Supplementary Materials for "Solving Random Systems of Quadratic Equations via Truncated Generalized Gradient Flow"

G. Wang*,† and G. B. Giannakis*
* ECE Dept. and Digital Tech. Center
University of Minnesota
Minneapolis, MN 55455, USA
† School of Automation
Beijing Institute of Technology
Beijing 100081, China
{gangwang, georgios}@umn.edu

October 24, 2016

1 Algorithm and main theorem

For convenience of presentation, we begin with repeating our model and main assumptions: Consider the noise-free real Gaussian model

$$\psi_i = \left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}\right|, \quad i \in [m] := \{1, 2, \cdots, m\} \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is the wanted unknown, and $\{\boldsymbol{a}_i\}_{i=1}^m$ are drawn independently and identically from the n-dimensional real Gaussian distribution, i.e., $\boldsymbol{a}_i \in \mathbb{R}^n \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. For notational brevity, define $\boldsymbol{A} := [\boldsymbol{a}_1 \ \cdots \boldsymbol{a}_m]^{\mathcal{T}}$, $\boldsymbol{\psi} := [\psi_1 \ \cdots \ \psi_m]^{\mathcal{T}}$, and $\boldsymbol{y} := [y_1 \ \cdots \ y_m]^{\mathcal{T}}$, where $y_i := \psi_i^2$ stands for the squared magnitudes or intensity. Assume that the quadratic system (1) admits a unique solution, which indeed holds true as long as $m \geq 2n-1$ generic measurements are taken [1]. Throughout the subsequent analysis, fix $\boldsymbol{x}$ to be any solution of the given system in (1). Note that if $\boldsymbol{x}$ satisfies the system in (1), so does $-\boldsymbol{x}$; i.e., the solution set for real-valued models becomes $\{\boldsymbol{x}, -\boldsymbol{x}\}$. Our theoretical analysis focuses on $\boldsymbol{x}$, rather than $-\boldsymbol{x}$. Introduce the notion of Euclidean distance of any estimate $\boldsymbol{z}$ to the solution set: $\mathrm{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min \|\boldsymbol{z} \pm \boldsymbol{x}\|$ for real signals; and $\mathrm{dist}(\boldsymbol{z}, \boldsymbol{x}) := \min_{\phi\in[0,2\pi)} \left\|\boldsymbol{z} - \boldsymbol{x}\mathrm{e}^{i\phi}\right\|$ for complex ones [2]. For concreteness, our results focus on real-valued models, define also the unrecoverable global phase factor for real-valued signals [2]

$$\phi(\boldsymbol{z}) := \begin{cases} 0, & \|\boldsymbol{z}-\boldsymbol{x}\| \leq \|\boldsymbol{z}+\boldsymbol{x}\| \\ \pi, & \text{otherwise.} \end{cases} \tag{2}$$

Henceforth, we always presume $\phi(\boldsymbol{z}) = 0$; otherwise, $\boldsymbol{z}$ is replaced by $\mathrm{e}^{-j\phi(\boldsymbol{z})}\boldsymbol{z}$, but for simplicity of presentation, the constant phase adaptation term is dropped whenever it is clear from the context [3].

Algorithm 1 and Theorem 1 are repeated next.

Algorithm 1 Truncated generalized gradient flow (TGGF) solver

1: **Input:** Data $\{\psi_i\}_{i=1}^m$ and features $\{\boldsymbol{a}_i\}_{i=1}^m$; maximum iterations T; by default, set constant step size $\mu = 0.6/1$ for real/complex-valued models, thresholds $|\overline{\mathcal{I}}_0| = \lceil \frac{1}{6}m \rceil$,[1]and $\gamma = 0.7$.

2: **Evaluate** $\psi_i/\|\boldsymbol{a}_i\|$, $i = 1, \ldots, m$, and find $\overline{\mathcal{I}}_0$ comprising indices associated with the $|\overline{\mathcal{I}}_0|$ largest $(\psi_i/\|\boldsymbol{a}_i\|)$'s.

3: **Initialize** $\boldsymbol{z}_0$ to $\sqrt{\sum_{i=1}^m \psi_i^2/m}\tilde{\boldsymbol{z}}_0$, where $\tilde{\boldsymbol{z}}_0$ is the unit leading eigenvector of $\bar{\boldsymbol{Y}}_0 := \frac{1}{|\overline{\mathcal{I}}_0|}\sum_{i\in\overline{\mathcal{I}}_0} \frac{\boldsymbol{a}_i\boldsymbol{a}_i^{\mathcal{T}}}{\|\boldsymbol{a}_i\|^2}$.

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

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

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

5: **Output:** $\boldsymbol{z}_T$

Theorem 1. *[4] Let $\boldsymbol{x} \in \mathbb{R}^n$ be an arbitrary signal, and consider (noise-free) measurements $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$, in which $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le i \le m$. Then with probability at least $1-(m+5)\mathrm{e}^{-n/2}-\mathrm{e}^{-c_0 m}-1/n^2$ for some universal constant $c_0 > 0$, the initialization $\boldsymbol{z}_0$ returned by the orthogonality-promoting method in Algorithm 1 satisfies*

$$\operatorname{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \le \rho\|\boldsymbol{x}\| \tag{3}$$

with $\rho = 1/10$ (or any sufficiently small positive constant), provided that $m \ge c_1|\overline{\mathcal{I}}_0| \ge c_2 n$ for some numerical constants c_1, $c_2 > 0$ and sufficiently large n. Further, choosing a constant step size $\mu \le \mu_0$ along with a fixed truncation level $\gamma \ge 1/2$, and starting from any guess $\boldsymbol{z}_0$ satisfying (5)*, successive estimates of the TGGF algorithm (tabulated in Algorithm 1) obey*

$$\operatorname{dist}(\boldsymbol{z}_t, \boldsymbol{x}) \le \rho(1-\nu)^t\|\boldsymbol{x}\|, \forall t = 1, 2, \ldots \tag{4}$$

for some $0 < \nu < 1$, which holds with probability exceeding $1-(m+5)\mathrm{e}^{-n/2}-8\mathrm{e}^{-c_0 m}-1/n^2$. Typical parameter values are $\mu = 0.6$, and $\gamma = 0.7$.

