Median-Truncated Nonconvex Approach for Phase Retrieval with Outliers*†

Huishuai Zhang[s], Yuejie Chi[o] and Yingbin Liang[s]
[s]Department of EECS, Syracuse University, Syracuse, NY 13244
[o]Department of ECE, Ohio State University, Columbus, OH 43210

February 1, 2017

Abstract

This paper investigates the phase retrieval problem, which aims to recover a signal from the magnitudes of its linear measurements. We develop statistically and computationally efficient algorithms for the situation when the measurements are corrupted by sparse outliers that can take arbitrary values. We propose a novel approach to robustify the gradient descent algorithm by using the sample median as a guide for pruning spurious samples in initialization and local search. Adopting the Poisson loss and the reshaped quadratic loss respectively, we obtain two algorithms termed *median-TWF* and *median-RWF*, both of which provably recover the signal from a near-optimal number of measurements when the measurement vectors are composed of i.i.d. Gaussian entries, up to a logarithmic factor, even when a constant fraction of the measurements are adversarially corrupted. We further show that both algorithms are stable in the presence of additional dense bounded noise. Our analysis is accomplished by developing non-trivial concentration results of median-related quantities, which may be of independent interest. We provide numerical experiments to demonstrate the effectiveness of our approach.

1 Introduction

Phase retrieval is a classical problem in signal processing, optics and machine learning that has a wide range of applications such as X-ray crystallography [21], astronomical imaging, and diffraction imaging. Mathematically, it is formulated as recovering a signal $\boldsymbol{x} \in \mathbb{R}^n/\mathbb{C}^n$ from the magnitudes of its linear measurements:

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

where m is the total number of measurements, and $\boldsymbol{a}_i \in \mathbb{R}^n/\mathbb{C}^n$ is the ith known measurement vector, $i = 1, \ldots, m$. Phase retrieval is known to be notoriously difficult due to the quadratic form of the measurements. Classical methods based on alternating minimization between the signal of interest and the phase information [22], though computationally simple, are often trapped at local minima and lack rigorous performance guarantees.

There has been, however, a recent line of work that successfully develops provably accurate algorithms for phase retrieval, in particular for the case when the measurement vectors $\boldsymbol{a}_i$'s are composed of *independent and identically distributed* (i.i.d.) Gaussian entries. Broadly speaking, two classes of approaches have been proposed based on convex and nonconvex optimization techniques, respectively. Using the lifting trick, the phase retrieval problem can be reformulated as estimating a rank-one positive semidefinite (PSD) matrix $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^T$ from linear measurements [4], for which convex relaxations into semidefinite programs have been studied [8, 10, 16, 19, 35, 50]. In particular, Phaselift [10] perfectly recovers the signal with high probability

*The work of H. Zhang and Y. Liang is supported in part by the grants AFOSR FA9550-16-1-0077 and NSF ECCS 16-09916. The work of Y. Chi is supported in part by the grants NSF ECCS-1650449, AFOSR FA9550-15-1-0205 and ONR N00014-15-1-2387.

†The material in this paper was presented in part at the International Conference of Machine Learning (ICML), New York, USA, 2016.

as long as the number of measurements m is on the order of n. However, the computational complexity of Phaselift is at least cubic in n, which becomes expensive when n is large. Very recently, another convex relaxation has been proposed in the natural parameter space without lifting [3, 24, 26], resulting in a linear program that can handle large problem dimensions as long as m is on the order of n.

Another class of approaches aims to find the signal that minimizes a loss function based on certain postulated noise model, which often results in a nonconvex optimization problem due to the quadratic measurements. Despite nonconvexity, it is demonstrated in [9, 42] that the so-called Wirtinger flow (WF) algorithm, based on gradient descent, works remarkably well: it converges to the global optima when properly initialized using the spectral method. Several variants of WF have been proposed thereafter to further improve its performance, including the truncated Wirtinger flow (TWF) algorithm [13], the reshaped Wirtinger flow (RWF) algorithm [55], and the truncated amplitude flow (TAF) algorithm [51] Notably, TWF, RWF and TAF are shown to converge globally at a linear rate as long as m is on the order of n, and attain ϵ-accuracy within $\mathcal{O}(mn\log(1/\epsilon))$ flops using a constant step size.[1]

1.1 Outlier-Robust Phase Retrieval

The aforementioned algorithms are evaluated based on their *statistical* and *computational* performances: statistically, we wish the sample complexity m to be as small as possible; computationally, we wish the run time to be as fast as possible. As can be seen, existing WF-type algorithms are already near-optimal both statistically and computationally. This paper introduces a third consideration, which is the *robustness to outliers*, where we wish the algorithm continues to work well even in the presence of outliers that may take arbitrary magnitudes. This bears great importance in practice, because outliers arise frequently from the phase imaging applications [53] due to various reasons such as detector failures, recording errors, and missing data. Specifically, suppose the set of m measurements are given as

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

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 values, i.e., $\|\boldsymbol{\eta}\|_0 \le sm$, 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 are corrupted by outliers.

The goal of this paper is to develop phase retrieval algorithms with both statistical and computational efficiency, and provable robustness to even a constant proportion of outliers. None of the existing algorithms meet all of the three considerations simultaneously. The performance of WF-type algorithms is very sensitive to outliers which introduce anomalous search directions when their values are excessively deviated. While a form of Phaselift [25] is robust to a constant portion of outliers, it is computationally too expensive.

1.2 Median-Truncated Gradient Descent

A natural idea is to recover the signal as a solution to the following loss minimization problem:

$$\min_{\boldsymbol{z}} \frac{1}{2m}\sum_{i=1}^{m} \ell(\boldsymbol{z}; y_i) \tag{3}$$

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. We consider two choices of $\ell(\boldsymbol{z}; y_i)$ in this paper. The first one is the Poisson loss function of $|\boldsymbol{a}_i^T\boldsymbol{z}|^2$ employed in TWF [13], which is given by

$$\ell(\boldsymbol{z}; y_i) = |\boldsymbol{a}_i^T\boldsymbol{z}|^2 - y_i \log |\boldsymbol{a}_i^T\boldsymbol{z}|^2. \tag{4}$$

The second one is the *reshaped*[2] quadratic loss of $|\boldsymbol{a}_i^T\boldsymbol{z}|$ employed in RWF [55], which is given by

$$\ell(\boldsymbol{z}; y_i) = \left(|\boldsymbol{a}_i^T\boldsymbol{z}| - \sqrt{y_i}\right)^2. \tag{5}$$

[1] Notation $f(n) = \mathcal{O}(g(n))$ or $f(n) \lesssim g(n)$ means that there exists a constant $c > 0$ such that $|f(n)| \le c|g(n)|$.

[2] It is called "reshaped" in order to distinguish it from the quadratic loss of $|\boldsymbol{a}_i^T\boldsymbol{z}|^2$ used in [9].

Though (5) is not smooth everywhere, it resembles more closely the least-squares loss as if the phase information is available than the quadratic loss of $|\boldsymbol{a}_i^T\boldsymbol{z}|^2$, resulting in a more amenable curvature.

In the presence of outliers, the signal of interest may no longer be the global optima of (3). Therefore, we wish to only include the clean samples that are not corrupted in the optimization (3), 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:

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{m}\sum_{i\in T_{t+1}} \nabla \ell(\boldsymbol{z}^{(t)}; y_i). \tag{6}$$

where $\boldsymbol{z}^{(t)}$ denotes the tth iterate of the algorithm, $\nabla\ell(\boldsymbol{z}^{(t)}; y_i)$ is the gradient of $\ell(\boldsymbol{z}^{(t)}; y_i)$, and μ is the step size, for $t = 0, 1, \ldots$. 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 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 [28], 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 of each update. Hiding some technical details, this gives the main ingredient of our *median-truncated gradient descent* update rule[3], i.e., for each iterate $t \geq 0$:

$$T_{t+1} := \{i : |y_i - |\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|^2| \lesssim \mathsf{med}(\{|y_i - |\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|^2\}_{i=1}^m)\}, \quad \text{for TWF}, \tag{7}$$

$$T_{t+1} := \{i : |\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|| \lesssim \mathsf{med}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|\}_{i=1}^m)\}, \quad \text{for RWF}, \tag{8}$$

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 [28]. 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. Finally, there still left the question of initialization, which is critical to the success of the algorithm. We use the spectral method, i.e., initialize $\boldsymbol{z}^{(0)}$ by a proper rescaling of the top eigenvector of a surrogate matrix

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i\in T_0} y_i \boldsymbol{a}_i\boldsymbol{a}_i^T, \tag{9}$$

where again T_0 includes only a subset of samples whose values are not excessively large compared with the sample median of the measurements, given as

$$T_0 = \{i : y_i \lesssim \mathsf{med}(\{y_i\}_{i=1}^m)\}. \tag{10}$$

Putting things together (the update rule (6) and the initialization (9)), we obtain two new median-truncated gradient descent algorithms, median-TWF and median-RWF, based on applying the median truncation strategy for the loss functions used in TWF and 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.

Main Result (informal): For the Gaussian measurement model, 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

[3]Please see the exact form of the algorithms in Section 2.

the best of our knowledge, this 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.

To establish the performance guarantees, we first show that the initialization is close enough to the ground truth, and then that within the neighborhood of the ground truth, the gradients satisfy certain *Regularity Condition* [9, 13] that guarantees linear convergence of the descent rule, as long as the fraction of outliers is small enough and the sample complexity is large enough. As a nonlinear operator, the sample median is much more difficult to analyze than the sample mean, which is a linear operator and many existing concentration inequalities are readily applicable. Therefore, considerable technical efforts are devoted to develop novel non-asymptotic concentrations of the sample median, and various statistical properties of the sample median related quantities, which may be of independent interest.

Finally, we note that while median-TWF and median-RWF share similar theoretical performance guarantees, their empirical performances vary under different scenarios, due to the use of different loss functions. Their theoretical analyses also have significant difference that worth separate treatments. While we only consider the loss functions used in TWF and RWF in this paper, we believe the median-truncation technique can be applied to gradient descent algorithms for solving other problems as well.

1.3 Related Work

Our work is closely related to the TWF algorithm [13], which is also 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 [25, 37, 53], 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 [12] and resilient data aggregation for sensor networks [49]. 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 [28], which has found success in robustifying sparse regression [15], subspace clustering [41], etc. However, using the trimmed mean requires knowledge of an upper bound on the number of outliers, whereas median does not require such information.

Developing non-convex algorithms with provable global convergence guarantees has attracted intensive research interest recently. A partial list of these studies include phase retrieval [9, 13, 38, 44, 51], matrix completion [18, 23, 27, 29–31, 45, 57], low-rank matrix recovery [6, 17, 33, 36, 40, 47, 52, 56], robust PCA [39, 54], robust tensor decomposition [1], dictionary learning [2, 43], community detection [5], phase synchronization [7], blind deconvolution [32, 34], joint alignment [14], etc. Our algorithm provides a new instance in this list that emphasizes robust high-dimensional signal estimation under minimal assumptions of outliers.

1.4 Paper Organization and Notations

The rest of this paper is organized as follows. Section 2 describes the proposed two algorithms, median-TWF and median-RWF, in details and their performance guarantees. Section 3 presents numerical experiments. Section 4 provides the preliminaries and the proof road map. Section 5 provides the proofs for median-TWF and Section 6 provides the proofs of median-RWF, respectively. Finally, we conclude in Section 7. Supporting proofs are given in the Appendix.

We adopt the following notations in this paper. Given a set of numbers $\{y_i\}_{i=1}^m$, the sample median is denoted as $\mathsf{med}(\{y_i\}_{i=1}^m)$. The indicator function $\mathbf{1}_A = 1$ if the event A holds, and $\mathbf{1}_A = 0$ otherwise. For a vector $\boldsymbol{y}$, $\|\boldsymbol{y}\|$ denotes the l_2 norm. For two matrices, $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is a positive semidefinite matrix.

2 Algorithms and Performance Guarantees

We consider the following model for phase retrieval, where the measurements are corrupted by not only sparse arbitrary outliers but also dense bounded noise. Under such a model, the measurements are given as

$$y_i = |\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle|^2 + w_i + \eta_i, \quad i = 1, \cdots, m, \tag{11}$$

Algorithm 1 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;
Initialization: Let $\boldsymbol{z}^{(0)} = \lambda_0 \tilde{\boldsymbol{z}}$, where $\lambda_0 = \sqrt{\mathsf{med}(\boldsymbol{y})/0.455}$ and $\tilde{\boldsymbol{z}}$ is the leading eigenvector of

$$\boldsymbol{Y} := \frac{1}{m}\sum_{i=1}^m y_i \boldsymbol{a}_i \boldsymbol{a}_i^T \mathbf{1}_{\{|y_i| \le \alpha_y^2 \lambda_0^2\}}. \tag{14}$$

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

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{m}\sum_{i=1}^m \frac{|\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}|^2 - y_i}{\boldsymbol{a}_i^T \boldsymbol{z}^{(t)}} \boldsymbol{a}_i \mathbf{1}_{\mathcal{E}_1^i \cap \mathcal{E}_2^i}, \tag{15}$$

where

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

Output $\boldsymbol{z}_T$.

where $\boldsymbol{x} \in \mathbb{R}^n$ is the unknown signal[4], $\boldsymbol{a}_i \in \mathbb{R}^n$ is the ith measurement vector composed of $i.i.d.$ Gaussian entries distributed as $\mathcal{N}(0,1)$, and $\eta_i \in \mathbb{R}$ for $i = 1, \ldots, m$ are outliers with arbitrary values satisfying $\|\boldsymbol{\eta}\|_0 \le sm$, where s is the fraction of outliers, and $\boldsymbol{w} = \{w_i\}_{i=1}^m$ is the bounded noise satisfying $\|\boldsymbol{w}\|_\infty \le c\|\boldsymbol{x}\|^2$ for some universal constant c.

