. 1, no. 2, pp. 293–314, 2014.

[9] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 57–60.

[10] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Low-rank sparse learning for robust visual tracking," in *European conference on computer vision*. Springer, 2012, pp. 470–484.

[11] W. Dong, G. Shi, and X. Li, "Nonlocal image restoration with bilateral variance estimation: a low-rank approach," *IEEE transactions on image processing*, vol. 22, no. 2, pp. 700–711, 2013.

[12] K. Zheng, K. Li, and S. Cong, "A reconstruction algorithm for compressive quantum tomography using various measurement sets," *Scientific Reports*, vol. 6, p. 38497, 2016.

[13] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[14] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM Journal on Optimization*, vol. 21, no. 2, pp. 572–596, 2011.

[15] H. Zhang, Y. Zhou, and Y. Liang, "Analysis of robust pca via local incoherence," in *Advances in Neural Information Processing Systems*, 2015, pp. 1819–1827.

[16] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[17] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980 –2998, June 2010.

[18] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[19] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, March 2011.

[20] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[21] Y. Chen, "Incoherence-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2909–2923, May 2015.

[22] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward, "Completing any low-rank matrix, provably." *Journal of Machine Learning Research*, vol. 16, pp. 2999–3034, 2015.

[23] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," in *Advances in Neural Information Processing Systems*, Vancouver, B.C., Canada, December 2009, pp. 952–960.

[24] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[25] V. Koltchinskii, K. Lounici, A. B. Tsybakov *et al.*, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, vol. 39, no. 5, pp. 2302–2329, 2011.

[26] J. Wright, A. Ganesh, K. Min, and Y. Ma, "Compressive principal component pursuit," *Information and Inference*, vol. 2, no. 1, pp. 32–68, 2013.

[27] T. T. Cai and A. Zhang, "Rop: Matrix recovery via rank-one projections," *The Annals of Statistics*, vol. 43, no. 1, pp. 102–138, 2015.

[28] D. Hsu, S. M. Kakade, and T. Zhang, "Robust matrix decomposition with sparse corruptions," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7221–7234, 2011.

[29] X. Li, "Compressed sensing and matrix completion with constant proportion of corruptions," *Constructive Approximation*, vol. 37, no. 1, pp. 73–99, 2013.

[30] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, "Low-rank matrix recovery from errors and erasures," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4324–4337, 2013.

[31] A. Ganesh, J. Wright, X. Li, E. J. Candes, and Y. Ma, "Dense error correction for low-rank matrices via principal component pursuit," in *IEEE International Symposium on Information Theory (ISIT)*, Austin, TX, US, June 2010, pp. 1513–1517.

[32] W. Ha and R. F. Barber, "Robust pca with compressed data," in *Advances in Neural Information Processing Systems*, 2015, pp. 1936–1944.

[33] M. B. McCoy and J. A. Tropp, "The achievable performance of convex demixing," *arXiv preprint arXiv:1309.7478*, 2013.

[34] ——, "Sharp recovery bounds for convex demixing, with applications," *Foundations of Computational Mathematics*, vol. 14, no. 3, pp. 503–567, 2014.

[35] M. B. McCoy, "A geometric analysis of convex demixing," Ph.D. dissertation, California Institute of Technology, 2013.

[36] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[37] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust PCA," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[38] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust pca via gradient descent," in *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[39] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[40] M. Hardt, "Understanding alternating minimization for matrix completion," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 651–660.

[41] Q. Zheng and J. Lafferty, "Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent," *arXiv preprint arXiv:1605.07051*, 2016.

[42] C. Jin, S. M. Kakade, and P. Netrapalli, "Provable efficient online matrix completion via non-convex stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2016, pp. 4520–4528.

[43] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[44] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, pp. 199–225, 2013.

[45] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[46] E. J. Candes and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, April 2009.

[47] E. Candes and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053 –2080, May 2010.

[48] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, "Stable low-rank matrix recovery via null space properties," *arXiv preprint arXiv:1507.07184*, 2015.

[49] J. Drenth, *X-Ray Crystallography*. Wiley Online Library, 2007.

[50] J. Miao, P. Charalambous, J. Kirz, and D. Sayre, "Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens," *Nature*, vol. 400, no. 6742, pp. 342–344, 1999.