2 Proofs

To prove Theorem 1, this section establishes a few lemmas and the main ideas, while technical details are deferred to the Appendix. Relative to WF and TWF, our objective function involves nonsmoothness and nonconvexity, rendering the proof of exact recovery of TGGF nontrivial. In addition, our initialization method starts from a rather different perspective than the spectral alternatives, so the thoughts and tools involved in proving performance of our initialization deviate from those of the spectral methods [5, 2, 3]. Part of the proof is adapted from [2, 3] and [6].

The proof of Theorem 1 consists of two parts: Section 2.1 justifies the performance of the proposed orthogonality-promoting initialization, which essentially achieves any given constant relative error as soon as the number of equations is on the order of the number of unknowns, namely, $m \asymp n$.[2]

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

Section 2.2 demonstrates theoretical convergence of TGGF to the solution of the quadratic system in (1) at a geometric rate provided that the initial estimate has a sufficiently small constant relative error as in (3). The two stages of TGGF can be performed independently, meaning that other better initialization methods, if available, could be adopted to initialize our truncated generalized gradient iterations; likewise, our initialization method can also be applied to initialize other iterative optimization algorithms.

2.1 Constant Relative Error by Orthogonality-promoting Initialization

This section concentrates on proving guaranteed performance of the proposed orthogonality-promoting initialization method, as asserted in the following proposition.

Proposition 1. *Fix $\boldsymbol{x} \in \mathbb{R}^n$ arbitrarily, and consider the noiseless case $\psi_i = |\boldsymbol{a}_i^{\mathcal{T}} \boldsymbol{x}|$, where $\boldsymbol{a}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, $1 \le i \le m$. Then with probability at least $1 - (m+5)\mathrm{e}^{-n/2} - \mathrm{e}^{-c_0 m} - 1/n^2$ for some universal constant $c_0 > 0$, the initialization $\boldsymbol{z}_0$ returned by the orthogonality-promoting method satisfies*

$$\operatorname{dist}(\boldsymbol{z}_0, \boldsymbol{x}) \le \rho \|\boldsymbol{x}\| \tag{5}$$

for $\rho = 1/10$ or any positive constant, with the proviso that $m \ge c_1 |\overline{\mathcal{I}}_0| \ge c_2 n$ for some numerical constants c_1, $c_2 > 0$ and sufficiently large n.

Due to homogeneity in (5), it suffices to work with the case where $\|\boldsymbol{x}\| = 1$. Assume for the moment that $\|\boldsymbol{x}\| = 1$ is known and $\boldsymbol{z}_0$ has been scaled such that $\|\boldsymbol{z}_0\| = 1$. Subsequently, the error between the employed $\boldsymbol{x}$'s norm estimate $\sqrt{\frac{1}{m}\sum_{i=1}^m y_i}$ and the unknown norm $\|\boldsymbol{x}\| = 1$ will be accounted for at the end of this Section. Instrumental in proving Proposition 1 is the following result, whose proof is deferred to Appendix A.1.

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

$$\frac{1}{2} \left\| \boldsymbol{x}\boldsymbol{x}^{\mathcal{T}} - \boldsymbol{z}_0 \boldsymbol{z}_0^{\mathcal{T}} \right\|_F^2 \le \frac{\left\| \overline{\boldsymbol{S}}_0 \boldsymbol{u} \right\|^2}{\left\| \overline{\boldsymbol{S}}_0 \boldsymbol{x} \right\|^2} \tag{6}$$

for $\boldsymbol{z}_0 = \tilde{\boldsymbol{z}}_0$, where the unit vector $\tilde{\boldsymbol{z}}_0$ is given by

$$\tilde{\boldsymbol{z}}_0 := \arg \max_{\|\boldsymbol{z}\|=1} \quad \boldsymbol{z}^{\mathcal{T}} \overline{\boldsymbol{Y}}_0 \boldsymbol{z} \tag{7}$$

with $\overline{\boldsymbol{Y}}_0 := \frac{1}{|\overline{\mathcal{I}}_0|} \overline{\boldsymbol{S}}_0^{\mathcal{T}} \overline{\boldsymbol{S}}_0$, and $\overline{\boldsymbol{S}}_0$ is formed by removing the rows of $\boldsymbol{S} := \left[\boldsymbol{a}_1 / \|\boldsymbol{a}_1\| \ \cdots \ \boldsymbol{a}_m / \|\boldsymbol{a}_m\| \right]^{\mathcal{T}} \in \mathbb{R}^{m \times n}$, if their indices do not belong to the set $\overline{\mathcal{I}}_0$ specified in Algorithm 1.

We now turn to prove Proposition 1. The first step consists in upper-bounding the term on the right-hand-side of (6). Specifically, its numerator term will be upper bounded, and the denominator term lower bounded, which are summarized in Lemma 2 and Lemma 3, whose proofs can be found in Appendix A.2 and Appendix A.3, respectively.

Lemma 2. *In the setup of Lemma 1, if $|\overline{\mathcal{I}}_0| \ge c_1' n$, then the next*

$$\left\| \overline{\boldsymbol{S}}_0 \boldsymbol{u} \right\|^2 \le 1.01 |\overline{\mathcal{I}}_0| / n \tag{8}$$

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

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

$$\left\|\bar{\boldsymbol{S}}_0\boldsymbol{x}\right\|^2 \geq \frac{0.99|\bar{\mathcal{I}}_0|}{2.3n}\left[1+\log\left(m/|\bar{\mathcal{I}}_0|\right)\right] \tag{9}$$

provided that $|\bar{\mathcal{I}}_0| \geq c_1' n$, $m \geq c_2'|\bar{\mathcal{I}}_0|$, *and* $m \geq c_3' n$ *for some absolute constants* c_1', c_2', $c_3' > 0$, *and sufficiently large* n.