It is straightforward that changing the sign of the signal does not affect the measurements. The goal is to recover the signal $\boldsymbol{x}$, up to a global sign difference, from the measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$ and the measurement vectors $\{\boldsymbol{a}_i\}_{i=1}^m$. To this end, we define the Euclidean distance between two vectors up to a global sign difference as the performance metric,

$$\mathrm{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min\{\|\boldsymbol{z} + \boldsymbol{x}\|, \|\boldsymbol{z} - \boldsymbol{x}\|\}. \tag{12}$$

We propose two median-truncated gradient descent algorithms, median-TWF in Section 2.1 and median-RWF in Section 2.2, based on different choices of the loss functions. This leads to applying the truncation based on the sample median of $\left\{\left|y_i - |\boldsymbol{a}_i^T \boldsymbol{z}|^2\right|\right\}_{i=1}^m$ in median-TWF, and the sample median of $\left\{\left|y_i - |\boldsymbol{a}_i^T \boldsymbol{z}|^2\right|\right\}_{i=1}^m$ in median-RWF. Section 2.3 provides the theoretical performance guarantees of median-TWF and median-RWF, which turn out to be almost the same at the order level except the choice of constants. The empirical comparisons of median-TWF and median-RWF are demonstrated in Section 3.

2.1 Median-TWF Algorithm

In median-TWF, we adopt the Poisson loss function of $|\boldsymbol{a}_i^T \boldsymbol{z}|^2$ employed in TWF [13], given as

$$\ell(\boldsymbol{z}) := \frac{1}{2m}\sum_{i=1}^m \left(|\boldsymbol{a}_i^T \boldsymbol{z}|^2 - y_i \log |\boldsymbol{a}_i^T \boldsymbol{z}|^2 \right). \tag{13}$$

The median-TWF algorithm, as described in Algorithm 1, gradually eliminates the influence of outliers on the way of minimizing (13). Specifically, it comprises an initialization step and a truncated gradient descent step.

1. **Initialization**: As in (14), we initialize $\boldsymbol{z}^{(0)}$ by the spectral method using a truncated set of samples, where the threshold is determined by $\mathsf{med}(\{y_i\}_{i=1}^m)$. As will be shown in Section 4.2, as long as the fraction

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

Algorithm 2 Median Reshaped Wirtinger Flow (median-RWF)

Input: $\boldsymbol{y} = \{y_i\}_{i=1}^m$, $\{\boldsymbol{a}_i\}_{i=1}^m$;
Parameters: threshold α_h', and step size μ;
Initialization: Same as median-TWF (see Algorithm 1).
Gradient loop: for $t = 0 : T-1$ do

$$\boldsymbol{z}^{(t+1)} = \boldsymbol{z}^{(t)} - \frac{\mu}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)} - \sqrt{y_i}\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i}, \tag{18}$$

where

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

Output $\boldsymbol{z}_T$.

of outliers is not too large and the sample complexity is large enough, our initialization 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

$$\nabla\ell_{tr}(\boldsymbol{z}^{(t)}) = \frac{1}{m}\sum_{i=1}^{m}\frac{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|^2 - y_i}{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}\boldsymbol{a}_i\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i}. \tag{16}$$

In order to remove the contribution of corrupted samples, from the definition of the set $\mathcal{E}_2^i$ (see Algorithm 1), it is clear that samples are truncated if their measurement residuals evaluated using the current iterate are much larger than the sample median. Samples are also truncated according to the set $\mathcal{E}_1^i$, which removes the contribution of samples outside some confidence interval to better control the search direction, since $\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{z}|] = \sqrt{2/\pi}\|\boldsymbol{z}\|$. The median-TWF algorithm closely resembles the TWF algorithm, except that the truncation is guided by the sample median, rather than the sample mean.

We set the step size in 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

$$\begin{aligned}
&\zeta_1 := \max\left\{\mathbb{E}\left[\xi^2\mathbf{1}_{\left\{|\xi|<\sqrt{1.01}\alpha_l \text{ or } |\xi|>\sqrt{0.99}\alpha_u\right\}}\right], \mathbb{E}\left[\mathbf{1}_{\left\{|\xi|<\sqrt{1.01}\alpha_l \text{ or } |\xi|>\sqrt{0.99}\alpha_u\right\}}\right]\right\},\\
&\zeta_2 := \mathbb{E}\left[\xi^2\mathbf{1}_{\{|\xi|>0.248\alpha_h\}}\right],\\
&2(\zeta_1+\zeta_2) + \sqrt{8/\pi}\alpha_h^{-1} < 1.99\\
&\alpha_y \ge 3,
\end{aligned} \tag{17}$$

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

2.2 Median-RWF Algorithm

In median-RWF, we adopt the reshaped quadratic loss function of $|\boldsymbol{a}_i^T\boldsymbol{z}|$ employed in RWF [55], given as

$$\mathcal{R}(\boldsymbol{z}) = \frac{1}{2m}\sum_{i=1}^{m}\left(\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}|\right)^2, \tag{19}$$

which has been shown to be advantageous over other loss functions for phase retrieval [55].

Similarly to median-TWF, the median-RWF algorithm as described in Algorithm 2, gradually eliminates the influence of outliers on the way of minimizing (19). Specifically, it also comprises an initialization step and a truncated gradient descent step.

1. **Initialization**: we initialize in the same manner as in median-TWF (Algorithm 1).

2. **Gradient loop**: for each iteration $0 \le t \le T-1$, median-RWF uses the following iteration-varying truncated gradient:

$$\nabla \mathcal{R}_{tr}(\boldsymbol{z}^{(t)}) = \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^T\boldsymbol{z}^{(t)} - \sqrt{y_i}\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}}{|\boldsymbol{a}_i^T\boldsymbol{z}^{(t)}|}\right)\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i}, \tag{20}$$

From the definition of the set $\mathcal{T}^i$ (see Algorithm 2), samples are truncated by the sample median of gradient components evaluated at the current iteration. We set the step size in median-RWF to be a fixed small constant, i.e., $\mu = 0.8$. Compared with median-TWF, the truncation rule is much simpler with fewer parameters. We simply set the truncation threshold $\alpha_h' = 5$. It is possible that including $\mathcal{E}_1^i$ may further improves the performance, but we wish to highlight that, in this paper, the simple truncation rule is already sufficient to guarantee both robustness and efficiency of median-RWF.

2.3 Performance Guarantees

In this section, we characterize the performance guarantees of median-TWF and median-RWF, which turn out to be very similar though the proofs in fact involve quite different techniques. To avoid repetition, we present the guarantees together for both algorithms. We note that the values of constants in the results can vary for median-TWF and median-RWF.

We first show that median-TWF/median-RWF performs 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 1 (**Exact recovery for the noise-free model**). *Suppose that the measurements are noise-free, i.e., $\eta_i = 0$ and $w_i = 0$ for $i = 1, \cdots, m$ in the model* (11). *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 $1 - c_1\exp(-c_2 m)$, median-TWF/median-RWF yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \le \nu(1-\rho)^t\|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \tag{21}$$

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

Proposition 1 suggests that median-TWF/median-RWF allows exact recovery at a linear rate of convergence as long as the sample complexity is on the order of $n\log n$, which is in fact slightly worse, by a logarithmic factor, than existing WF-type algorithms (TWF, RWF and TAF) for the noise-free model. This is a price due to working with the nonlinear operator of median in the proof, and it is not clear whether it is possible to further improve the result. Nonetheless, as the median is quite stable as long as the number of outliers is not so large, the following main theorem indeed establishes that median-TWF/median-RWF still performs well even in the presence of a constant fraction of sparse outliers with the same sample complexity.

Theorem 1 (**Exact recovery with sparse arbitrary outliers**). *Suppose that the measurements are corrupted by sparse outliers, i.e., $w_i = 0$ for $i = 1, \cdots, m$ in the model* (11). *There exist constants μ_0, $s_0 > 0$, $0 < \rho, \nu < 1$ and $c_0, c_1, c_2 > 0$ such that if $m \ge c_0 n\log n$, $s < s_0$, $\mu \le \mu_0$, then with probability at least $1 - c_1\exp(-c_2 m)$, median-TWF/median-RWF yields*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \le \nu(1-\rho)^t\|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \tag{22}$$

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

Theorem 1 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 sample complexity satisfies $m \gtrsim n\log n$. Moreover, median-TWF/median-RWF converges at a linear rate using a constant step size, with per-iteration cost $\mathcal{O}(mn)$ (note that the median can be computed in linear time [46]). To reach ϵ-accuracy, i.e., $\text{dist}(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \le \epsilon$, only $\mathcal{O}(\log 1/\epsilon)$ iterations are needed, yielding the total computational cost as $\mathcal{O}(mn\log 1/\epsilon)$, which is highly efficient. Empirically in the numerical experiments in Section 3, median-RWF converges faster and tolerates a larger fraction of outliers than median-TWF, which can be due to the use of the reshaped quadratic loss function.

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 stable to coexistence of the two types of noises.

Theorem 2 (**Stability to sparse arbitrary outliers and dense bounded noises**). *Consider the phase retrieval problem given in* (11) *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 and median-RWF respectively yield*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|} + (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \quad \text{for median-TWF}, \tag{23}$$

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \sqrt{\|\boldsymbol{w}\|_\infty} + (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \quad \text{for median-RWF}, \tag{24}$$

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

Theorem 2 immediately implies the stability of median-TWF/median-RWF when the measurements are only corrupted by dense bounded noise.

Corollary 1. *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* (11)*. 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 and median-RWF respectively yield*

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|} + (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \quad \text{for median-TWF}, \tag{25}$$

$$dist(\boldsymbol{z}^{(t)}, \boldsymbol{x}) \lesssim \sqrt{\|\boldsymbol{w}\|_\infty} + (1-\rho)^t \|\boldsymbol{x}\|, \quad \forall t \in \mathbb{N} \quad \text{for median-RWF}, \tag{26}$$

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

With both sparse arbitrary outliers and dense bounded noises, Theorem 2 and Corollary 1 imply that median-TWF/median-RWF achieves the same convergence rate and the same level of estimation error as the model with only bounded noise. In fact, together with Theorem 1 and Proposition 1, it can be seen that applying median-TWF/median-RWF does not require the knowledge of the existence of outliers. When there do exist outliers, median-TWF/median-RWF achieves almost the same performance *as if outliers do not exist.*

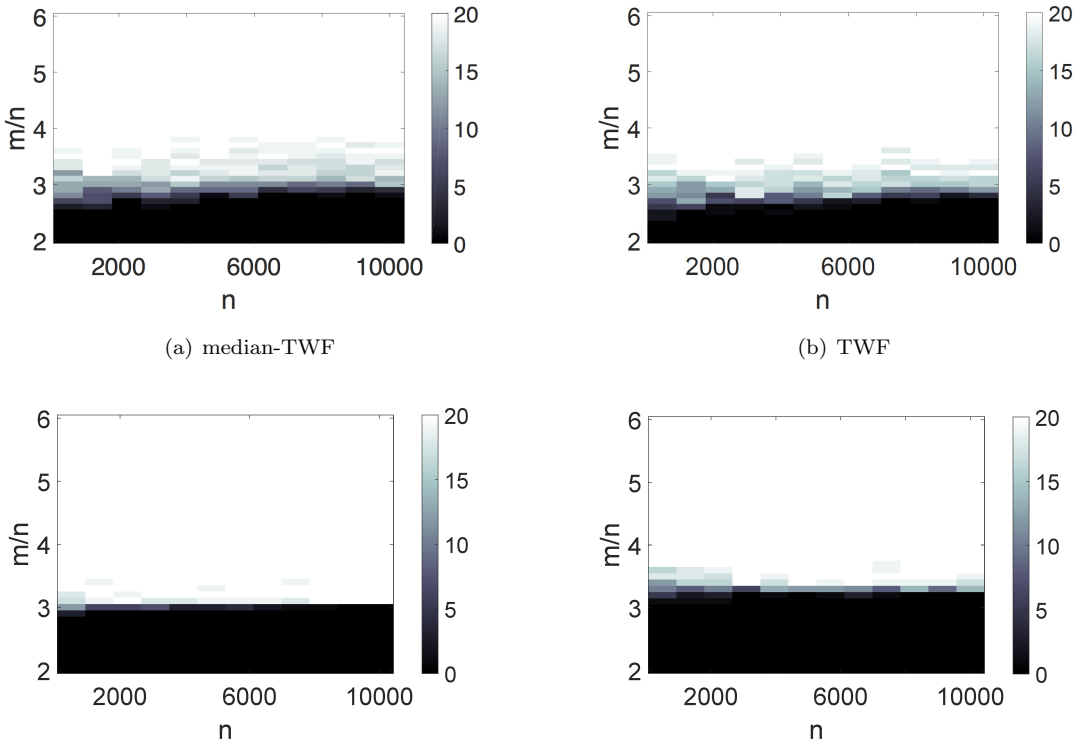
3 Numerical Experiments

In this section, we provide numerical experiments to demonstrate the effectiveness of median-TWF and median-RWF, which corroborate our theoretical findings.

3.1 Exact Recovery for Noise-free Data

We first show that, in the noise-free case, median-TWF and median-RWF provide similar performance as TWF [13] and RWF [55] for exact recovery. We set the parameters of median-TWF and median-RWF as specified in Section 2.1 and Section 2.2, and those of TWF and RWF as suggested in [13] and [55], respectively. Let the signal length n take values from 1000 to 10000 by a step size of 1000, and the ratio of the number of measurements to the signal dimension, m/n, take values from 2 to 6 by a step size of 0.1. For each pair of $(n, m/n)$, we generate a signal $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n\times n})$, and the measurement vectors $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{n\times n})$ i.i.d. for $i = 1, \ldots, m$. For all 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 1 shows the number of successful trials out of 20 trials for all algorithms, with respect to m/n and n. It can be seen that, as soon as m is above $4n$, exact recovery is achieved for all four algorithms. Around the phase transition boundary, the empirical sample complexity 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 [28]. Interestingly, the empirical sample complexity of median-RWF is slightly better than RWF because the truncation rule used in median-RWF allows sample pruning that improves the performance.[5]



(a) median-TWF (b) TWF

(c) median-RWF (d) RWF

Figure 1: Sample complexity of median-TWF, TWF, median-RWF, and RWF for noise-free data: the gray scale of each cell $(n, m/n)$ indicates the number of successful recovery out of 20 trials.