[51] J. Miao, T. 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, 2008.

[52] R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, p. 237, 1972.

[53] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied Optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[54] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345–356, 2006.

[55] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Mathematical Programming*, vol. 149, no. 1-2, pp. 47–81, 2015.

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

[57] L. Demanet and P. Hand, "Stable optimizationless recovery from phaseless linear measurements," *Journal of Fourier Analysis and Applications*, vol. 20, no. 1, pp. 199–221, 2014.

[58] E. J. Candès and X. Li, "Solving quadratic equations via phaselift when there are about as many equations as unknowns," *Foundations of Computational Mathematics*, vol. 14, no. 5, pp. 1017–1026, 2014.

[59] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM Journal on Mathematical Analysis*, 2013.

[60] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: a contemporary overview," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 87–109, 2015.

[61] Y. Chen and E. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[62] K. Wei, "Solving systems of phaseless equations via Kaczmarz methods: a proof of concept study," *Inverse Problems*, vol. 31, no. 12, p. 125008, 2015.

[63] G. Li, Y. Gu, and Y. M. Lu, "Phase retrieval using iterative projections: Dynamics in the large systems limit," in *The 53rd Annual Allerton Conference on Communication, Control, and Computing*, 2015.

[64] R. Kolte and A. Özgür, "Phase retrieval via incremental truncated Wirtinger flow," *arXiv preprint arXiv:1606.03196*, 2016.

[65] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "Reshaped Wirtinger flow and incremental algorithms for solving quadratic systems of equations," *arXiv preprint arXiv:1605.07719v2*, 2016.

[66] H. Zhang, Y. Chi, and Y. Liang, "Median-truncated nonconvex approach for phase retrieval with outliers," *arXiv preprint arXiv:1603.03805v2*, 2016.

[67] ——, "Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow," in *International conference on machine learning (ICML)*, 2016, pp. 1022–1031.

[68] H. Zhang and Y. Liang, "Reshaped wirtinger flow for solving quadratic system of equations," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 2622–2630.

[69] S. Sanghavi, R. Ward, and C. D. White, "The local convexity of solving systems of quadratic equations," *Results in Mathematics*, pp. 1–40, 2016.

[70] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *arXiv preprint arXiv:1605.08285*, 2016.

[71] D. Weller, A. Pnueli, G. Divon, O. Radzyner, Y. Eldar, and J. Fessler, "Undersampled phase retrieval with outliers," *IEEE Transactions on Computational Imaging*, vol. 1, no. 4, pp. 247–258, Dec 2015.

[72] P. Hand, "Phaselift is robust to a constant fraction of arbitrary errors," *Applied and Computational Harmonic Analysis*, vol. 42, no. 3, pp. 550–562, 2017.

[73] P. J. Huber, *Robust statistics*. Springer, 2011.

[74] K. Chen, "On k-median clustering in high dimensions," in *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. Society for Industrial and Applied Mathematics, 2006, pp. 1177–1185.

[75] D. Wagner, "Resilient aggregation in sensor networks," in *Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks*. ACM, 2004, pp. 78–87.

[76] Y. Chen, C. Caramanis, and S. Mannor, "Robust sparse regression under adversarial corruption," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[77] C. Qu and H. Xu, "Subspace clustering with irrelevant features via robust dantzig selector," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[78] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[79] T. T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow," *arXiv preprint arXiv:1506.03382*, 2015.

[80] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *arXiv preprint arXiv:1602.06664*, 2016.

[81] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, 2013.

[82] C. De Sa, K. Olukotun, and C. Ré, "Global convergence of stochastic gradient descent for some non-convex matrix problems," *arXiv preprint arXiv:1411.1134v3*, 2015.

[83] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[84] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *arXiv preprint arXiv:1509.03025*, 2015.

[85] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *International Conference on Machine Learning (ICML)*, 2016.

[86] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[87] D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi, "Provable non-convex projected gradient descent for a class of constrained matrix optimization problems," *arXiv preprint arXiv:1606.01316*, 2016.

[88] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, "Guarantees of riemannian optimization for low rank matrix recovery," *arXiv preprint arXiv:1511.01562*, 2015.