Therefore, putting the upper and lower bounds in (8) and (9) together, one arrives at

$$\frac{\left\|\bar{\boldsymbol{S}}_0\boldsymbol{u}\right\|^2}{\left\|\bar{\boldsymbol{S}}_0\boldsymbol{x}\right\|^2} \leq \frac{2.4}{1+\log\left(m/|\bar{\mathcal{I}}_0|\right)} \triangleq \kappa \tag{10}$$

which holds with probability at least $1-(m+3)\mathrm{e}^{-n/2}-\mathrm{e}^{-c_0 m}-1/n^2$, with the proviso that $m \geq c_1'|\bar{\mathcal{I}}_0|$, and $m \geq c_2' n$, $|\bar{\mathcal{I}}_0| \geq c_3' n$ for some absolute constants c_1', c_2', $c_3' > 0$, and sufficiently large n.

Apparently, the bound κ in (10) is meaningful only when the ratio $\log(m/|\bar{\mathcal{I}}_0|) > 1.4$, i.e., $m/|\bar{\mathcal{I}}_0| > 4$, because the left hand side expressible in terms of $\sin^2\theta$ enjoys a trivial upper bound 1. Henceforth, we will work with the case where $m/|\bar{\mathcal{I}}_0| > 4$. Empirically, $\lfloor m/|\bar{\mathcal{I}}_0| \rfloor = 6$ or equivalently $|\bar{\mathcal{I}}_0| = \lceil \frac{1}{6}m \rceil$ in Algorithm 1 works well when m/n is relatively small. Note further that the bound κ can be made arbitrarily small by letting $m/|\bar{\mathcal{I}}_0|$ be large enough. Without any loss of generality, let us take $\kappa := 0.001$. An additional step leads to the wanted bound on the distance between $\tilde{\boldsymbol{z}}_0$ and $\boldsymbol{x}$; similar arguments can be found in [2, Section 7.8]. Recall that

$$|\boldsymbol{x}^{\mathcal{T}}\tilde{\boldsymbol{z}}_0|^2 = \cos^2\theta = 1-\sin^2\theta \geq 1-\kappa, \tag{11}$$

so one has

$$\begin{aligned}\mathrm{dist}^2(\tilde{\boldsymbol{z}}_0,\, \boldsymbol{x}) &\leq \|\tilde{\boldsymbol{z}}_0\|^2 + \|\boldsymbol{x}\|^2 - 2|\boldsymbol{x}^{\mathcal{T}}\tilde{\boldsymbol{z}}_0| \\ &\leq \left(2-2\sqrt{1-\kappa}\right)\|\boldsymbol{x}\|^2 \\ &\approx \kappa\,\|\boldsymbol{x}\|^2 .\end{aligned} \tag{12}$$

Coming back to the case in which $\|\boldsymbol{x}\|$ is unknown stated prior to Lemma 1, the unit eigenvector $\tilde{\boldsymbol{z}}_0$ is scaled by the estimate of $\|\boldsymbol{x}\|$ to yield the initial guess $\boldsymbol{z}_0 = \sqrt{\frac{1}{m}\sum_{i=1}^m y_i}\tilde{\boldsymbol{z}}_0$. Using the results in Lemma 7.8 in [2], the following holds with high probability

$$\|\boldsymbol{z}_0 - \tilde{\boldsymbol{z}}_0\| = |\|\boldsymbol{z}_0\| - 1| \leq (1/20)\,\|\boldsymbol{x}\|\,. \tag{13}$$

Summarizing the two inequalities, we conclude that

$$\mathrm{dist}(\boldsymbol{z}_0,\, \boldsymbol{x}) \leq \|\boldsymbol{z}_0 - \tilde{\boldsymbol{z}}_0\| + \mathrm{dist}(\tilde{\boldsymbol{z}}_0,\, \boldsymbol{x}) \leq (1/10)\,\|\boldsymbol{x}\|\,. \tag{14}$$

The initialization thus obeys $\mathrm{dist}(\boldsymbol{z}_0,\, \boldsymbol{x})/\|\boldsymbol{x}\| \leq 1/10$ for any $\boldsymbol{x} \in \mathbb{R}^n$ with high probability provided that $m \geq c_1|\bar{\mathcal{I}}_0| \geq c_2 n$ holds for some universal constants c_1, $c_2 > 0$ and sufficiently large n.

2.2 Exact Recovery from Noiseless Data

We now prove that with accurate enough initial estimates, TGGF converges at a geometric rate to $\boldsymbol{x}$ with high probability (i.e., the second part of Theorem 1). To be specific, with initialization

obeying (5) in Proposition 1, TGGF reconstructs the solution $\boldsymbol{x}$ exactly in linear time. In this direction, it suffices to demonstrate that the TGGF's update rule (i.e., Step 4 in Algorithm 1) is locally contractive within a sufficiently small neighborhood of $\boldsymbol{x}$, as asserted in the following proposition.

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

$$\mathrm{dist}^2\left(\boldsymbol{z} + \frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}),\, \boldsymbol{x}\right) \le (1-\nu)\mathrm{dist}^2\left(\boldsymbol{z},\, \boldsymbol{x}\right) \tag{15}$$

for all $\boldsymbol{x},\, \boldsymbol{z} \in \mathbb{R}^n$ obeying the condition (5) for sufficiently small $\rho > 0$ with the proviso that $m \ge c_1 n$ and that the constant step size μ satisfying $0 < \mu \le \mu_0$ for some $\mu_0 > 0$.

Proposition 2 demonstrates that the distance of TGGF's successive iterates to $\boldsymbol{x}$ is monotonically decreasing once the algorithm enters a small-size neighborhood around $\boldsymbol{x}$. This neighborhood is commonly referred to as the *basin of attraction*; see further discussions in [2, 7, 3]. In other words, as soon as one lands within the basin of attraction, TGGF's iterates remain in this region and will be attracted to $\boldsymbol{x}$ exponentially fast. To substantiate Proposition 2, recall the concept of the *local regularity condition*, which was first developed in [2] and plays a fundamental role in establishing linear convergence to global optimum of nonconvex optimization approaches such as WF/TWF [2, 7, 3]. Now consider the update rule of TGGF

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}_t), \quad \forall t \ge 0, \tag{16}$$

where the truncated gradient $\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}_t)$ (See Remark 1 in [8] for more discussion) evaluated at some point $\boldsymbol{z}_t \in \mathbb{R}^n$ is given by

$$\frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}_t) \stackrel{\triangle}{=} \frac{1}{m}\sum_{i\in\mathcal{I}}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t - \psi_i\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t}{\left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}_t\right|}\right)\boldsymbol{a}_i.$$