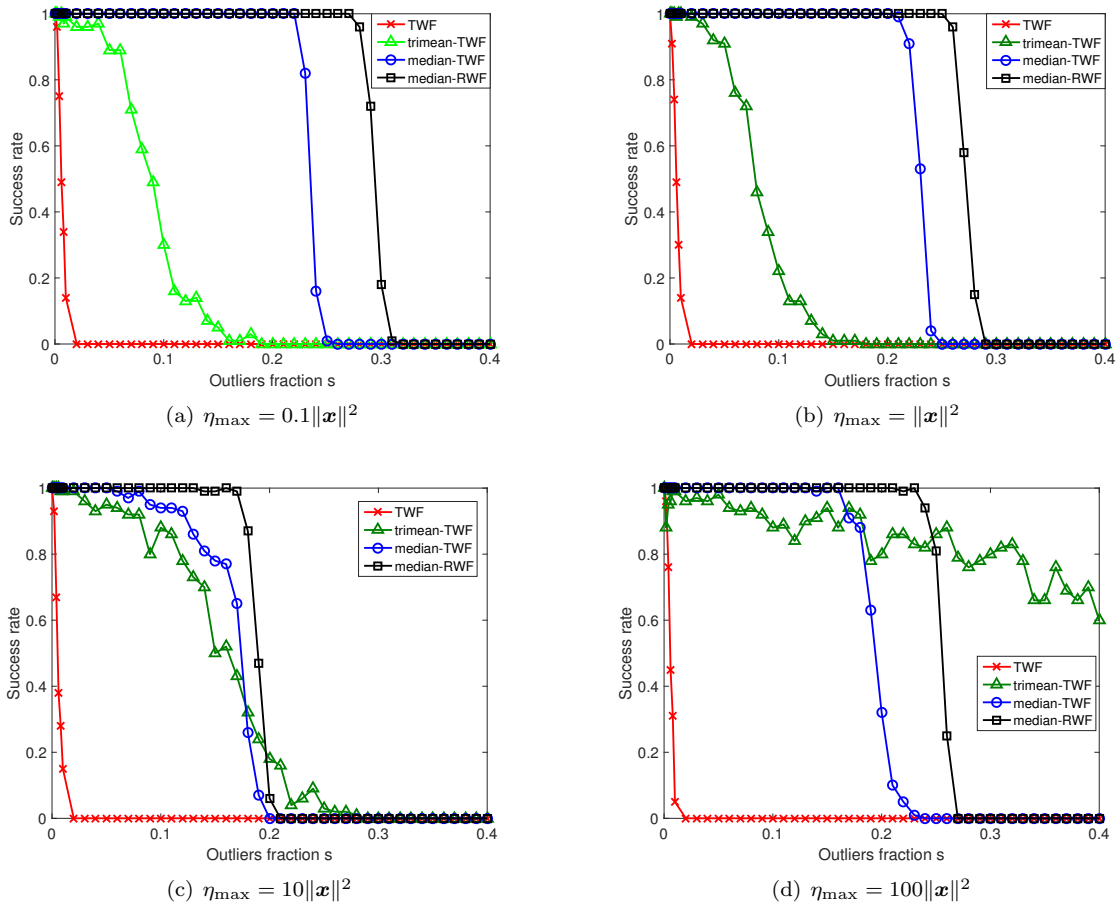
3.2 Exact Recovery with Sparse Outliers

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 TWF, and also an alternative which we call *trimean-TWF*, based on replacing the sample mean in TWF by the *trimmed mean* for truncation purpose. Specifically, trimean-TWF requires knowing the fraction s of outliers so that samples corresponding to sm largest values in the measurements or gradients are first removed, and truncation is then applied based on the mean of the remaining samples similar to TWF.

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

From Figure 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 assuming any knowledge of the outliers, which validates our theoretical analysis. Empirically, median-RWF can tolerate a larger fraction of outliers than median-TWF. This could be due to the fact that the lower-order objective adopted in median-RWF reduces the variance and allows more stable search direction. Unsurprisingly, TWF fails quickly even with a very

[5]The original RWF in [55] does not have sample truncation.



(a) $\eta_{\max} = 0.1\|\boldsymbol{x}\|^2$ (b) $\eta_{\max} = \|\boldsymbol{x}\|^2$

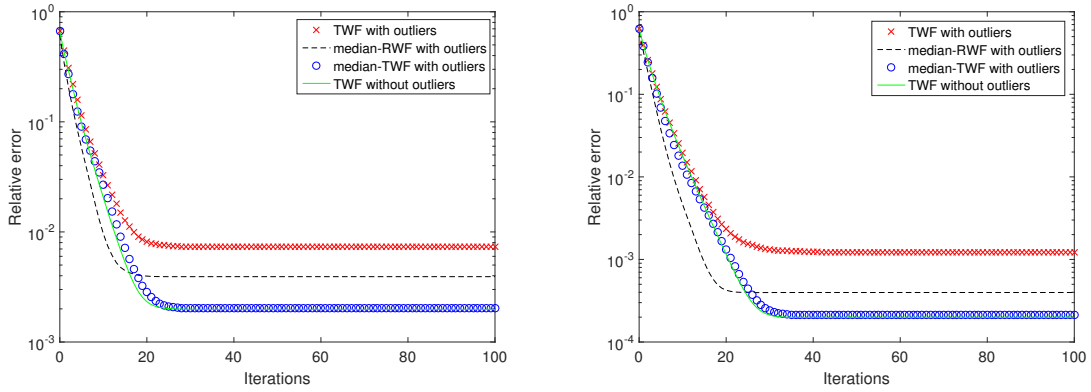
(c) $\eta_{\max} = 10\|\boldsymbol{x}\|^2$ (d) $\eta_{\max} = 100\|\boldsymbol{x}\|^2$

Figure 2: Success rate of exact recovery with respect to the fraction of sparse outliers for median-TWF, median-RWF, trimean-TWF, and TWF at different levels of outlier magnitudes.

small fraction of outliers. No successful instance is observed for TWF when $s \geq 0.02$ irrespective of the value of $\eta_{\max}$. Trimean-TWF, even knowing the number of outliers, still does not exhibit a sharp phase transition, and in general underperforms the proposed median-TWF and median-RWF, except when both $\eta_{\max}$ and s get very large, see Figure 2(d). This is because in this range of large outliers, knowing the fraction s of outliers provides substantial advantage for trimean-TWF to eliminate them, while the sample median can deviate 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 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 2(d); and meanwhile, the outlier magnitudes in Figure 2(c) are large enough to affect the accuracy heavily in contrast to the cases in Figures 2(a) and 2(b) where outliers are less prominent.

3.3 Stable Recovery with Sparse Outliers and Dense Noise

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 term $\boldsymbol{w}$ are generated independently from $\mathcal{U}(0, w_{\max})$. The entries of the sparse outlier are then generated as $\eta_i \sim \|\boldsymbol{w}\| \cdot \text{Bernoulli}(0.1)$ independently. Figure 3(a) and Figure 3(b) depict the relative error $\text{dist}(\boldsymbol{z}^{(t)}, \boldsymbol{x})/\|\boldsymbol{x}\|$ with respect to the iteration count t, when $w_{\max}/\|\boldsymbol{x}\|^2 = 0.001$ and 0.01 respectively. In the presence of sparse outliers, it can be seen that both median-TWF and median-RWF clearly outperforms TWF under the same situation, and acts as if



(a) $w_{\max} = 0.01\|\boldsymbol{x}\|^2$ (b) $w_{\max} = 0.001\|\boldsymbol{x}\|^2$

Figure 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. In (a) and (b), the dense noise is generated uniformly at different levels.

the outliers do not exist by achieving almost the same accuracy as TWF without outliers. Moreover, the relative error of the reconstruction using median-TWF or median-RWF has 10 times gain from Figure 3(a) to Figure 3(b) as $w_{\max}$ shrinks by a factor of 10, which corroborates Theorem 2 nicely. Furthermore, it can be seen that median-RWF converges faster than the other algorithms, due to the improved curvature of using low-order objectives, corroborating the result in [55]. On the other hand, median-TWF returns more accurate estimates, due to employing more delicate truncation rules that may help reduce the noise.

Finally, we consider the case when the measurements are corrupted by both Poisson noise and outliers, modeling 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 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 outliers and Poisson noise have almost the same accuracy as, if not better than, TWF under only the Poisson noise.

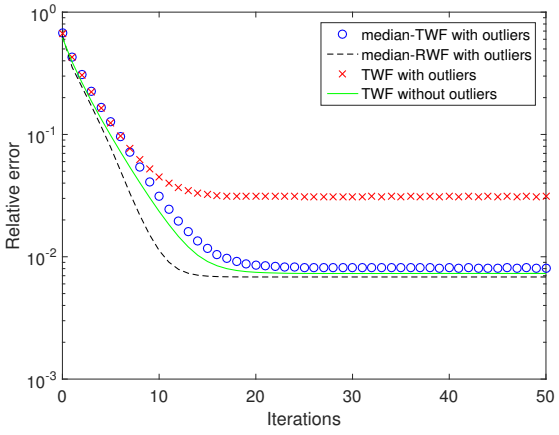


Figure 4: The relative error with respect to the iteration count for median-TWF, median-RWF and TWF with both Poisson noise and sparse outliers, and TWF with only Poisson noise.

4 Preliminaries and 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* [9], 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 that will be useful throughout our analysis in Section 4.1. Section 4.2 analyzes the initialization that is used in both algorithms. We then state the definition of Regularity Condition in Section 4.3 and explain how it leads to the linear convergence rate. We provide separate detailed proofs for two algorithms in Section 5 and Section 6, respectively, because they involve different bounding techniques that may be of independent interest due to different loss functions.

4.1 Properties of Median

We start by the definitions of the quantile of a population distribution and its sample version.

Definition 1 (Generalized quantile function)**.** *Let $0 < p < 1$. For a cumulative distribution function (CDF) F, the generalized quantile function is defined as*

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}. \tag{27}$$

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 $\mathsf{med}(\{X_i\}) = \theta_{1/2}(\{X_i\})$, 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 [11]), as in Lemma 1.

Lemma 1. *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*

$$|\theta_p(\{X_i\}_{i=1}^m) - \theta_p(F)| < \epsilon \tag{28}$$

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

Proof. See Appendix A.1. □

Lemma 2 bounds the distance between the median of two sequences.

Lemma 2. *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

$$|X_{(k)} - Y_{(k)}| \leq \|\boldsymbol{X} - \boldsymbol{Y}\|_\infty, \tag{29}$$

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

Proof. See Appendix A.2. □

Lemma 3, 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 3. *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 Appendix A.3. □

Finally, Lemma 4 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. *Let $u, v \sim \mathcal{N}(0,1)$ which can be correlated with the correlation coefficient $|\rho| \le 1$. 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 Appendix A.4. □

4.2 Robust Initialization with Outliers

Considering the model that the measurements are corrupted by both bounded noise and sparse outliers given by (11), we show that the initialization provided by the median-truncated spectral method in (14) is close enough to the ground truth, i.e., $\mathrm{dist}(\boldsymbol{z}^{(0)}, \boldsymbol{x}) \le \delta\|\boldsymbol{x}\|$.

Proposition 2. *Fix $\delta > 0$ and $\boldsymbol{x} \in \mathbb{R}^n$, and consider the model given by (11). Suppose that $\|\mathbf{w}\|_\infty \le c\|\boldsymbol{x}\|^2$ for some sufficiently small constant $c > 0$ and that $\|\eta\|_0 \le sm$ for some sufficiently small constant s. With probability at least $1 - \exp(-\Omega(m))$, the initialization given by the median-truncated spectral method obeys*[6]

$$dist(\boldsymbol{z}^{(0)}, \boldsymbol{x}) \le \delta\|\boldsymbol{x}\|, \tag{30}$$

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

Proof. See Appendix B.

4.3 Regularity Condition

Once the initialization is guaranteed to be within a small neighborhood of the ground truth, we only need to show that the truncated gradient (16) and (20) satisfy the *Regularity Condition* (RC) [9, 13], which guarantees the geometric convergence of median-TWF/median-RWF once the initialization lands into this neighborhood.

Definition 2. *The gradient $\nabla\ell(\boldsymbol{z})$ is said to satisfy the Regularity Condition $\mathsf{RC}(\mu, \lambda, c)$ if*

$$\langle \nabla\ell(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}\rangle \ge \frac{\mu}{2}\|\nabla\ell(\boldsymbol{z})\|^2 + \frac{\lambda}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2 \tag{31}$$

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

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 [13] if $\mu\lambda < 1$. We repeat this argument below for completeness.

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

5 Proofs for Median-TWF

We first show that $\nabla\ell_{tr}(\boldsymbol{z})$ in (16) satisfies the RC for the noise-free case in Section 5.1, and then extend it to the model with only sparse outliers in Section 5.2, thus together with Proposition 2 establishing the global convergence of median-TWF in both cases. Section 5.3 proves Theorem 2 in the presence of both sparse outliers and dense bounded noise.

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

5.1 Proof of Proposition 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 3. *If $m > c_0 n \log n$, then with probability at least $1 - c_1 \exp(-c_2 m)$,*

$$0.6\|\boldsymbol{z}\|\|\boldsymbol{z}-\boldsymbol{x}\| \le \theta_{0.49}, \theta_{0.5}, \theta_{0.51}(\{\big||\boldsymbol{a}_i^T\boldsymbol{x}|^2 - |\boldsymbol{a}_i^T\boldsymbol{z}|^2\big|\}_{i=1}^m) \le \|\boldsymbol{z}\|\|\boldsymbol{z}-\boldsymbol{x}\|, \tag{32}$$

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

Proof. Detailed proof is provided in Appendix C.1. □

We note that a similar property for the sample mean has been shown in [13] 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, and that $\|\nabla \ell_{tr}(\boldsymbol{z})\|$ is upper bounded on the order of $\|\boldsymbol{z}-\boldsymbol{x}\|$, as in Proposition 5.

Proposition 4 (Adapted version of Proposition 2 of [13])**.** *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* (17)*, if $m > c_0 n \log n$, then with probability at least $1 - c_1 \exp(-c_2\epsilon^{-2} m)$,*

$$\langle \nabla \ell_{tr}(\boldsymbol{z}), \boldsymbol{z}-\boldsymbol{x}\rangle \ge \left\{1.99 - 2(\zeta_1+\zeta_2) - \sqrt{8/\pi}\alpha_h^{-1} - \epsilon\right\}\|\boldsymbol{z}-\boldsymbol{x}\|^2 \tag{33}$$

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

$$\frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{z}\|} \le \min\left\{\frac{1}{11}, \frac{\alpha_l}{\alpha_h}, \frac{\alpha_l}{6}, \frac{\sqrt{98/3}(\alpha_l)^2}{2\alpha_u + \alpha_l}\right\}, \tag{34}$$

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 (17)*.*

The proof of Proposition 4 adapts the proof of Proposition 2 of [13], by properly setting parameters based on the properties of sample median. For completeness, we include a short outline of the proof in Appendix C.2.