[89] A. Anandkumar, P. Jain, Y. Shi, and U. Niranjan, "Tensor vs matrix methods: Robust tensor decomposition under block sparse perturbations," *arXiv preprint arXiv:1510.04747*, 2015.

[90] S. Arora, R. Ge, T. Ma, and A. Moitra, "Simple, efficient, and neural algorithms for sparse coding." in *COLT*, 2015, pp. 113–149.

[91] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery using nonconvex optimization," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[92] A. S. Bandeira, N. Boumal, and V. Voroninski, "On the low-rank approach for semidefinite programs arising in synchronization and community detection," in *29th Annual Conference on Learning Theory*, 2016.

[93] N. Boumal, "Nonconvex phase synchronization," *SIAM Journal on Optimization*, vol. 26, no. 4, pp. 2355–2377, 2016.

[94] K. Lee, Y. Li, M. Junge, and Y. Bresler, "Blind recovery of sparse signals from subsampled convolution," *arXiv preprint arXiv:1511.06149*, 2015.

[95] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *arXiv preprint arXiv:1606.04933*, 2016.

[96] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems (NIPS)*, 2012.

[97] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[98] Y. Li, Y. Sun, and Y. Chi, "Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 397–408, Jan 2017.

[99] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography-Part I: Secret sharing," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1121–1132, Jul. 1993.

[100] U. M. Maurer, "Secrete key agreement by public discussion based on common information," *IEEE Trans. Inform. Theory*, vol. 39, no. 5, pp. 733–742, May 1993.

[101] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, 1973.

[102] I. Csiszár and P. Narayan, "Secrecy capacities for mulitple terminals," *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 3047–3061, Dec. 2004.

[103] U. M. Maurer and S. Wolf, "Unconditionally secure key agreement and the intrinsic conditional information," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 499–514, Feb. 1999.

[104] I. Csiszár and P. Narayan, "Common randomness and secret key generation with a helper," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 344–366, Mar. 2000.

[105] U. M. Maurer and S. Wolf, "Secret-key agreement over unauthenticated public channels-Part I. definitions and a completeness result," *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 822–831, Apr. 2003.

[106] ——, "Secret-key agreement over unauthenticated public channels-Part II. the simulatability condition," *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 832–838, Apr. 2003.

[107] ——, "Secret-key agreement over unauthenticated public channels-Part III. privacy amplification," *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 839–851, Apr. 2003.

[108] C. Ye and P. Narayan, "The secret key-private key capacity region for three terminals," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Adelaide, Australia, September 2005.

[109] H. Zhang, L. Lai, Y. Liang, and H. Wang, "The secret key-private key generation over three terminals: Capacity region," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Honolulu, HI, USA, June 2014.

[110] ——, "The capacity region of the source-type model for secret key and private key generation," *IEEE Trans. Inform. Theory*, vol. 60, no. 10, pp. 6389–6398, Oct. 2014.

[111] C. Ye and P. Narayan, "The private key capacity region for three terminals," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Chicago, Illinois, USA, June 2004.

[112] L. Lai and L. Huie, "Simultaneously generating multiple keys in many to one networks," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Istanbul, Turkey, July 2013.

[113] H. Zhang, Y. Liang, and L. Lai, "Key capacity region for a cellular source model," in *Proc. IEEE Information Theory Workshop (ITW)*, Hobart, Tasmania, Australia, November 2014.

[114] H. Zhang, Y. Liang, L. Lai, and S. Shamai (Shitz), "Two-key generation for a cellular model with a helper," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Hong Kong, China, June 2015.

[115] H. Zhang, Y. Liang, L. Lai, and S. S. Shitz, "Multi-key generation over a cellular model with a helper," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3804–3822, 2017.

[116] M. Mukherjee, N. Kashyap, and Y. Sankarasubramaniam, "Achieving sk capacity in the source model: When must all terminals talk?" in *2014 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2014, pp. 1156–1160.

[117] C. Chan, A. Al-Bashabsheh, Q. Zhou, N. Ding, T. Liu, and A. Sprintson, "Successive omniscience," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3270–3289, 2016.