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

$$\left\langle \frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}),\, \boldsymbol{h}\right\rangle \ge \frac{\mu}{2}\left\|\frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z})\right\|^2 + \frac{\lambda}{2}\left\|\boldsymbol{h}\right\|^2 \tag{17}$$

holds for all $\boldsymbol{z} \in \mathbb{R}^n$ such that $\|\boldsymbol{h}\| = \|\boldsymbol{z} - \boldsymbol{x}\| \le \epsilon\|\boldsymbol{x}\|$ for some constant $0 < \epsilon < 1$, where the ball $\|\boldsymbol{z} - \boldsymbol{x}\| \le \epsilon\|\boldsymbol{x}\|$ is the so-called *basin of attraction*. Simple linear algebra along with the regularity condition in (17) leads to

$$\begin{aligned}
\mathrm{dist}^2\left(\boldsymbol{z} - \frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}), \boldsymbol{x}\right) &= \left\|\boldsymbol{z} - \frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z}) - \boldsymbol{x}\right\|^2 \\
&= \left\|\boldsymbol{h}\right\|^2 - 2\mu\left\langle \boldsymbol{h}, \frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z})\right\rangle + \left\|\frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z})\right\|^2 \qquad (18) \\
&\le \left\|\boldsymbol{h}\right\|^2 - 2\mu\left(\frac{\mu}{2}\left\|\frac{1}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z})\right\|^2 + \frac{\lambda}{2}\left\|\boldsymbol{h}\right\|^2\right) + \left\|\frac{\mu}{m}\nabla\ell_{\mathrm{tr}}(\boldsymbol{z})\right\|^2 \\
&= (1-\lambda\mu)\left\|\boldsymbol{h}\right\|^2 = (1-\lambda\mu)\,\mathrm{dist}^2(\boldsymbol{z}, \boldsymbol{x}) \qquad (19)
\end{aligned}$$

for all $\boldsymbol{z}$ obeying $\|\boldsymbol{h}\| \le \epsilon\|\boldsymbol{x}\|$. Clearly, if the $\mathrm{LRC}(\mu, \lambda, \epsilon)$ is proved for TGGF, our goal (15) follows upon letting $\nu := \lambda\mu$.

2.2.1 Proof of the local regularity condition in (17)

By definition, justifying the local regularity condition in (17) entails controlling the norm of the truncated gradient $\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})$, i.e., bounding the last term in (18). Recall that

$$\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})=\frac{1}{m}\sum_{i\in\mathcal{I}}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}-\psi_i\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}}{\left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}\right|}\right)\boldsymbol{a}_i\triangleq\frac{1}{m}\boldsymbol{A}\boldsymbol{v} \tag{20}$$

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

$$|v_i|^2=\left|\left(|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|-|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\right)\mathbb{1}_{\left\{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|\ge|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|/(1+\gamma)\right\}}\right|^2\le\left||\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{z}|-|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|\right|^2\le\left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right|^2=a_{i,1}^2\|\boldsymbol{h}\|^2, \tag{21}$$

where $\boldsymbol{h}=\boldsymbol{z}-\boldsymbol{x}$. Observe that $a_{i,1}^2$ obeys the *Chi-square* distribution with $k=1$ degrees of freedom; yet due to our working assumption $\|\boldsymbol{a}_i\|\le\sqrt{2.3n}$, it has mean $\mathbb{E}[a_{i,1}^2]\le k=1$. So fixing any $0<\delta'<1$ and applying the one-sided Bernstein-type inequality, the following holds with probability at least $1-\mathrm{e}^{-m\delta'^2/2}$ [9, Proposition 5.16]

$$\|\boldsymbol{v}\|^2=\sum_{i=1}^m v_i^2\le\sum_{i=1}^m a_{i,1}^2\|\boldsymbol{h}\|^2\le(1+\delta')m\|\boldsymbol{h}\|^2. \tag{22}$$

On the other hand, standard matrix concentration results confirm that the largest singular value of $\boldsymbol{A}=[\boldsymbol{a}_1\ \cdots\ \boldsymbol{a}_m]^{\mathcal{T}}$ with i.i.d. Gaussian $\{\boldsymbol{a}_i\}$ satisfies $\sigma_1:=\|\boldsymbol{A}\|\le(1+\delta'')\sqrt{m}$ for some $\delta''>0$ with probability exceeding $1-2\mathrm{e}^{-c_0m}$ as soon as $m\ge c_1n$ for sufficiently large $c_1>0$, where $c_1>0$ is a universal constant depending on δ'' [9, Remark 5.25]. Putting together (20), (21), and (22) yields

$$\left\|\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})\right\|\le\frac{1}{m}\|\boldsymbol{A}\|\cdot\|\boldsymbol{v}\|\le(1+\delta')(1+\delta'')\|\boldsymbol{h}\|\le(1+\delta)^2\|\boldsymbol{h}\|,\quad\delta:=\max\{\delta',\delta''\} \tag{23}$$

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

Notice that to prove the LRC, it suffices to show that the truncated gradient $\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})$ ensures sufficient descent, i.e., it obeys a uniform lower bound along the search direction $\boldsymbol{h}$ taking the form

$$\left\langle\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z}),\,\boldsymbol{h}\right\rangle\gtrsim\|\boldsymbol{h}\|^2 \tag{24}$$

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

Proposition 3. *Consider the noiseless measurements $\psi_i=|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|$ and fix any sufficiently small constant $\epsilon>0$. There exist universal constants c_0, $c_1>0$ such that if $m>c_1n$, then the following holds with probability exceeding $1-4\mathrm{e}^{-c_0m}$*

$$\left\langle\boldsymbol{h},\frac{1}{m}\nabla\ell_{\rm tr}(\boldsymbol{z})\right\rangle\ge2\left(1-\zeta_1-\zeta_2-2\epsilon\right)\|\boldsymbol{h}\|^2 \tag{25}$$

for all $\boldsymbol{x}$, $\boldsymbol{z}\in\mathbb{R}^n$ such that $\|\boldsymbol{h}\|/\|\boldsymbol{x}\|\le\rho$ for $0<\rho\le1/10$ and any fixed $\gamma\ge1/2$.