Proposition 5 (Lemma 7 of [13])**.** *Under the same condition as in Proposition 4, if $m > c_0 n$, then there exist some constants $c_1, c_2 > 0$ such that with probability at least $1 - c_1 \exp(-c_2 m)$,*

$$\|\nabla \ell_{tr}(\boldsymbol{z})\| \le (1+\delta)\cdot 2\sqrt{1.02 + 2/\alpha_h}\|\boldsymbol{z}-\boldsymbol{x}\| \tag{35}$$

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

$$\frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{z}\|} \le \min\left\{\frac{1}{11}, \frac{\alpha_l}{\alpha_h}, \frac{\alpha_l}{6}, \frac{\sqrt{98/3}(\alpha_l)^2}{2\alpha_u + \alpha_l}\right\}, \tag{36}$$

where δ can be arbitrarily small as long as m/n sufficiently large, and α_l, α_u and α_h are given in (17)*.*

Proof. See the proof of Lemma 7 in [13]. □

With these two propositions and (17), RC is guaranteed by setting

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

5.2 Proof of Theorem 1

We next consider the model (11) 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 3 yields

$$\theta_{\frac{1}{2}-s}(\{|(\boldsymbol{a}_i^T\boldsymbol{x})^2-(\boldsymbol{a}_i^T\boldsymbol{z})^2|\}) \le \theta_{\frac{1}{2}}(\{|y_i-(\boldsymbol{a}_i^T\boldsymbol{z})^2|\}) \le \theta_{\frac{1}{2}+s}(\{|(\boldsymbol{a}_i^T\boldsymbol{x})^2-(\boldsymbol{a}_i^T\boldsymbol{z})^2|\}. \tag{37}$$

For simplicity of notation, we let $\boldsymbol{h} := \boldsymbol{z}-\boldsymbol{x}$. Then for the instance of $s=0.01$, by Proposition 3, we have with probability at least $1-2\exp(-\Omega(m))$,

$$0.6\|\boldsymbol{z}\|\|\boldsymbol{h}\| \le \theta_{\frac{1}{2}}(\{|y_i-(\boldsymbol{a}_i^T\boldsymbol{z})^2|\}) \le \|\boldsymbol{z}\|\|\boldsymbol{h}\|. \tag{38}$$

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

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

Choosing ϵ small enough, it is easy to verify that Propositions 4 and 5 are still valid on $\nabla^{clean}\ell_{tr}(\boldsymbol{z})$. Thus, one has

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

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 $|q_i| \le 2\alpha_h\|\boldsymbol{h}\|$. Thus $\|\boldsymbol{q}\| \le \sqrt{sm}\cdot 2\alpha_h\|\boldsymbol{h}\|$, and

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

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

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

and

$$\begin{aligned}
\|\nabla \ell_{tr}(\boldsymbol{z})\| &\le \left\|\nabla^{clean}\ell_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{extra}\ell_{tr}(\boldsymbol{z})\right\| \\
&\le 2(1+\delta)\left(\sqrt{1.02+2/\alpha_h}+\sqrt{s}\alpha_h\right)\|\boldsymbol{h}\|.
\end{aligned} \tag{39}$$

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

5.3 Proof of Theorem 2

We consider the model (11), and split our analysis of the gradient loop into two regimes.

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

$$\operatorname{dist}\left(\boldsymbol{z} - \mu\nabla\ell_{tr}(\boldsymbol{z}), \boldsymbol{x}\right) \leq (1-\rho)\operatorname{dist}(\boldsymbol{z}, \boldsymbol{x}).$$

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 3, we have

$$\theta_{\frac{1}{2}-s}\left\{\left|\tilde{y}_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} \leq \mathsf{med}\left\{\left|y_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} \leq \theta_{\frac{1}{2}+s}\left\{\left|\tilde{y}_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\}.$$

Moreover, by Lemma 2 we have

$$\begin{aligned}
&\left|\theta_{\frac{1}{2}+s}\left\{\left|\tilde{y}_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} - \theta_{\frac{1}{2}+s}\left\{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\}\right| \leq \|\boldsymbol{w}\|_\infty,\\
&\left|\theta_{\frac{1}{2}-s}\left\{\left|\tilde{y}_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} - \theta_{\frac{1}{2}-s}\left\{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\}\right| \leq \|\boldsymbol{w}\|_\infty.
\end{aligned}$$

Assume that $s = 0.01$. By Proposition 3, if c_3 is sufficiently large (i.e., $c_3 > 100$), we still shave

$$0.6\|\boldsymbol{x}-\boldsymbol{z}\|\|\boldsymbol{z}\| \leq \mathsf{med}\left\{\left|y_i - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right|\right\} \leq \|\boldsymbol{x}-\boldsymbol{z}\|\|\boldsymbol{z}\|. \tag{40}$$

Furthermore, recall $\tilde{\mathcal{E}}_2^i := \left\{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \leq \alpha_h\mathsf{med}\left\{\left|(\boldsymbol{a}_i^T\boldsymbol{z})^2 - y_i\right|\right\}\frac{|\boldsymbol{a}_i^T\boldsymbol{z}|}{\|\boldsymbol{z}\|}\right\}$. 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 $|w_i| \leq \frac{1}{c_3}\|\boldsymbol{h}\|\|\boldsymbol{z}\|$ 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 (40). All the proof arguments for Propositions 4 and 5 are also valid for $\nabla^{clean}\ell_{tr}(\boldsymbol{z})$, and thus we have

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

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

$$\|\nabla^{noise}\ell_{tr}(\boldsymbol{z})\| = \left\|\frac{1}{m}\boldsymbol{A}^T\tilde{\boldsymbol{w}}\right\| \leq \left\|\frac{1}{\sqrt{m}}\boldsymbol{A}^T\right\|\left\|\frac{\tilde{\boldsymbol{w}}}{\sqrt{m}}\right\| \leq (1+\delta)\|\tilde{\boldsymbol{w}}\|_\infty \leq (1+\delta)\frac{\|\boldsymbol{w}\|_\infty}{\alpha_l\|\boldsymbol{z}\|},$$

when m/n is sufficiently large. Given the regime condition $\|\boldsymbol{h}\| \geq c_3\frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{z}\|}$, we further have

$$\begin{aligned}
&\|\nabla^{noise}\ell_{tr}(\boldsymbol{z})\| \leq \frac{(1+\delta)}{c_3\alpha_l}\|\boldsymbol{h}\|,\\
&\left|\left\langle\nabla^{noise}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \leq \left\|\nabla^{noise}\ell_{tr}(\boldsymbol{z})\right\|\cdot\|\boldsymbol{h}\| \leq \frac{(1+\delta)}{c_3\alpha_l}\|\boldsymbol{h}\|^2.
\end{aligned}$$

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

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

Putting these together, one has

$$\begin{aligned} \langle \nabla\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle &\ge \left\langle \nabla^{clean}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle - \left|\left\langle \nabla^{noise}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| - \left|\left\langle \nabla^{extra}\ell_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \\ &\ge \left(1.99 - 2(\zeta_1+\zeta_2) - \sqrt{8/\pi}\alpha_h^{-1} - \epsilon - (1+\delta)(1/(c_3\alpha_z^l) + 2\sqrt{s}\alpha_h)\right)\|\boldsymbol{h}\|^2, \end{aligned} \tag{41}$$

and

$$\begin{aligned} \|\nabla\ell_{tr}(\boldsymbol{z})\| &\le \left\|\nabla^{clean}\ell_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{noise}\ell_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{extra}\ell_{tr}(\boldsymbol{z})\right\| \\ &\le (1+\delta)\left(2\sqrt{1.02 + 2/\alpha_h} + 1/(c_3\alpha_z^l) + 2\sqrt{s}\alpha_h\right)\|\boldsymbol{h}\|. \end{aligned} \tag{42}$$

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

$$\operatorname{dist}\left(\boldsymbol{z} - \mu\nabla\ell_{tr}(\boldsymbol{z}), \boldsymbol{x}\right) \le c_5\frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}\|}$$

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

6 Proofs for Median-RWF

We first show that $\nabla\mathcal{R}_{tr}(\boldsymbol{z})$ in (20) satisfies the RC for the noise-free case in Section 6.1, and then extend it to the model with only sparse outliers in Section 6.2, thus together with Proposition 2 establishing the global convergence of median-RWF in both cases. Section 6.3 proves Theorem 2 in the presence of both sparse outliers and dense bounded noise.

6.1 Proof of Proposition 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 6. *If $m > c_0 n\log n$, then with probability at least $1 - c_1\exp(-c_2 m)$,*

$$0.5\|\boldsymbol{z}-\boldsymbol{x}\| \le \theta_{0.49}, \theta_{1/2}, \theta_{0.51}\left(\left\{\left||\boldsymbol{a}_i^T\boldsymbol{z}| - |\boldsymbol{a}_i^T\boldsymbol{x}|\right|\right\}_{i=1}^m\right) \le 0.8\|\boldsymbol{z}-\boldsymbol{x}\|, \tag{43}$$

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

Proof. See Appendix D.1. □

Next we give a bound on the left hand side of RC.

Proposition 7 (Adapted version of Proposition 2 of [13]). *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 $1 - c_1 \exp(-c_2 m)$,*

$$\langle \nabla \mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x} \rangle \geq \{0.88 - \zeta_1' - \zeta_2' - \epsilon\} \|\boldsymbol{z} - \boldsymbol{x}\|^2 \tag{44}$$

holds uniformly over all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$ satisfying $\frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{z}\|} \leq \frac{1}{20}$, where $c_0, c_1, c_2 > 0$ are some universal constants, and ζ_1', ζ_2' are given by

$$\zeta_1' := 1 - \min\left\{ \mathbb{E}\left[\xi^2 \mathbf{1}_{\{\xi \geq 0.5\sqrt{1.01}\alpha_h' \frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\}} \right], \mathbb{E}\left[\mathbf{1}_{\{\xi \geq 0.5\sqrt{1.01}\alpha_h' \frac{\|\boldsymbol{z}-\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\}} \right] \right\}$$
$$\zeta_2' := \mathbb{E}\left[\xi^2 \mathbf{1}_{\{|\xi| > 0.5\sqrt{0.99}\alpha_h'\}} \right]$$

for some $\xi \sim \mathcal{N}(0,1)$ and $\alpha_h' = 5$.

Proof. See Appendix D.2. □

Proposition 7 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 8 (Lemma 7 of [13]). *If $m > c_0 n$, then there exist some constants $c_1, c_2 > 0$ such that with probability at least $1 - c_1 \exp(-c_2 m)$,*

$$\|\nabla \mathcal{R}_{tr}(\boldsymbol{z})\| \leq (1.8 + \delta)\|\boldsymbol{z} - \boldsymbol{x}\| \tag{45}$$

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 Appendix D.3. □

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 $\lambda + \mu \cdot (1.8 + \delta)^2 < 2(0.88 - \zeta_1' - \zeta_2' - \epsilon)$.

6.2 Proof of Theorem 1

We consider the model (11) 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 3 yields

$$\theta_{\frac{1}{2}-s}(\{||\boldsymbol{a}_i^T \boldsymbol{x}| - |\boldsymbol{a}_i^T \boldsymbol{z}||\}) \leq \theta_{\frac{1}{2}}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}||\}) \leq \theta_{\frac{1}{2}+s}(\{||\boldsymbol{a}_i^T \boldsymbol{x}| - |\boldsymbol{a}_i^T \boldsymbol{z}||\}). \tag{46}$$

For the simplicity of notation, we let $\boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$. Then for the instance of $s = 0.01$, Proposition 6 yields that if $m > c_0 n \log n$, then

$$0.5\|\boldsymbol{h}\| \leq \theta_{\frac{1}{2}}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}||\}) \leq 0.8\|\boldsymbol{h}\| \tag{47}$$

holds with probability at least $1 - 2\exp(-\Omega(m))$.

To differentiate from $\mathcal{T}^i$, we define $\tilde{\mathcal{T}}^i := \left\{ \left| |\boldsymbol{a}_i^T \boldsymbol{x}| - |\boldsymbol{a}_i^T \boldsymbol{z}| \right| \leq \alpha_h' \mathsf{med}\left\{ \left| \sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}| \right| \right\} \right\}$. 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 (47), 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 7 and 8 are also valid to $\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})$. Thus, one has

$$\begin{aligned}
&\left\langle \nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h} \right\rangle \geq (0.88 - \zeta_1' - \zeta_2' - \epsilon)\, \|\boldsymbol{h}\|^2 \\
&\left\| \nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}) \right\| \leq (1.8+\delta)\|\boldsymbol{h}\|.
\end{aligned}$$

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 $|q_i| \leq 1.6\alpha_h' \|\boldsymbol{h}\|$. Thus, $\|\boldsymbol{q}\| \leq \sqrt{sm} \cdot 1.6\alpha_h' \|\boldsymbol{h}\|$, and

$$\begin{aligned}
\left\| \nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}) \right\| &= \frac{1}{m} \left\| \boldsymbol{A}^T \boldsymbol{q} \right\| \leq 1.6(1+\delta)\sqrt{s}\alpha_h' \|\boldsymbol{h}\|, \\
\left| \left\langle \nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h} \right\rangle \right| &\leq \|\boldsymbol{h}\| \cdot \left\| \nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}) \right\| \leq 1.6(1+\delta)\sqrt{s}\alpha_h' \|\boldsymbol{h}\|^2,
\end{aligned}$$

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

$$\begin{aligned}
\langle \nabla \mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h} \rangle &\geq \left\langle \nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h} \right\rangle - \left| \left\langle \nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h} \right\rangle \right| \\
&\geq \left(0.88 - \zeta_1' - \zeta_2' - \epsilon - 1.6(1+\delta)\sqrt{s}\alpha_h' \right) \|\boldsymbol{h}\|^2,
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla \mathcal{R}_{tr}(\boldsymbol{z})\| &\leq \left\| \nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}) \right\| + \left\| \nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}) \right\| \\
&\leq \left(1.8 + \delta + 1.6(1+\delta)\sqrt{s}\alpha_h' \right) \|\boldsymbol{h}\|.
\end{aligned}$$

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

6.3 Proof of Theorem 2

We consider the model (11) with outliers and bounded noise. We split our analysis of the gradient loop into two regimes.