[118] H. Zhang, Y. Liang, and L. Lai, "Secret key capacity: Talk or keep silent?" in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 291–295.

[119] S. Oymak and B. Hassibi, "Finding dense clusters via "low rank+ sparse" decomposition," *arXiv preprint arXiv:1104.5186*, 2011.

[120] Y. Chen, S. Sanghavi, and H. Xu, "Clustering sparse graphs," in *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, US, December 2012, pp. 2204–2212.

[121] ——, "Improved graph clustering," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6440–6455, Oct 2014.

[122] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu, "Clustering partially observed graphs via convex optimization," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2213–2238, 2014.

[123] N. Alon, M. Krivelevich, and B. Sudakov, "Finding a large hidden clique in a random graph," *Random Structures and Algorithms*, vol. 13, no. 3-4, pp. 457–466, 1998.

[124] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *Conference on Learning Theory*, Atlanta, GA, USA, June 2013, pp. 1046–1066.

[125] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie, "Testing k-wise and almost k-wise independence," in *Proceedings of ACM symposium on Theory of computing (STOC)*, San Diego, CA, USA, June 2007, pp. 496–505.

[126] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[127] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[128] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[129] T. T. Cai, J. Fan, and T. Jiang, "Distributions of angles in random packing on spheres." *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1837–1864, 2013.

[130] A. Y. Kruger, "On Fréchet subdifferentials," *Journal of Mathematical Sciences*, vol. 116, no. 3, pp. 3325–3358, 2003.

[131] Y. Chi and Y. M. Lu, "Kaczmarz method for solving quadratic equations," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1183–1187, 2016.

[132] E. Moulines and F. R. Bach, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[133] D. Needell, N. Srebro, and R. Ward, "Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm," *Mathematical Programming*, vol. 155, no. 1-2, pp. 549–573, 2016.

[134] S. Kaczmarz, “Angenäherte auflösung von systemen linearer gleichungen,” *Bulletin International de lŠAcademie Polonaise des Sciences et des Lettres*, vol. 35, pp. 355–357, 1937.

[135] T. Strohmer and R. Vershynin, “A randomized kaczmarz algorithm with exponential convergence,” *Journal of Fourier Analysis and Applications*, vol. 15, no. 2, pp. 262–278, 2009.

[136] A. Zouzias and N. M. Freris, “Randomized extended Kaczmarz for solving least squares,” *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 2, pp. 773–793, 2013.

[137] A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson, “Saving phase: Injectivity and stability for phase retrieval,” *Applied and Computational Harmonic Analysis*, vol. 37, no. 1, pp. 106–125, 2014.

[138] F. Fogel, I. Waldspurger, and A. dŠAspremont, “Phase retrieval for imaging problems,” *Mathematical programming computation*, vol. 8, no. 3, pp. 311–335, 2016.

[139] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, “Gradient descent converges to minimizers,” *arXiv preprint arXiv:1602.04915*, 2016.

[140] R. Ge, F. Huang, C. Jin, and Y. Yuan, “Escaping from saddle points-online stochastic gradient for tensor decomposition.” in *Conference on Learning Theory (COLT)*, 2015, pp. 797–842.

[141] Z. Xu, “The minimal measurement number for low-rank matrices recovery,” *arXiv preprint arXiv:1505.07204*, 2015.

[142] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *Compressed Sensing, Theory and Applications*, pp. 210 – 268, 2012.

[143] J. D. Donahue, "Products and quotients of random variables and their applications," DTIC Document, Tech. Rep., 1964.

[144] R. J. Tibshirani, "Fast computation of the median by successive binning," *arXiv preprint arXiv:0806.3301*, 2008.

[145] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," in *Automata, languages and programming*. Springer, 2002, pp. 693–703.

[146] Y. Li, Y. Sun, and Y. Chi, "Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements," *IEEE Transactions on Signal Processing*, 2016.

VITA

NAME OF AUTHOR: Huishuai Zhang

PLACE OF BIRTH: Puyang, Henan, China

DATE OF BIRTH: September 28, 1988

UNDERGRADUATE SCHOOLS ATTENDED:

University of Science and Technology of China, Hefei, China

GRADUATE SCHOOLS ATTENDED:

Graduate School of Chinese Academy of Sciences, Beijing, China