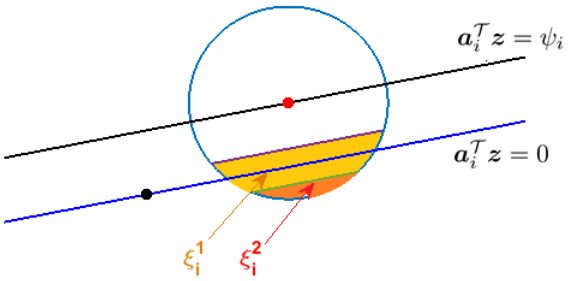

Figure 1: The geometric understanding of the proposed truncation rule on the i-th gradient component involving $\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}=\psi_i$, where the red dot denotes the solution $\boldsymbol{x}$ and the black one is the origin. Hyperplanes $\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}=\psi_i$ and $\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}=0$ (of $\boldsymbol{z}\in\mathbb{R}^n$) passing through points $\boldsymbol{z}=\boldsymbol{x}$ and $\boldsymbol{z}=\boldsymbol{0}$, respectively, are shown.

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

$$\mathcal{E}_i := \left\{\frac{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}|}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}|}\geq\frac{1}{1+\gamma}\right\}, \tag{26}$$

$$\mathcal{D}_i := \left\{\frac{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}|}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}|}\geq\frac{2+\gamma}{1+\gamma}\right\}, \tag{27}$$

$$\text{and}\quad \mathcal{K}_i := \left\{\frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}|}\neq\frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}|}\right\} \tag{28}$$

where $\boldsymbol{h}=\boldsymbol{z}-\boldsymbol{x}$. Under the condition $\|\boldsymbol{h}\|/\|\boldsymbol{x}\|\leq\rho$, the following inclusion holds

$$\mathcal{E}_i\cap\mathcal{K}_i\subseteq\mathcal{D}_i. \tag{29}$$

Proof. From Fig. 1, it is clear that if $\boldsymbol{z}\in\xi_i^2$, then the sign of $\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}$ will be different than that of $\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}$. The region ξ_i^2, however, can be specified by the conditions that $\frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}|}\neq\frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}|}$ and $\frac{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}|}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}|}\geq 1+\frac{1}{1+\gamma}=\frac{2+\gamma}{1+\gamma}$. Under our initialization condition $\|\boldsymbol{h}\|/\|\boldsymbol{x}\|\leq\rho$, it is self-evident that $\mathcal{D}_i$ describes two spherical caps that contain ξ_i^2. Hence, it holds that $\mathcal{E}_i\cap\mathcal{K}_i=\xi_i^2\subseteq\mathcal{D}_i$. □

Rewrite the truncated gradient as

$$\begin{aligned}\frac{1}{2m}\nabla\ell_{\text{tr}}(\boldsymbol{z}) &= \frac{1}{m}\sum_{i=1}^m\left(\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}-|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}|\frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}|}\right)\boldsymbol{a}_i\mathbb{1}_{\mathcal{E}_i}\\ &= \frac{1}{m}\sum_{i=1}^m\boldsymbol{a}_i\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}\mathbb{1}_{\mathcal{E}_i}-\frac{1}{m}\sum_{i=1}^m\left(\frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}|}-\frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}}{|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}|}\right)|\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}|\,\boldsymbol{a}_i\mathbb{1}_{\mathcal{E}_i}.\end{aligned} \tag{30}$$

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

$$\begin{aligned}\left\langle \frac{1}{2m}\nabla \ell_{\mathrm{tr}}(\boldsymbol{z}),\, \boldsymbol{h} \right\rangle &\geq \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2 \mathbb{1}_{\mathcal{E}_i} - \frac{2}{m}\sum_{i=1}^{m}\left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}\right|\left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right|\mathbb{1}_{\mathcal{E}_i\cap\mathcal{K}_i} \\ &\geq \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2 \mathbb{1}_{\mathcal{E}_i} - \frac{2}{m}\sum_{i=1}^{m}\left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}\right|\left|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right|\mathbb{1}_{\mathcal{D}_i} \\ &\geq \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2 \mathbb{1}_{\mathcal{E}_i} - \frac{1+\gamma}{2+\gamma}\cdot\frac{2}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2 \mathbb{1}_{\mathcal{D}_i} \end{aligned} \tag{31}$$

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

Proving the regularity condition boils down to lower bounding the right-hand side of (31), specifically, to lower bounding the first term and to upper bounding the second one. Apparently, the first term approximately gives $\|\boldsymbol{h}\|^2$ by the SLLN as long as our truncation procedure does not eliminate too many generalized gradient components (i.e., summands in the first term). Regarding the second, one would expect its contribution to be small under our initialization condition in (5) and as the relative error $\|\boldsymbol{h}\| / \|\boldsymbol{x}\|$ decreases. Specifically, under our initialization, $\mathcal{D}_i$ is provably a rare event, thus eliminating the possibility of the second term exerting a noticeable influence on the first term. Rigorous analyses concerning the two terms are elaborated in Lemma 5 and Lemma 6, whose proofs can be found in Appendix A.4 and Appendix A.5, respectively.

Lemma 5. *Fix $\gamma \geq 1/2$ and $\rho \leq 1/10$, and let $\mathcal{E}_i$ be defined in (26). For independent random variables $W \sim \mathcal{N}(0,1)$ and $Z \sim \mathcal{N}(0,1)$, set*

$$\zeta_1 := 1 - \min\left\{ \mathbb{E}\left[\mathbb{1}_{\left\{\left|\frac{1-\rho}{\rho}+\frac{W}{Z}\right| \geq \frac{\sqrt{1.01}}{\rho(1+\gamma)}\right\}}\right],\ \mathbb{E}\left[Z^2\mathbb{1}_{\left\{\left|\frac{1-\rho}{\rho}+\frac{W}{Z}\right| \geq \frac{\sqrt{1.01}}{\rho(1+\gamma)}\right\}}\right]\right\}. \tag{32}$$

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

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2 \mathbb{1}_{\mathcal{E}_i} \geq (1 - \zeta_1 - \epsilon)\|\boldsymbol{h}\|^2, \tag{33}$$

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

To have a sense of how large the quantities involved in (5) are, when $\gamma = 0.7$ and $\rho = 1/10$, it holds $\mathbb{E}\left[\mathbb{1}_{\left\{\left|\frac{1-\rho}{\rho}+\frac{W}{Z}\right| \geq \frac{\sqrt{1.01}}{\rho(1+\gamma)}\right\}}\right] \approx 0.92$, and $\mathbb{E}\left[Z^2\mathbb{1}_{\left\{\left|\frac{1-\rho}{\rho}+\frac{W}{Z}\right| \geq \frac{\sqrt{1.01}}{\rho(1+\gamma)}\right\}}\right] \approx 0.99$, hence leading to $\zeta_1 \approx 0.08$.