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

$$\operatorname{dist}\left(\boldsymbol{z} - \mu \nabla \mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{x} \right) \leq (1-\rho)\operatorname{dist}(\boldsymbol{z}, \boldsymbol{x}). \tag{48}$$

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 3, we have

$$\theta_{\frac{1}{2}-s}\left\{ \left| \sqrt{\tilde{y}_i} - |\boldsymbol{a}_i^T \boldsymbol{z}| \right| \right\} \leq \mathsf{med}\left\{ \left| \sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}| \right| \right\} \leq \theta_{\frac{1}{2}+s}\left\{ \left| \sqrt{\tilde{y}_i} - |\boldsymbol{a}_i^T \boldsymbol{z}| \right| \right\}.$$

Moreover, by Lemma 2 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 6, if c_3 is sufficiently large (i.e., $c_3 > 100$), we still have

$$0.5\|\boldsymbol{h}\| \leq \mathsf{med}\left\{ \left| \sqrt{y_i} - |\boldsymbol{a}_i^T \boldsymbol{z}| \right| \right\} \leq 0.8\|\boldsymbol{h}\|. \tag{49}$$

Furthermore, recall $\tilde{\mathcal{T}}^i := \left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right| \le \alpha_h' \mathsf{med}\left\{\left||\boldsymbol{a}_i^T\boldsymbol{z}| - \sqrt{y_i}\right|\right\}\right\}$. 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 (49). All the proof arguments for Propositions 7 and 8 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 $|q_i| \le 1.6\alpha_h'\|\boldsymbol{h}\|$, and $\|\boldsymbol{q}\| \le \sqrt{sm}\cdot 1.6\alpha_h'\|\boldsymbol{h}\|$. We thus have

$$\begin{aligned}
&\left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\| = \frac{1}{m}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\| \le 1.6(1+\delta)\sqrt{s}\alpha_h'\|\boldsymbol{h}\|, \\
&\left|\left\langle\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \le \|\boldsymbol{h}\| \cdot \left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\| \le 1.6(1+\delta)\sqrt{s}\alpha_h'\|\boldsymbol{h}\|^2.
\end{aligned}$$

Putting these together, one has

$$\begin{aligned}
\langle\nabla\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle &\ge \left\langle\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle - \left|\left\langle\nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| - \left|\left\langle\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\right\rangle\right| \\
&\ge \left(0.88 - \zeta_1' - \zeta_2' - \epsilon - (1+\delta)(1/c_3 - 1.6\sqrt{s}\alpha_h')\right)\|\boldsymbol{h}\|^2,
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla\mathcal{R}_{tr}(\boldsymbol{z})\| &\le \left\|\nabla^{clean}\mathcal{R}_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{noise}\mathcal{R}_{tr}(\boldsymbol{z})\right\| + \left\|\nabla^{extra}\mathcal{R}_{tr}(\boldsymbol{z})\right\| \\
&\le \left(1.8 + \delta + (1+\delta)\cdot(1/c_3 + 1.6\sqrt{s}\alpha_h')\right)\|\boldsymbol{h}\|. \qquad (50)
\end{aligned}$$

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 $\|\boldsymbol{h}\| \le c_3\sqrt{\|\boldsymbol{w}\|_\infty}$, 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

$$\operatorname{dist}\left(\boldsymbol{z} - \mu\nabla\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{x}\right) \le c_5\sqrt{\|\boldsymbol{w}\|_\infty} \tag{51}$$

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

7 Conclusions

In this paper, we propose 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 in this paper will be useful for designing provably robust algorithms for other inference problems under sparse corruptions.

Appendix

A Proof of Properties of Median

A.1 Proof of Lemma 1

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 $\mathbb{P}(X \le \theta_p - \epsilon) = p - \delta_1$, where $\delta_1 \in (\epsilon l, \epsilon L)$. Then one has

$$\begin{aligned}\mathbb{P}\left(\hat{\theta}_p < \theta_p - \epsilon\right) &\overset{(a)}{=} \mathbb{P}\left(\sum_{i=1}^m \mathbf{1}_{\{X_i \le \theta_p - \epsilon\}} \ge pm\right) = \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{X_i \le \theta_p - \epsilon\}} \ge (p-\delta_1) + \delta_1\right) \\ &\overset{(b)}{\le} \exp(-2m\delta_1^2) \le \exp(-2m\epsilon^2 l^2),\end{aligned}$$

where (a) is due to the definition of the quantile function in (27) and (b) is due to the fact that $\mathbf{1}_{\{X_i \le \theta_p - \epsilon\}} \sim$ Bernoulli$(p - \delta_1)$ i.i.d., followed by the Hoeffding inequality. Similarly, one can show for some $\delta_2 \in (\epsilon l, \epsilon L)$,

$$\mathbb{P}\left(\hat{\theta}_p > \theta_p + \epsilon\right) \le \exp(-2m\delta_2^2) \le \exp(-2m\epsilon^2 l^2).$$

Combining these two inequalities, one has the conclusion.

A.2 Proof of Lemma 2

It suffices to show that

$$|X_{(k)} - Y_{(k)}| \le \max_l |X_l - Y_l|, \quad \forall k = 1, \cdots, n. \tag{52}$$

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 $X_j \le X_i \le Y_j$ or $Y_i \le Y_j \le X_i$. Hence,

$$|X_{(n)} - Y_{(n)}| = |X_i - Y_j| \le \max\{|X_i - Y_i|, |X_j - Y_j|\}.$$

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.

A.3 Proof of Lemma 3

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.

A.4 Proof of Lemma 4

First we introduce some general facts for the distribution of the product of two correlated standard Gaussian random variables [20]. 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

$$\psi_\rho(x) = \frac{1}{\pi\sqrt{1-\rho^2}} \left[\exp\left(\frac{\rho x}{1-\rho^2}\right) + \exp\left(-\frac{\rho x}{1-\rho^2}\right)\right] K_0\left(\frac{|x|}{1-\rho^2}\right), \quad x > 0, \tag{53}$$

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 (53) 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 (53), due to the continuity of $\psi_\rho(x)$ and $\theta_{1/2}(\psi_\rho)$, we can calculate them numerically, as illustrated in Figure 5. From the numerical calculation, one can see that both $\psi_\rho(\theta_{1/2})$ and $\theta_{1/2}(\psi_\rho)$ 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

$$0.348 < \theta_{1/2}(\psi_\rho) < 0.455, \quad 0.47 < \psi_\rho(\theta_{1/2}) < 0.76. \tag{54}$$

B Proof of Proposition 2

Denote $\tilde{y}_i := |\boldsymbol{a}_i^T \boldsymbol{x}|^2 + w_i$ for convenience. We first bound the concentration of $\mathsf{med}(\{y_i\})$, also denoted by $\theta_{\frac{1}{2}}(\{y_i\})$. Lemma 3 yields

$$\theta_{\frac{1}{2}-s}(\{\tilde{y}_i\}) < \theta_{\frac{1}{2}}(\{y_i\}) < \theta_{\frac{1}{2}+s}(\{\tilde{y}_i\}). \tag{55}$$

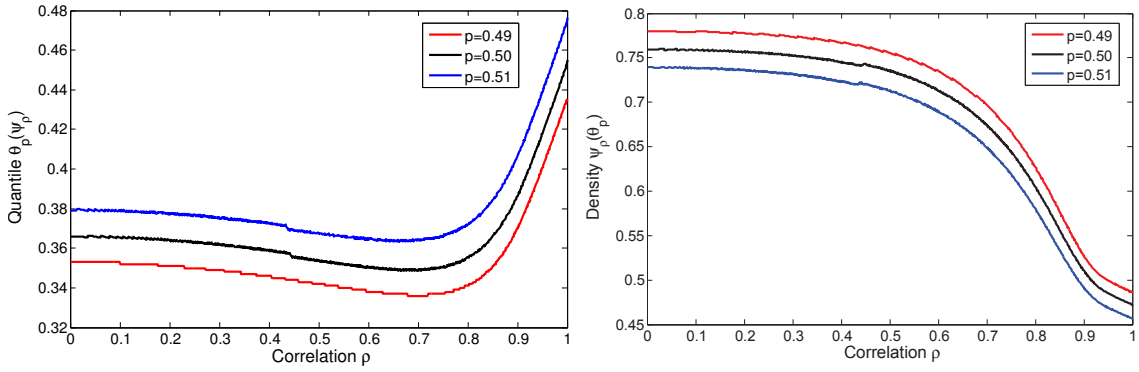


Figure 5: Quantiles and density at quantiles of $\psi_\rho(x)$ across ρ.

Moreover, Lemma 2 indicates that

$$\theta_{\frac{1}{2}-s}(\{\tilde{y}_i\}) \geq \theta_{\frac{1}{2}-s}(\{|\boldsymbol{a}_i^T\boldsymbol{x}|^2\}) - \|\boldsymbol{w}\|_\infty, \tag{56}$$

$$\theta_{\frac{1}{2}+s}(\{\tilde{y}_i\}) \leq \theta_{\frac{1}{2}+s}(\{|\boldsymbol{a}_i^T\boldsymbol{x}|^2\}) + \|\boldsymbol{w}\|_\infty. \tag{57}$$

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 1, 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 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

$$\left(\theta_{\frac{1}{2}-s}(K) - \epsilon - c\right)\|\boldsymbol{x}\|^2 < \theta_{\frac{1}{2}}(\{y_i\}) < \left(\theta_{\frac{1}{2}+s}(K) + \epsilon + c\right)\|\boldsymbol{x}\|^2, \tag{58}$$

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 (58) holds, the truncation function has the following bounds,

$$\begin{aligned}
\mathbf{1}_{\{y_i \leq \alpha_y^2\theta_{1/2}(\{y_i\})/0.455\}} &\leq \mathbf{1}_{\left\{y_i \leq \alpha_y^2(\zeta^s+\epsilon)/0.455\right\}} \leq \mathbf{1}_{\left\{(\boldsymbol{a}_i^T\boldsymbol{x})^2 \leq \alpha_y^2(\zeta^s+\epsilon+c)/0.455\right\}} \\
\mathbf{1}_{\{y_i \leq \alpha_y^2\theta_{1/2}(\{y_i\})/0.455\}} &\geq \mathbf{1}_{\left\{y_i \leq \alpha_y^2(\zeta_s-\epsilon)/0.455\right\}} \geq \mathbf{1}_{\left\{(\boldsymbol{a}_i^T\boldsymbol{x})^2 \leq \alpha_y^2(\zeta_s-\epsilon-c)/0.455\right\}}.
\end{aligned}$$

On the other hand, denote the support of the outliers as S, and we have

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T\tilde{y}_i\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})^2 \leq \alpha_y^2\theta_{1/2}(\{y_i\})/0.455\}} + \frac{1}{m}\sum_{i\in S}\boldsymbol{a}_i\boldsymbol{a}_i^T y_i\mathbf{1}_{\{y_i \leq \alpha_y^2\theta_{1/2}(\{y_i\})/0.455\}}.$$

Consequently, one can bound $\boldsymbol{Y}$ as

$$\begin{aligned}
\boldsymbol{Y}_1 :=& \frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T(\boldsymbol{a}_i^T\boldsymbol{x})^2\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})^2 \leq \alpha_y^2(\zeta_s-\epsilon-c)/0.455\}} - c\cdot\frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T \preceq \boldsymbol{Y} \\
\preceq& \frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T(\boldsymbol{a}_i^T\boldsymbol{x})^2\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})^2 \leq \alpha_y^2(\zeta^s+\epsilon+c)/0.455\}} + c\cdot\frac{1}{m}\sum_{i\notin S}\boldsymbol{a}_i\boldsymbol{a}_i^T + \frac{1}{m}\sum_{i\in S}\boldsymbol{a}_i\boldsymbol{a}_i^T\alpha_y^2(\zeta^s+\epsilon+c)/0.455 =: \boldsymbol{Y}_2,
\end{aligned}$$

where we have

$$\mathbb{E}[\boldsymbol{Y}_1] = (1-s)(\beta_1\boldsymbol{x}\boldsymbol{x}^T + \beta_2\boldsymbol{I} - c\boldsymbol{I}), \quad \mathbb{E}[\boldsymbol{Y}_2] = (1-s)(\beta_3\boldsymbol{x}\boldsymbol{x}^T + \beta_4\boldsymbol{I} + c\boldsymbol{I}) + s\alpha_y^2\frac{(\zeta^s+\epsilon)}{0.455}\boldsymbol{I}, \tag{59}$$

with

$$\beta_1 := \mathbb{E}\left[\xi^4 \mathbf{1}_{\left\{|\xi|\le \alpha_y\sqrt{(\zeta_s-\epsilon-c)/0.455}\right\}}\right] - \mathbb{E}\left[\xi^2 \mathbf{1}_{\left\{|\xi|\le \alpha_y\sqrt{(\zeta_s-\epsilon-c)/0.455}\right\}}\right]$$
$$\beta_2 := \mathbb{E}\left[\xi^2 \mathbf{1}_{\left\{|\xi|\le \alpha_y\sqrt{(\zeta_s-\epsilon-c)/0.455}\right\}}\right]$$
$$\beta_3 := \mathbb{E}\left[\xi^4 \mathbf{1}_{\left\{|\xi|\le \alpha_y\sqrt{(\zeta^s+\epsilon+c)/0.455}\right\}}\right] - \mathbb{E}\left[\xi^2 \mathbf{1}_{\left\{|\xi|\le \alpha_y\sqrt{(\zeta^s+\epsilon+c)/0.455}\right\}}\right]$$
$$\beta_4 := \mathbb{E}\left[\xi^2 \mathbf{1}_{\left\{|\xi|\le \alpha_y\sqrt{(\zeta^s+\epsilon+c)/0.455}\right\}}\right]$$

where $\xi \sim \mathcal{N}(0,1)$.

Applying standard results on random matrices with non-isotropic sub-Gaussian rows [48, 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

$$\|\boldsymbol{Y}_1 - \mathbb{E}[\boldsymbol{Y}_1]\| \le \delta, \quad \|\boldsymbol{Y}_2 - \mathbb{E}[\boldsymbol{Y}_2]\| \le \delta \tag{60}$$

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 $\|\mathbb{E}[\boldsymbol{Y}_1] - \mathbb{E}[\boldsymbol{Y}_2]\| \le \delta$. Putting these together, one has

$$\|\boldsymbol{Y} - (1-s)(\beta_1\boldsymbol{x}\boldsymbol{x}^T + \beta_2\boldsymbol{I} - c\boldsymbol{I})\| \le 3\delta. \tag{61}$$

Let $\tilde{\boldsymbol{z}}^{(0)}$ be the normalized leading eigenvector of $\boldsymbol{Y}$. Repeating the same argument as in [9, Section 7.8] and taking δ, ϵ to be sufficiently small, one has

$$\mathrm{dist}(\tilde{\boldsymbol{z}}^{(0)}, \boldsymbol{x}) \le \tilde{\delta}, \tag{62}$$

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

Furthermore let $\boldsymbol{z}^{(0)} = \sqrt{\mathsf{med}\{y_i\}/0.455}\tilde{\boldsymbol{z}}^{(0)}$ to handle cases $\|\boldsymbol{x}\| \ne 1$. By the bound (58), one has

$$\left|\frac{\mathsf{med}(\{y_i\})}{0.455} - \|\boldsymbol{x}\|^2\right| \le \max\left\{\left|\frac{\zeta_s-\epsilon-c}{0.455} - 1\right|, \left|\frac{\zeta^s+\epsilon+c}{0.455} - 1\right|\right\}\|\boldsymbol{x}\|^2 \le \frac{\zeta^s - \zeta_s + 2\epsilon + 2c}{0.455}\|\boldsymbol{x}\|^2. \tag{63}$$

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. □

C Supporting Proofs for median-TWF

C.1 Proof of Proposition 3

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 $\boldsymbol{x}$, (32) 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, we have

$$0.47 < \psi_\rho(\theta_{1/2}) < 0.76, \quad 0.348 < \theta_{1/2}(\psi_\rho) < 0.455. \tag{64}$$

By Lemma 1, we have with probability at least $1 - 2\exp(-cm\epsilon^2)$ (here c is around 2×0.47^2),

$$0.348 - \epsilon < \mathsf{med}(\{r_i'\}_{i=1}^m) < 0.455 + \epsilon. \tag{65}$$

The same arguments carry over to other quantiles $\theta_{0.49}(\{r_i'\})$ and $\theta_{0.51}(\{r_i'\})$. From Figure. 5, we observe that for $\rho \in [0,1]$

$$0.45 < \psi_\rho(\theta_{0.49}), \psi_\rho(\theta_{0.51}) < 0.78, \quad 0.336 < \theta_{0.49}(\psi_\rho), \theta_{0.51}(\psi_\rho) < 0.477 \tag{66}$$

and then we have with probability at least $1 - 2\exp(-cm\epsilon^2)$ (here c is around 2×0.45^2),

$$0.336 - \epsilon < \theta_{0.49}(\{r_m'\}), \theta_{0.51}(\{r_m'\}) < 0.477 + \epsilon. \tag{67}$$

Hence, by multiplying by $\sqrt{\mathsf{Var}(\tilde{v})}$, we have with probability $1 - 2\exp(-cm\epsilon^2)$,

$$(0.65 - \epsilon)\|\boldsymbol{z} - \boldsymbol{x}\|\|\boldsymbol{z}\| \le \mathsf{med}\left(\left\{|(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2|\right\}\right) \le (0.91 + \epsilon)\|\boldsymbol{z} - \boldsymbol{x}\|\|\boldsymbol{z}\|, \tag{68}$$

$$(0.63 - \epsilon)\|\boldsymbol{z} - \boldsymbol{x}\|\|\boldsymbol{z}\| \le \theta_{0.49}, \theta_{0.51}\left(\left\{|(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2|\right\}\right) \le (0.95 + \epsilon)\|\boldsymbol{z} - \boldsymbol{x}\|\|\boldsymbol{z}\|. \tag{69}$$

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 (32) 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 (68) as

$$(0.65 - \epsilon) \le \mathsf{med}\left(\left\{\left|\left(\frac{2(\boldsymbol{a}_i^T\boldsymbol{z})}{\|\boldsymbol{z}\|} - \frac{\boldsymbol{a}_i^T\boldsymbol{h}}{\|\boldsymbol{h}\|}\frac{\|\boldsymbol{h}\|}{\|\boldsymbol{z}\|}\right)\frac{\boldsymbol{a}_i^T\boldsymbol{h}}{\|\boldsymbol{h}\|}\right|\right\}\right) \le (0.91 + \epsilon) \tag{70}$$

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

$$\mathcal{N}_\tau = \{(\boldsymbol{z}_0, \boldsymbol{h}_0, t_0) : (\boldsymbol{z}_0, \boldsymbol{h}_0, t_0) \in \mathcal{S}_\tau \times \mathcal{S}_\tau \times \mathcal{L}_\tau\}. \tag{71}$$

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

$$(0.65-\epsilon) \le \mathsf{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}_0) - (\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0||\boldsymbol{a}_i^T\boldsymbol{h}_0|\right\}\right) \le (0.91+\epsilon), \quad \forall (\boldsymbol{z}_0, \boldsymbol{h}_0, t_0) \in \mathcal{N}_\epsilon \tag{72}$$

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

We next argue that (72) holds with probability $1-c_1\exp(-c_2 m\epsilon^2)$ for some constants c_1, c_2 as long as $m \ge c_0(\epsilon^{-2}\log\epsilon^{-1})n\log n$ 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 $m \ge c_0(\epsilon^{-2}\log\epsilon^{-1})n\log n$, then

$$2n\log m < c' m\epsilon^2. \tag{73}$$

For any fixed n, if (73) holds for some m and $m > (2/c')\epsilon^{-2}n$, then (73) always holds for larger m, because

$$\begin{aligned} 2n\log(m+1) &= 2n\log m + 2n(\log(m+1) - \log m) = 2n\log m + \frac{2n}{m}\log(1+\frac{1}{m})^m \\ &\le 2n\log m + \frac{2n}{m} \le c' m\epsilon^2 + c'\epsilon^2 = c'(m+1)\epsilon^2. \end{aligned}$$

Next, for any n, we can always find a constant c_0 such that (73) holds for $m = c_0(\epsilon^{-2}\log\epsilon^{-1})n\log n$. Such c_0 can be easily found for large n. For example, $c_0 = 4/c'$ is a valid option if

$$(4/c')(\epsilon^{-2}\log\epsilon^{-1})n\log n < n^2. \tag{74}$$

Moreover, since the number of n that violates (74) is finite, the maximum over all such c_0 serves the purpose.

Next, one needs to bound

$$\left|\mathsf{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}_0) - (\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0||\boldsymbol{a}_i^T\boldsymbol{h}_0|\right\}\right) - \mathsf{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{h})t||\boldsymbol{a}_i^T\boldsymbol{h}|\right\}\right)\right|$$

for any $\|\boldsymbol{z}-\boldsymbol{z}_0\| < \tau, \|\boldsymbol{z}-\boldsymbol{z}_0\| < \tau$ and $\|t-t_0\| < \tau$.

By Lemma 2 and the inequality $\big||x|-|y|\big| \le |x-y|$, we have

$$\begin{aligned} &\left|\mathsf{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}_0) - (\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0||\boldsymbol{a}_i^T\boldsymbol{h}_0|\right\}\right) - \mathsf{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{h})t||\boldsymbol{a}_i^T\boldsymbol{h}|\right\}\right)\right| \\ &\le \max_{i\in[m]} \left|\left(2(\boldsymbol{a}_i^T\boldsymbol{z}_0) - (\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0\right)(\boldsymbol{a}_i^T\boldsymbol{h}_0) - \left(2(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{h})t\right)(\boldsymbol{a}_i^T\boldsymbol{h})\right| \\ &\le \max_{i\in[m]} \left|\left(2(\boldsymbol{a}_i^T\boldsymbol{z}_0) - (\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0\right)(\boldsymbol{a}_i^T\boldsymbol{h}_0) - \left(2(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{h})t\right)(\boldsymbol{a}_i^T\boldsymbol{h}_0)\right| \\ &\quad + \max_{i\in[m]} \left|\left(2(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{h})t\right)(\boldsymbol{a}_i^T\boldsymbol{h}_0) - \left(2(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{h})t\right)(\boldsymbol{a}_i^T\boldsymbol{h})\right| \\ &\le \max_{i\in[m]} \left(\left|2\boldsymbol{a}_i^T(\boldsymbol{z}_0-\boldsymbol{z})\right| + \left|(\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0 - (\boldsymbol{a}_i^T\boldsymbol{h})t\right|\right)\left|\boldsymbol{a}_i^T\boldsymbol{h}_0\right| + \max_{i\in[m]}\left|2(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{h})t\right|\left|\boldsymbol{a}_i^T(\boldsymbol{h}_0-\boldsymbol{h})\right| \\ &\le \max_{i\in[m]} \|\boldsymbol{a}_i\|^2(3+t)\tau + \max_{i\in[m]}\|\boldsymbol{a}_i\|^2(2+t)\tau \\ &\le \max_{i\in[m]} \|\boldsymbol{a}_i\|^2(5+2t)\tau \end{aligned}$$

On the event $E_1 := \left\{\max_{i\in[m]}\|\boldsymbol{a}_i\|^2 \le m+n\right\}$, one can show that

$$\left|\mathsf{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}_0) - (\boldsymbol{a}_i^T\boldsymbol{h}_0)t_0||\boldsymbol{a}_i^T\boldsymbol{h}_0|\right\}\right) - \mathsf{med}\left(\left\{|2(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{h})t||\boldsymbol{a}_i^T\boldsymbol{h}|\right\}\right)\right| < 6(m+n)\tau < \epsilon. \tag{75}$$

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 [48, Corollary 5.17] and observing that the sub-exponential norm of χ_1^2 is smaller than 2, we have

$$\mathbb{P}\left\{\|\boldsymbol{a}_i\|^2 \geq m + n\right\} \leq \exp(-m/8). \tag{76}$$

Then a union bound concludes the claim.

Further note that (72) holds on an event E_2, which has probability $1 - c_1\exp(-c_2 m\epsilon^2)$ 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 (32) holds. Such net covering arguments can also carry over to show that inequalities of $\theta_{0.49}$ and $\theta_{0.51}$ in (32) also hold for all $\boldsymbol{x}$ and $\boldsymbol{z}$ obeying $\|\boldsymbol{x} - \boldsymbol{z}\| \leq \frac{1}{11}\|\boldsymbol{z}\|$.

C.2 Proof of Proposition 4

The proof adapts that of [13, 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

$$\begin{aligned}\nabla\ell_{tr}(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^m \frac{(\boldsymbol{a}_i^T\boldsymbol{z})^2 - (\boldsymbol{a}_i^T\boldsymbol{x})^2}{\boldsymbol{a}_i^T\boldsymbol{z}}\boldsymbol{a}_i\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i} \\ &= \frac{1}{m}\sum_{i=1}^m 2(\boldsymbol{a}_i^T\boldsymbol{h})\boldsymbol{a}_i\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i} - \frac{1}{m}\sum_{i=1}^m \frac{(\boldsymbol{a}_i^T\boldsymbol{h})^2}{\boldsymbol{a}_i^T\boldsymbol{z}}\boldsymbol{a}_i\mathbf{1}_{\mathcal{E}_1^i\cap\mathcal{E}_2^i}. \end{aligned}\tag{77}$$

One expects the contribution of the second term in (77) 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 3 holds, the following inclusion property

$$\mathcal{E}_3^i \subseteq \mathcal{E}_2^i \subseteq \mathcal{E}_4^i \tag{78}$$

is true for all i, where $\mathcal{E}_2^i$ is defined in Algorithm 1. 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 5 (Lemma 3 in [13]). *For any $\gamma > 0$, define*

$$\mathcal{D}_\gamma^i := \{\left|(\boldsymbol{a}_i^T\boldsymbol{x})^2 - (\boldsymbol{a}_i^T\boldsymbol{z})^2\right| \leq \gamma\|\boldsymbol{h}\|\|\boldsymbol{a}_i^T\boldsymbol{z}|\}, \tag{79}$$

$$\mathcal{D}_\gamma^{i,1} := \left\{\frac{|\boldsymbol{a}_i^T\boldsymbol{h}|}{\|\boldsymbol{h}\|} \leq \gamma\right\}, \tag{80}$$

$$\mathcal{D}_\gamma^{i,2} := \left\{\left|\frac{\boldsymbol{a}_i^T\boldsymbol{h}}{\|\boldsymbol{h}\|} - \frac{2\boldsymbol{a}_i^T\boldsymbol{z}}{\|\boldsymbol{h}\|}\right| \leq \gamma\right\}. \tag{81}$$

On the event $\mathcal{E}_1^i$ defined in Algorithm 1, 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

$$\left(\mathcal{D}_{\frac{\gamma}{1+\sqrt{2}}}^{i,1}\cap\mathcal{E}_1^i\right)\cup\left(\mathcal{D}_{\frac{\gamma}{1+\sqrt{2}}}^{i,2}\cap\mathcal{E}_1^i\right) \subseteq \mathcal{D}_\gamma^i\cap\mathcal{E}_1^i \subseteq \left(\mathcal{D}_\gamma^{i,1}\cap\mathcal{E}_1^i\right)\cup\left(\mathcal{D}_\gamma^{i,2}\cap\mathcal{E}_1^i\right). \tag{82}$$

Specifically, following the two inclusion properties (78) and (82), we have

$$\mathcal{D}^{i,1}_{\gamma_3} \cap \mathcal{E}^i_{1,\gamma_3} \subseteq \mathcal{E}^i_3 \cap \mathcal{E}^i_1 \subseteq \mathcal{E}^i_2 \cap \mathcal{E}^i_1 \subseteq \mathcal{E}^i_4 \cap \mathcal{E}^i_1 \subseteq (\mathcal{D}^{i,1}_{\gamma_4} \cup \mathcal{D}^{i,2}_{\gamma_4}) \cap \mathcal{E}^i_1 \tag{83}$$

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 (77), we have the following lower bound

$$\langle \nabla \ell_{tr}(\boldsymbol{z}), \boldsymbol{h} \rangle \geq \frac{2}{m}\sum_{i=1}^m (\boldsymbol{a}_i^T\boldsymbol{h})^2 \mathbf{1}_{\mathcal{E}^i_1 \cap \mathcal{D}^{i,1}_{\gamma_3}} - \frac{1}{m}\sum_{i=1}^m \frac{|\boldsymbol{a}_i^T\boldsymbol{h}|^3}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\mathbf{1}_{\mathcal{D}^{i,1}_{\gamma_4} \cap \mathcal{E}^i_1} - \frac{1}{m}\sum_{i=1}^m \frac{|\boldsymbol{a}_i^T\boldsymbol{h}|^3}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\mathbf{1}_{\mathcal{D}^{i,2}_{\gamma_4} \cap \mathcal{E}^i_1}. \tag{84}$$

The three terms in (84) can be bounded following Lemmas 4, 5, and 6 in [13], which concludes the proof.