Having derived a lower bound for the first term in the right-hand side of (31), it remains to deal with the second one.

Lemma 6. *Fix $\gamma > 0$ and $\rho \leq 1/10$, and let $\mathcal{D}_i$ be defined in (27). For any constant $\epsilon > 0$, there exists some universal constants c_5, $c_6 > 0$ such that*

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2 \mathbb{1}_{\mathcal{D}_i} \leq (\zeta_2' + \epsilon)\|\boldsymbol{h}\|^2 \tag{34}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_5\epsilon^2 m}$ provided that $m/n > \left(c_6 \cdot \epsilon^{-2}\log\epsilon^{-1}\right)$ for some universal constants c_5, $c_6 > 0$, where $\zeta_2' = 1.3785\sqrt{\rho\tau/(0.99\tau^2 - \rho^2)}$ with $\tau := (2+\gamma)/(1+\gamma)$.

With our TGGF default parameters $\rho = 1/10$ and $\gamma = 0.7$, we have $\zeta_2' \approx 0.3483$. Taking results in (31), (33), and (34) together, choosing m/n exceeding some sufficiently large constant such that $c_0 \le c_5\epsilon^2$, and denoting $\zeta_2 := \zeta_2'(1+\gamma)/(2+\gamma)$, then with probability exceeding $1 - 4\mathrm{e}^{-c_0 m}$, the following

$$\left\langle \boldsymbol{h}, \frac{1}{2m}\nabla \ell_{\mathrm{tr}}(\boldsymbol{z}) \right\rangle \ge (1 - \zeta_1 - \zeta_2 - 2\epsilon) \left\|\boldsymbol{h}\right\|^2 \tag{35}$$

holds for all $\boldsymbol{x}$ and $\boldsymbol{z}$ such that $\|\boldsymbol{h}\| / \|\boldsymbol{x}\| \le \rho$ for $0 < \rho \le 1/10$ and any fixed $\gamma \ge 1/2$. This combining with (17) and (19) proves Proposition 2 for appropriately chosen $\mu > 0$ and $\lambda > 0$.

To conclude this section, an estimate for the working step size is provided next. To be specific, plugging the results in (23) and (25) into (18) suggests that

$$\begin{aligned} \mathrm{dist}^2\left(\boldsymbol{z} - \frac{\mu}{m}\nabla \ell_{\mathrm{tr}}(\boldsymbol{z}), \boldsymbol{x}\right) &= \left\|\boldsymbol{h}\right\|^2 - 2\mu \left\langle \boldsymbol{h}, \frac{1}{m}\nabla \ell_{\mathrm{tr}}(\boldsymbol{z}) \right\rangle + \left\| \frac{\mu}{m}\nabla \ell_{\mathrm{tr}}(\boldsymbol{z}) \right\|^2 \qquad (36) \\ &\le \left\{1 - \mu\left[4\left(1 - \xi_1 - \xi_2 - 2\epsilon\right) - \mu(1+\delta)^4\right]\right\} \left\|\boldsymbol{h}\right\|^2 \\ &\overset{\triangle}{=} (1 - \nu)\left\|\boldsymbol{h}\right\|^2. \qquad (37) \end{aligned}$$

Taking ϵ and δ to be sufficiently small, one obtains the feasible range of the step size for TGGF

$$\mu \le \frac{4\left(0.99 - \xi_1 - \xi_2\right)}{1.02} \overset{\triangle}{=} \mu_0, \tag{38}$$

thus concluding the proof of Theorem 1.

A Proof details

By homogeneity, it suffices to work with the case where $\|\boldsymbol{x}\| = 1$.

A.1 Proof of Lemma 1

It is easy to check that

$$\begin{aligned} \frac{1}{2}\left\|\boldsymbol{x}\boldsymbol{x}^{\mathcal{T}} - \tilde{\boldsymbol{z}}_0\boldsymbol{z}_0^{\mathcal{T}}\right\|_F^2 &= \frac{1}{2}\|\boldsymbol{x}\|^4 + \frac{1}{2}\|\tilde{\boldsymbol{z}}_0\|^4 - \left|\boldsymbol{x}^{\mathcal{T}}\tilde{\boldsymbol{z}}_0\right|^2 \\ &= 1 - \left|\boldsymbol{x}^{\mathcal{T}}\tilde{\boldsymbol{z}}_0\right|^2 \\ &= 1 - \cos^2\theta \end{aligned} \tag{39}$$

where $0 \le \theta \le \pi$ is the angle between the spaces spanned by $\boldsymbol{x}$ and $\tilde{\boldsymbol{z}}_0$. Then one can write

$$\boldsymbol{x} = \cos\theta\, \tilde{\boldsymbol{z}}_0 + \sin\theta\, \tilde{\boldsymbol{z}}_0^{\perp}, \tag{40}$$

where $\tilde{\boldsymbol{z}}_0^{\perp} \in \mathbb{R}^n$ is a unit vector that is orthogonal to $\tilde{\boldsymbol{z}}_0$ and has a nonnegative inner product with $\boldsymbol{x}$. Likewise, one can express

$$\boldsymbol{x}^{\perp} := -\sin\theta\, \tilde{\boldsymbol{z}}_0 + \cos\theta\, \tilde{\boldsymbol{z}}_0^{\perp}, \tag{41}$$

in which $\boldsymbol{x}^{\perp} \in \mathbb{R}^n$ is a unit vector orthogonal to $\boldsymbol{x}$.

Since $\tilde{\boldsymbol{z}}_0$ is the solution to the maximum eigenvalue problem

$$\tilde{\boldsymbol{z}}_0 := \arg\max_{\|\boldsymbol{z}\|=1} \quad \boldsymbol{z}^{\mathcal{T}}\overline{\boldsymbol{Y}}_0\boldsymbol{z} \tag{42}$$

for $\bar{\boldsymbol{Y}}_0 := \frac{1}{|\bar{\mathcal{I}}_0|}\bar{\boldsymbol{S}}_0^{\mathcal{T}}\bar{\boldsymbol{S}}_0$, it is the leading eigenvector of $\bar{\boldsymbol{Y}}_0$, i.e., $\bar{\boldsymbol{Y}}_0\tilde{\boldsymbol{z}}_0 = \lambda_1\tilde{\boldsymbol{z}}_0$, where $\lambda_1 > 0$ is the largest eigenvalue of $\bar{\boldsymbol{Y}}_0$. Premultiplying (40) and (41) by $\bar{\boldsymbol{S}}_0$ yields

$$\bar{\boldsymbol{S}}_0\boldsymbol{x} = \cos\theta\,\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0 + \sin\theta\,\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}, \tag{43a}$$

$$\bar{\boldsymbol{S}}_0\boldsymbol{x}^{\perp} = -\sin\theta\,\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0 + \cos\theta\,\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}. \tag{43b}$$

Pythagoras' relationship now gives

$$\big\|\bar{\boldsymbol{S}}_0\boldsymbol{x}\big\|^2 = \cos^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0\big\|^2 + \sin^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\big\|^2, \tag{44a}$$

$$\big\|\bar{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2 = \sin^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0\big\|^2 + \cos^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\big\|^2, \tag{44b}$$

where the cross-terms vanish because $\tilde{\boldsymbol{z}}_0^{\mathcal{T}}\bar{\boldsymbol{S}}_0^{\mathcal{T}}\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp} = |\bar{\mathcal{I}}_0|\tilde{\boldsymbol{z}}_0^{\mathcal{T}}\bar{\boldsymbol{Y}}_0\tilde{\boldsymbol{z}}_0^{\perp} = \lambda_1|\bar{\mathcal{I}}_0|\tilde{\boldsymbol{z}}_0^{\mathcal{T}}\tilde{\boldsymbol{z}}_0^{\perp} = 0$ following from the definition of $\tilde{\boldsymbol{z}}_0^{\perp}$.

We next construct the following expression

$$\begin{aligned}
&\sin^2\theta\big\|\bar{\boldsymbol{S}}_0\boldsymbol{x}\big\|^2 - \big\|\bar{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2 \\
&= \sin^2\theta\cos^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0\big\|^2 + \sin^4\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\big\|^2 - \sin^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0\big\|^2 - \cos^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\big\|^2 \\
&= \sin^2\theta\left(\cos^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0\big\|^2 - \big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0\big\|^2 + \sin^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\big\|^2\right) - \cos^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\big\|^2 \\
&= \sin^4\theta\big(\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\big\|^2 - \big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0\big\|^2\big) - \cos^2\theta\big\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\big\|^2 \\
&\le 0
\end{aligned}$$

where $\bar{\boldsymbol{S}}_0^{\mathcal{T}}\bar{\boldsymbol{S}}_0 \succeq \boldsymbol{0}$, so $\|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0^{\perp}\|^2 - \|\bar{\boldsymbol{S}}_0\tilde{\boldsymbol{z}}_0\|^2 \le 0$ holds for any unit vector $\tilde{\boldsymbol{z}}_0^{\perp} \in \mathbb{R}^n$ arising from the fact that $\tilde{\boldsymbol{z}}_0$ maximizes the term in (7), hence yielding

$$\sin^2\theta = 1 - \cos^2\theta \le \frac{\big\|\bar{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2}{\big\|\bar{\boldsymbol{S}}_0\boldsymbol{x}\big\|^2}. \tag{45}$$

Upon letting $\boldsymbol{u} = \boldsymbol{x}^{\perp}$, the last inequality taken together with (39) concludes the proof of (6).

A.2 Proof of Lemma 2

Recall that rows in $\bar{\boldsymbol{S}}_0 \in \mathbb{R}^{|\bar{\mathcal{I}}_0|\times n}$, hereafter denoted by $\boldsymbol{s}_i^{\mathcal{T}} \in \mathbb{R}^{1\times n}$, $\forall i \in [|\bar{\mathcal{I}}_0|]$, are drawn uniformly on the unit sphere. The uniformly spherical distribution is rotationally invariant, so it suffices to prove the results in the case where $\boldsymbol{x} = \boldsymbol{e}_1$ with $\boldsymbol{e}_1$ being the first canonical vector in $\mathbb{R}^n$. Indeed, any unit vector $\boldsymbol{x}$ can be expressed as $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{e}_1$ for some orthogonal transformation $\boldsymbol{U} \in \mathbb{R}^{n\times n}$. To see this, consider the following [10]

$$|\langle\boldsymbol{s}_i, \boldsymbol{x}\rangle|^2 = |\langle\boldsymbol{s}_i, \boldsymbol{U}\boldsymbol{e}_1\rangle|^2 = \big|\big\langle\boldsymbol{U}^{\mathcal{T}}\boldsymbol{s}_i, \boldsymbol{e}_1\big\rangle\big|^2 \overset{d}{=} |\langle\boldsymbol{s}_i, \boldsymbol{e}_1\rangle|^2, \tag{46}$$

where $\overset{d}{=}$ means terms involved on both sides of the equality have the same distribution. Thus, the problem of finding any unit-normed $\boldsymbol{x}$ is equivalent to that of finding $\boldsymbol{e}_1$. Henceforth, we assume without any loss of generality that $\boldsymbol{x} = \boldsymbol{e}_1$.