D Supporting Proofs for Median-RWF

D.1 Proof of Proposition 6

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, $\mathsf{med}\left(\left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}_{i=1}^m\right)$ can be viewed as $\mathsf{med}(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}_{i=1}^m)$ with a small perturbation.

Lemma 6. *If $m > c_0 n$, then with probability at least $1 - c_1\exp(-c_2 m)$,*

$$\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<0\}} < 0.05 \tag{85}$$

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

Proof. See Appendix D.4. □

By Lemma 3 and Lemma 6, we have

$$\theta_{p-0.05}\left(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}\right) \leq \theta_p\left(\left\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\right\}\right) \leq \theta_{p+0.05}\left(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}\right) \tag{86}$$

for all $\boldsymbol{x}$ and $\boldsymbol{z}$ satisfying $\|\boldsymbol{x} - \boldsymbol{z}\| \leq \frac{1}{11}\|\boldsymbol{z}\|$ with high probability.

For the model (2) with a fraction s of outliers, due to Lemma 3, we have that

$$\theta_{\frac{1}{2}-s}(\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\}) \leq \theta_{\frac{1}{2}}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}||\}) \leq \theta_{\frac{1}{2}+s}(\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\}). \tag{87}$$

Combining with (86), we obtain that

$$\theta_{0.45-s}(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}) \leq \theta_{\frac{1}{2}}(\{|\sqrt{y_i} - |\boldsymbol{a}_i^T\boldsymbol{z}||\}) \leq \theta_{0.55+s}(\{|\boldsymbol{a}_i^T\boldsymbol{h}|\}). \tag{88}$$

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

$$\phi(\theta_{0.44}) = 0.67, \quad \phi(\theta_{0.45}) = 0.67, \quad \phi(\theta_{0.55}) = 0.60, \quad \phi(\theta_{0.56}) = 0.59 \tag{89}$$

$$\theta_{0.44}(\phi) = 0.58, \quad \theta_{0.45}(\phi) = 0.6, \quad \theta_{0.55}(\phi) = 0.76, \quad \theta_{0.56}(\phi) = 0.78. \tag{90}$$

By Lemma 1, the sample quantiles concentrate on population quantiles. Thus, for any fixed pair $(\boldsymbol{x}, \boldsymbol{z})$,

$$(0.6 - \epsilon)\|\boldsymbol{h}\| \leq \theta_{1/2}(\{\left||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\right|\}_{i=1}^m) \leq (0.76 + \epsilon)\|\boldsymbol{h}\|, \tag{91}$$

holds with probability at least $1 - 2\exp(-cm\epsilon^{-2})$.

Following the argument of net covering similarly to that in Appendix C.1, the proposition is proved.

D.2 Proof of Proposition 7

The proof adapts the proof of Proposition 2 in [13]. We outline the main steps for completeness. Observe that for the noise-free case, $y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|$. We obtain

$$\nabla \mathcal{R}_{tr}(\boldsymbol{z}) = \frac{1}{m}\sum_{i=1}^{m}\left((\boldsymbol{a}_i^T\boldsymbol{z}) - |\boldsymbol{a}_i^T\boldsymbol{x}|\cdot\frac{\boldsymbol{a}_i^T\boldsymbol{z}}{|\boldsymbol{a}_i^T\boldsymbol{z}|}\right)\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i} = \frac{1}{m}\sum_{i\notin\mathcal{B}}(\boldsymbol{a}_i^T\boldsymbol{h})\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i} + \frac{1}{m}\sum_{i\in\mathcal{B}}(\boldsymbol{a}_i^T\boldsymbol{z}+\boldsymbol{a}_i^T\boldsymbol{x})\boldsymbol{a}_i\mathbf{1}_{\mathcal{T}^i}, \quad (92)$$

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 (92) 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

$$\mathcal{T}_1^i := \{\big||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\big| \le 0.5\alpha_h'\|\boldsymbol{h}\|\}, \quad (93)$$

$$\mathcal{T}_2^i := \{\big||\boldsymbol{a}_i^T\boldsymbol{x}| - |\boldsymbol{a}_i^T\boldsymbol{z}|\big| \le 0.8\alpha_h'\|\boldsymbol{h}\|\}. \quad (94)$$

Due to Proposition 6, the following inclusion property

$$\mathcal{T}_1^i \subseteq \mathcal{T}^i \subseteq \mathcal{T}_2^i \quad (95)$$

holds for all i, where $\mathcal{T}^i$ is defined in Algorithm 2. 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 (95), we have

$$\langle \nabla\mathcal{R}_{tr}(\boldsymbol{z}), \boldsymbol{h}\rangle \ge \frac{1}{m}\sum_{i\notin\mathcal{B}}(\boldsymbol{a}_i^T\boldsymbol{h})^2\mathbf{1}_{\mathcal{T}_1^i} - \frac{1}{m}\sum_{i\in\mathcal{B}}|\boldsymbol{a}_i^T\boldsymbol{z}+\boldsymbol{a}_i^T\boldsymbol{x}|\cdot|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\mathcal{T}_2^i}. \quad (96)$$

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 (96) 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 [13] gives that if $m > c_0 n$, with probability at least $1 - c_1\exp(-c_2 m\epsilon^2)$

$$\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{a}_i^T\boldsymbol{h})^2\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\le\gamma_1\|\boldsymbol{h}\|\}}\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|\ge\gamma_1\|\boldsymbol{h}\|\}} \ge (1-\zeta_1'-\zeta_2'-\epsilon)\|\boldsymbol{h}\|^2 \quad (97)$$

holds for all $\boldsymbol{h}\in\mathbb{R}^n$, where $\zeta_1' := 1 - \min\left\{\mathbb{E}\left[\xi^2\mathbf{1}_{\{\xi\ge\sqrt{1.01}\gamma_1\frac{\|\boldsymbol{h}\|}{\|\boldsymbol{x}\|}\}}\right], \mathbb{E}\left[\mathbf{1}_{\{\xi\ge\sqrt{1.01}\gamma_1\frac{\|\boldsymbol{h}\|}{\|\boldsymbol{x}\|}\}}\right]\right\}$ and $\zeta_2' := \mathbb{E}\left[\xi^2\mathbf{1}_{\{|\xi|>\sqrt{0.99}\gamma_1\}}\right]$ for $\xi\sim\mathcal{N}(0,1)$.

For the second term, we have

$$\frac{1}{m}\sum_{i\in\mathcal{B}}|\boldsymbol{a}_i^T\boldsymbol{z}+\boldsymbol{a}_i^T\boldsymbol{x}|\cdot|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\mathcal{T}_2^i} \le \gamma_2\|\boldsymbol{h}\|\frac{1}{m}\sum_{i\in\mathcal{B}}|\boldsymbol{a}_i^T\boldsymbol{h}| \le \gamma_2\|\boldsymbol{h}\|\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}}, \quad (98)$$

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

$$\frac{1}{m}\sum_{i=1}^{m} |\boldsymbol{a}_i^T\boldsymbol{h}| \cdot \mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}} \le (0.12+\epsilon)\,\|\boldsymbol{h}\| \tag{99}$$

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 Appendix D.5. □

Thus, putting together (97), (98) and Lemma 7 concludes the proof.

D.3 Proof of Proposition 8

This proof adapts the proof of Lemma 7 in [13]. 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

$$\|\nabla\mathcal{R}_{tr}(\boldsymbol{z})\| = \left\|\frac{1}{m}\boldsymbol{A}^T\boldsymbol{v}\right\| \le \frac{1}{m}\|\boldsymbol{A}\|\cdot\|\boldsymbol{v}\| \le (1+\delta)\frac{\|\boldsymbol{v}\|}{\sqrt{m}}$$

where the last inequality is due to the spectral norm bound $\|\boldsymbol{A}\| \le \sqrt{m}(1+\delta)$ following from [48, 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 $\|\boldsymbol{v}\| \le \|\boldsymbol{v}^{(1)}\| + \|\boldsymbol{v}^{(2)}\|$. Furthermore, given $m > c_0 n$, by [10, Lemma 3.1] with probability $1-\exp(-cm)$, we have

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

By Lemma 6, we have with probability $1 - C\exp(-c_1 m)$

$$\frac{1}{m}\|\boldsymbol{v}^{(2)}\|^2 \le (0.8\alpha_h'\|\boldsymbol{h}\|)^2\cdot\left(\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<0\}}\right) \le 0.8\|\boldsymbol{h}\|^2$$

holds, where the last inequality is due to Lemma 6. Hence,

$$\frac{\|\boldsymbol{v}\|}{\sqrt{m}} \le \left(\sqrt{1+\delta}+\sqrt{0.8}\right)\|\boldsymbol{h}\|.$$

This concludes the proof.

D.4 Proof of Lemma 6

Denote correlation $\rho := \frac{\boldsymbol{z}^T\boldsymbol{x}}{\|\boldsymbol{z}\|\|\boldsymbol{x}\|}$. Under the condition $\|\boldsymbol{z}-\boldsymbol{x}\| \le \frac{1}{11}\|\boldsymbol{x}\|$, simple calculation yields $0.995 < \rho \le 1$. 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

$$\chi(t) := \begin{cases} 1, & \text{if } t < 0; \\ -\frac{1}{\delta}\cdot t + 1, & \text{if } 0 \le t \le \delta; \\ 0, & \text{else;} \end{cases}$$

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

$$\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<0\}} \le \chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) \le \mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}. \tag{100}$$

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.

$$\mathbb{E}[\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}] = \mathbb{P}\left\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}) < \delta\right\} = \iint_{\tau_1\cdot\tau_2<\delta} f(\tau_1,\tau_2)d\tau_1 d\tau_2, \tag{101}$$

where $f(\tau_1,\tau_2)$ is the density of the jointly Gaussian random variables $(b_i, \tilde{b}_i)$. Note that $\mathbb{E}[\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}]$ is decreasing on ρ and for the case $\rho = 0.995$ we calculate $\mathbb{E}[\mathbf{1}_{\{(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})<\delta\}}] = 0.045$ numerically. This implies that

$$\mathbb{E}[\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right)] \le 0.045$$

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 [48], we have

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m \chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) > (0.045+\epsilon)\right] < \exp(-cm\epsilon^2), \tag{102}$$

for some universal constant c, as long as $\rho \ge 0.995$.

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

$$\mathcal{N}_\epsilon = \{(\boldsymbol{x}_0, \boldsymbol{z}_0) : (\boldsymbol{x}_0, \boldsymbol{z}_0) \in \mathcal{N}'_\epsilon \times \mathcal{N}'_\epsilon\}. \tag{103}$$

One has cardinality bound (i.e., the upper bound on the covering number) $|\mathcal{N}_\epsilon| \le (1+2/\epsilon)^{2n}$. 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 $\|\boldsymbol{x}-\boldsymbol{x}_0\| \le \epsilon$ and $\|\boldsymbol{z}-\boldsymbol{z}_0\| \le \epsilon$. Taking the union bound for all the points on the net, we claim that

$$\frac{1}{m}\sum_{i=1}^m \chi\left((\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right) \le 0.045+\epsilon, \quad \forall(\boldsymbol{x}_0,\boldsymbol{z}_0) \in \mathcal{N}_\epsilon \tag{104}$$

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

$$\left|\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) - \chi\left((\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right)\right| \le \frac{1}{\delta}\left|(\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z}) - (\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right|. \tag{105}$$

Moreover, by [13, Lemma 1],

$$\frac{1}{m}\|\mathcal{A}(\boldsymbol{M})\|_1 \le c_2\|\boldsymbol{M}\|_F, \qquad \text{for all symmetric rank-2 matrices } \boldsymbol{M} \in \mathbb{R}^{n\times n}, \tag{106}$$

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 (106) holds, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^m \chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) - \frac{1}{m}\sum_{i=1}^m \chi\left((\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right)\right| \\
&\le \frac{1}{m}\sum_{i=1}^m \left|\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right) - \chi\left((\boldsymbol{a}_i^T\boldsymbol{x}_0)(\boldsymbol{a}_i^T\boldsymbol{z}_0)\right)\right| \\
&\le \frac{1}{\delta}\cdot\frac{1}{m}\|\mathcal{A}(\boldsymbol{x}\boldsymbol{z}^T - \boldsymbol{x}_0\boldsymbol{z}_0^T)\|_1 \qquad \text{due to (105)} \\
&\le \frac{1}{\delta}\cdot c_2\|\boldsymbol{x}\boldsymbol{z}^T - \boldsymbol{x}_0\boldsymbol{z}_0^T\|_F \qquad \text{due to (106)} \\
&\le \frac{1}{\delta}\cdot c_2(\|\boldsymbol{x}-\boldsymbol{x}_0\|\cdot\|\boldsymbol{z}\| + \|\boldsymbol{z}-\boldsymbol{z}_0\|\cdot\|\boldsymbol{x}_0\|) \le 2c_3\epsilon/\delta.
\end{aligned}$$

On the intersection of events that (104) and (106) hold, we have

$$\frac{1}{m}\sum_{i=1}^{m}\chi\left((\boldsymbol{a}_i^T\boldsymbol{x})(\boldsymbol{a}_i^T\boldsymbol{z})\right)\le(0.045+\epsilon+2c_3\epsilon/\delta)\,, \tag{107}$$

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.

D.5 Proof of Lemma 7

We first observe that for any γ,

$$\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<|\boldsymbol{a}_i^T\boldsymbol{h}|\}}\le\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma\|\boldsymbol{x}\|\}}+\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\ge\gamma\|\boldsymbol{x}\|\}}\le\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma\|\boldsymbol{x}\|\}}+\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|\ge 20\gamma\|\boldsymbol{h}\|\}} \tag{108}$$

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

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma\|\boldsymbol{x}\|\}}\le(0.11+\epsilon)\|\boldsymbol{h}\| \tag{109}$$

for all $\boldsymbol{x},\boldsymbol{h}\in\mathbb{R}^n$, and

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'\|\boldsymbol{h}\|\}}\le(0.01+\epsilon)\|\boldsymbol{h}\| \tag{110}$$

for all $\boldsymbol{h}\in\mathbb{R}^n$.

We first prove (109). 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

$$\chi_x(t):=\begin{cases}1, & \text{if } |t|<\gamma;\\ \frac{1}{\delta}(\gamma-|t|)+1, & \text{if } \gamma\le|t|\le\gamma+\delta;\\ 0, & \text{else.}\end{cases}$$

Then $\chi_x(t)$ is a Lipschitz function with constant $\frac{1}{\delta}$. We further have

$$\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma\}}\le\chi_x(\boldsymbol{a}_i^T\boldsymbol{x})\le\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}. \tag{111}$$

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 $|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}$,

$$\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]=\iint_{-\infty}^{\infty}|\tau_1|\mathbf{1}_{\{|\tau_2|<\gamma+\delta\}}\cdot f(\tau_1,\tau_2)d\tau_1d\tau_2, \tag{112}$$

where $f(\tau_1,\tau_2)$ is the density of two jointly Gaussian random variables with correlation $\rho=\frac{\boldsymbol{h}^T\boldsymbol{x}}{\|\boldsymbol{h}\|\|\boldsymbol{x}\|}\ne\pm1$. We then continue to derive

$$\begin{aligned}\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}] &= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty}|\tau_1|\exp\left(-\frac{\tau_1^2}{2}\right)\cdot\int_{-(\gamma+\delta)}^{\gamma+\delta}\exp\left(-\frac{(\tau_2-\rho\tau_1)^2}{2(1-\rho^2)}\right)d\tau_2d\tau_1 &(113)\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}|\tau_1|\exp\left(-\frac{\tau_1^2}{2}\right)\cdot\int_{\frac{-\gamma-\delta-\rho\tau_1}{\sqrt{2(1-\rho^2)}}}^{\frac{\gamma+\delta-\rho\tau_1}{\sqrt{2(1-\rho^2)}}}\exp\left(-\tau^2\right)d\tau d\tau_1 \qquad \text{by changing variables}\\ &= \frac{1}{\sqrt{8\pi}}\int_{-\infty}^{\infty}|\tau_1|\exp\left(-\frac{\tau_1^2}{2}\right)\cdot\left(\operatorname{erf}\left(\frac{\gamma+\delta-\rho\tau_1}{\sqrt{2(1-\rho^2)}}\right)-\operatorname{erf}\left(\frac{-\gamma-\delta-\rho\tau_1}{\sqrt{2(1-\rho^2)}}\right)\right)d\tau_1 &(114)\end{aligned}$$

For $|\rho| < 1$, $\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]$ is a continuous function of ρ. The last integral (114) can be calculated numerically. Figure 6 plots $\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]$ for $\gamma = 0.15$ and $\delta = 0.01$ over $\rho \in (-1, 1)$. Furthermore, (113) indicates that $\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]$ is monotonically increasing with both θ and δ. Thus, we obtain a universal bound

$$\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}] \leq 0.11\|\boldsymbol{h}\| \quad \text{for } \gamma < 0.15 \text{ and } \delta = 0.01, \tag{115}$$

which further implies $\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x(\boldsymbol{a}_i^T\boldsymbol{x})] \leq 0.11\|\boldsymbol{h}\|$ for $\gamma < 0.15$ and $\delta = 0.01$.

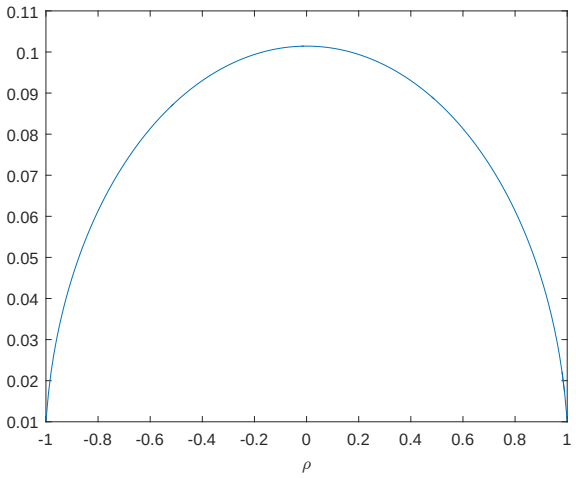


Figure 6: $\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{x}|<\gamma+\delta\}}]$ with respect to ρ

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 [48], we have

$$\mathcal{P}\left[\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x(\boldsymbol{a}_i^T\boldsymbol{x}) > (0.11+\epsilon)\,\|\boldsymbol{h}\|\right] < \exp(-cm\epsilon^2), \tag{116}$$

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

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}_0|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right) \leq 0.11+\epsilon, \quad \forall(\boldsymbol{x}_0, \boldsymbol{h}_0) \in \mathcal{N}_\epsilon \tag{117}$$

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

$$\left|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}\right) - \chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)\right| \leq \frac{1}{\delta}\left|\boldsymbol{a}_i^T(\boldsymbol{x}-\boldsymbol{x}_0)\right|. \tag{118}$$

Consequently, on the event that (106) holds, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)-\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}_0|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)\right| \\
&\le \frac{1}{m}\sum_{i=1}^{m}\left||\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}\right)-|\boldsymbol{a}_i^T\boldsymbol{h}_0|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)\right| \\
&\le \frac{1}{m}\sum_{i=1}^{m}\left|\boldsymbol{a}_i^T(\boldsymbol{h}-\boldsymbol{h}_0)\right|+\frac{1}{m}\sum_{i=1}^{m}\frac{1}{\delta}\left|\boldsymbol{a}_i^T\boldsymbol{h}_0\right|\cdot\left|\boldsymbol{a}_i^T\boldsymbol{x}-\boldsymbol{a}_i^T\boldsymbol{x}_0\right| \qquad \text{due to (118)} \\
&\le c_2'\|\boldsymbol{h}-\boldsymbol{h}_0\|+\frac{1}{\delta}\cdot c_2\|\boldsymbol{h}_0(\boldsymbol{x}-\boldsymbol{x}_0)^T\|_F \qquad \text{due to (106)} \\
&\le c_3\epsilon/\delta.
\end{aligned}$$

On the intersection of events that (117) and (106) hold, we have

$$\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{a}_i^T\boldsymbol{h}|\chi_x\left(\boldsymbol{a}_i^T\boldsymbol{x}_0\right)\le\left(0.11+\epsilon+2c_3\epsilon/\delta\right), \tag{119}$$

for all $\boldsymbol{x}$ and $\boldsymbol{h}$ with unit length.

We next prove (110). 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

$$|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'\|\boldsymbol{h}\|\}}\le\chi_h(\boldsymbol{a}_i^T\boldsymbol{h})\le|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'(1-\delta)\|\boldsymbol{h}\|\}}. \tag{120}$$

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:

$$\begin{aligned}
\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'(1-\delta)\|\boldsymbol{h}\|\}}] &= \int_{-\infty}^{\infty}|\tau|\mathbf{1}_{\{|\tau|>\gamma'(1-\delta)\}}\cdot f(\tau)d\tau, \\
&= 2\cdot\frac{1}{\sqrt{2\pi}}\int_{\gamma'(1-\delta)}^{\infty}\tau\exp\left(-\frac{\tau^2}{2}\right)d\tau \\
&= \sqrt{\frac{2}{\pi}}\exp(-\gamma'^2(1-\delta)^2/2)<0.01 \qquad \text{for } \gamma'=3,\delta=0.01,
\end{aligned} \tag{121}$$

where $f(\tau)$ is the density of the standard Gaussian distrution. We note that $\mathbb{E}[|\boldsymbol{a}_i^T\boldsymbol{h}|\mathbf{1}_{\{|\boldsymbol{a}_i^T\boldsymbol{h}|>\gamma'(1-\delta)\|\boldsymbol{h}\|\}}]$ is monotonically increasing with δ and decreasing with γ'. Furthermore, $\mathbb{E}[\chi_h(\boldsymbol{a}_i^T\boldsymbol{h})]\le0.01\|\boldsymbol{h}\|$ for $\gamma'\ge3$ and $\delta\le0.01$.

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 [48], we have

$$\mathcal{P}\left[\frac{1}{m}\sum_{i=1}^{m}\chi_h(\boldsymbol{a}_i^T\boldsymbol{h})>\left(0.01+\epsilon\right)\|\boldsymbol{h}\|\right]<\exp(-cm\epsilon^2), \tag{122}$$

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) $|\mathcal{N}_\epsilon| \le (1+2/\epsilon)^n$. Then for any $\boldsymbol{h}$ with unit length, there exists a $\boldsymbol{h}_0 \in \mathcal{N}_\epsilon$ such that $\|\boldsymbol{h}-\boldsymbol{h}_0\| \le \epsilon$. Taking the union bound for all the points on the net, one can show

$$\frac{1}{m}\sum_{i=1}^m \chi_h(\boldsymbol{a}_i^T\boldsymbol{h}_0) \le 0.01+\epsilon, \quad \forall \boldsymbol{h}_0 \in \mathcal{N}_\epsilon \tag{123}$$

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

Consequently, we have

$$\begin{aligned}
&\left|\frac{1}{m}\sum_{i=1}^m \chi_h(\boldsymbol{a}_i^T\boldsymbol{h}) - \frac{1}{m}\sum_{i=1}^m \chi_h(\boldsymbol{a}_i^T\boldsymbol{h}_0)\right| \\
&\le \frac{1}{m}\sum_{i=1}^m \left|\chi_h(\boldsymbol{a}_i^T\boldsymbol{h}) - \chi_h(\boldsymbol{a}_i^T\boldsymbol{h}_0)\right| \\
&\le \frac{1}{\delta}\cdot\frac{1}{m}\sum_{i=1}^m \left|\boldsymbol{a}_i^T(\boldsymbol{h}-\boldsymbol{h}_0)\right| \\
&\le \frac{1}{\delta}c_2'\|\boldsymbol{h}-\boldsymbol{h}_0\| \le c_3\epsilon/\delta,
\end{aligned}$$

where the second inequality is because $\chi_h(t)$ is Lipschitz continuous with constant $1/\delta$.

On the intersection of events that (123) and (106) hold, we have

$$\frac{1}{m}\sum_{i=1}^m \chi_h(\boldsymbol{a}_i^T\boldsymbol{h}) \le \left(0.01+\epsilon+c_3\epsilon/\delta\right), \tag{124}$$

for all $\boldsymbol{h}$ with unit length.

Putting together (119) and (124), and since ϵ can be arbitrarily small, the proof is completed.

References

[1] 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.

[2] S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. *arXiv preprint arXiv:1503.00778*, 2015.

[3] S. Bahmani and J. Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. *arXiv preprint arXiv:1610.04210*, 2016.

[4] R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

[5] 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.

[6] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.

[7] N. Boumal. Nonconvex phase synchronization. *arXiv preprint arXiv:1601.06114*, 2016.

[8] E. J. Candès and X. Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2014.

[9] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[10] 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*, 66(8):1241–1274, 2013.

[11] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Automata, languages and programming*, pages 693–703. Springer, 2002.

[12] K. Chen. On k-median clustering in high dimensions. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 2006.

[13] Y. Chen and E. Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems (NIPS)*. 2015.

[14] Y. Chen and E. Candes. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *arXiv preprint arXiv:1609.05820*, 2016.

[15] 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.

[16] Y. Chen, Y. Chi, and A. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, July 2015.

[17] 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.

[18] 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.

[19] L. Demanet and P. Hand. Stable optimizationless recovery from phaseless linear measurements. *Journal of Fourier Analysis and Applications*, 20(1):199–221, 2014.

[20] J. D. Donahue. Products and quotients of random variables and their applications. Technical report, DTIC Document, 1964.

[21] J. Drenth. *X-Ray Crystallography*. Wiley Online Library, 2007.

[22] J. R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982.

[23] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[24] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. *arXiv preprint arXiv:1610.07531*, 2016.

[25] P. Hand. Phaselift is robust to a constant fraction of arbitrary errors. *arXiv preprint arXiv:1502.04241*, 2015.

[26] P. Hand and V. Voroninski. An elementary proof of convex phase retrieval in the natural parameter space via the linear program phasemax. *arXiv preprint arXiv:1611.03935*, 2016.

[27] M. Hardt. Understanding alternating minimization for matrix completion. In *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 651–660. IEEE, 2014.

[28] P. J. Huber. *Robust statistics*. Springer, 2011.

[29] 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.

[30] C. Jin, S. M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *arXiv preprint arXiv:1605.08370*, 2016.

[31] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980 –2998, June 2010.

[32] K. Lee, Y. Li, M. Junge, and Y. Bresler. Blind recovery of sparse signals from subsampled convolution. *arXiv preprint arXiv:1511.06149*, 2015.

[33] Q. Li and G. Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv preprint arXiv:1611.03060*, 2016.

[34] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.

[35] X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 2013.

[36] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.

[37] Y. Li, Y. Sun, and Y. Chi. Low-rank positive semidefinite matrix recovery from corrupted rank-one measurements. *IEEE Transactions on Signal Processing*, 65(2):397–408, Jan 2017.

[38] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[39] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[40] 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.

[41] C. Qu and H. Xu. Subspace clustering with irrelevant features via robust dantzig selector. In *Advances in Neural Information Processing Systems (NIPS)*. 2015.

[42] M. Soltanolkotabi. *Algorithms and theory for clustering and nonconvex quadratic programming*. PhD thesis, Stanford University, 2014.

[43] 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.

[44] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

[45] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via nonconvex factorization. In *IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 270–289, 2015.

[46] R. J. Tibshirani. Fast computation of the median by successive binning. *arXiv preprint arXiv:0806.3301*, 2008.

[47] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[48] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*, pages 210 – 268, 2012.

[49] D. Wagner. Resilient aggregation in sensor networks. In *Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks*, pages 78–87. ACM, 2004.

[50] I. Waldspurger, A. d'Aspremont, and S. Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[51] 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.

[52] 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.

[53] D. Weller, A. Pnueli, G. Divon, O. Radzyner, Y. Eldar, and J. Fessler. Undersampled phase retrieval with outliers. *IEEE Transactions on Computational Imaging*, 1(4):247–258, Dec 2015.

[54] X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust pca via gradient descent. *arXiv preprint arXiv:1605.07784*, 2016.

[55] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. Reshaped Wirtinger Flow and Incremental Algorithm for Solving Quadratic System of Equations. *ArXiv 1605.07719*, May 2016.

[56] 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.

[57] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.