Considering a unit vector $\boldsymbol{x}^{\perp}$ such that $\boldsymbol{x}^{\mathcal{T}}\boldsymbol{x}^{\perp} = \boldsymbol{e}_1^{\mathcal{T}}\boldsymbol{x}^{\perp} = 0$, there exists a unit vector $\boldsymbol{d} \in \mathbb{R}^{n-1}$ such that $\boldsymbol{x}^{\perp} = \big[0\ \boldsymbol{d}^{\mathcal{T}}\big]^{\mathcal{T}}$. So it holds that

$$\big\|\bar{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\big\|^2 = \Big\|\bar{\boldsymbol{S}}_0\big[0\ \boldsymbol{d}^{\mathcal{T}}\big]^{\mathcal{T}}\Big\|^2 = \|\boldsymbol{F}\boldsymbol{d}\|^2, \tag{47}$$

where $\boldsymbol{F} \in \mathbb{R}^{|\overline{\mathcal{I}}_0|\times(n-1)}$ is obtained through deleting the first column in $\overline{\boldsymbol{S}}_0$, denoted by $\overline{\boldsymbol{S}}_{0,1}$, i.e., $\overline{\boldsymbol{S}}_0 = \left[\overline{\boldsymbol{S}}_{0,1} \ \boldsymbol{F}\right]$. Letting $\boldsymbol{F} := \left[\boldsymbol{f}_1 \ \cdots \ \boldsymbol{f}_{|\overline{\mathcal{I}}_0|}\right]^{\mathcal{T}}$, one can readily write $\boldsymbol{s}_i = \left[s_{i,1} \ \boldsymbol{f}_i^{\mathcal{T}}\right]^{\mathcal{T}}$, $\forall i \in [|\overline{\mathcal{I}}_0|] := \{1, \ldots, |\overline{\mathcal{I}}_0|\}$. Uniformly spherically distributed $\boldsymbol{s}_i \in \mathbb{R}^n$ has statistics $\mathbb{E}[\boldsymbol{s}_i] = \boldsymbol{0}$, and $\mathbb{E}[\boldsymbol{s}_i\boldsymbol{s}_i^{\mathcal{T}}] = \frac{1}{n}\boldsymbol{I}_n$ [11]. Leveraging the linearity of expectation operator, one arrives at

$$\mathbb{E}\left[\boldsymbol{s}_i\right] = \mathbb{E}\left[\begin{array}{c} s_{i,1} \\ \boldsymbol{f}_i \end{array}\right] = \left[\begin{array}{c} \mathbb{E}[s_{i,1}] \\ \mathbb{E}[\boldsymbol{f}_i] \end{array}\right] = \boldsymbol{0}, \ \forall i \tag{48}$$

to yield

$$\mathbb{E}[\boldsymbol{f}_i] = \boldsymbol{0}, \ \forall i. \tag{49}$$

A similar argument holds for the second-order moment

$$\mathbb{E}\left[\boldsymbol{s}_i\boldsymbol{s}_i^{\mathcal{T}}\right] = \left[\begin{array}{cc} \mathbb{E}[s_{i,1}^2] & \mathbb{E}[s_{i,1}\boldsymbol{f}_i^{\mathcal{T}}] \\ \mathbb{E}[s_{i,1}\boldsymbol{f}_i] & \mathbb{E}[\boldsymbol{f}_i\boldsymbol{f}_i^{\mathcal{T}}] \end{array}\right] = \frac{1}{n}\boldsymbol{I}_n, \ \forall i \tag{50}$$

leading to

$$\mathbb{E}\left[\boldsymbol{f}_i\boldsymbol{f}_i^{\mathcal{T}}\right] = \frac{1}{n}\boldsymbol{I}_{n-1}, \ \forall i. \tag{51}$$

Recall that a random vector $\boldsymbol{z} \in \mathbb{R}^n$ is said to be *isotropic* if it has zero-mean and identity covariance matrix [9, Definition 5.19]. Then recognize, from (49) and (51), that a proper scaling of $\boldsymbol{f}_i$ renders $\sqrt{n}\boldsymbol{f}_i$ isotropic. Further, it is known that a spherical random vector is subgaussian, and its subgaussian norm is bounded by an absolute constant [9]. Indeed, this comes from the following geometric argument: using rotational invariance of the uniform spherical distribution $\mathcal{S}^{n-1}$ in $\mathbb{R}^n$, it holds that, given any $\epsilon \geq 0$, the spherical cap $\{\boldsymbol{s}_i \in \mathcal{S}^{n-1} : s_{i,1} > \epsilon\}$ consists of at most $\mathrm{e}^{-\epsilon^2 n/2}$ proportion of the total area on the sphere. A similar argument carries over to $\boldsymbol{f}_i$, and thus, $\boldsymbol{f}_i$ is subgaussian as well.

Standard concentration inequalities on the sum of random positive semi-definite matrices composed of independent isotropic subgaussian rows [9, Remark 5.40] confirm that

$$\left\| \frac{1}{|\overline{\mathcal{I}}_0|}\left(\sqrt{n}\boldsymbol{F}\right)^{\mathcal{T}}\left(\sqrt{n}\boldsymbol{F}\right) - \boldsymbol{I}_{n-1} \right\| \leq \sigma \left\|\boldsymbol{I}_{n-1}\right\| \tag{52}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_K n}$ as long as $|\overline{\mathcal{I}}_0|/n$ is sufficiently large, where σ is a numerical constant that can take arbitrarily small values and $c_K > 0$ is a universal constant. Without loss of generality, let us work with $\sigma := 0.01$ in (52), so for any unit vector $\boldsymbol{d} \in \mathbb{R}^{n-1}$, the following inequality holds with probability at least $1 - 2\mathrm{e}^{-c_K n}$,

$$\left| \frac{n}{|\overline{\mathcal{I}}_0|}\boldsymbol{d}^{\mathcal{T}}\boldsymbol{F}^{\mathcal{T}}\boldsymbol{F}\boldsymbol{d} - \boldsymbol{d}^{\mathcal{T}}\boldsymbol{d} \right| \leq 0.01\boldsymbol{d}^{\mathcal{T}}\boldsymbol{d}, \tag{53}$$

or equivalently,

$$\|\boldsymbol{F}\boldsymbol{d}\|^2 = \left|\boldsymbol{d}^{\mathcal{T}}\boldsymbol{F}^{\mathcal{T}}\boldsymbol{F}\boldsymbol{d}\right| \leq 1.01|\overline{\mathcal{I}}_0|/n. \tag{54}$$

Combining the last with (47), one readily concludes that

$$\left\|\overline{\boldsymbol{S}}_0\boldsymbol{x}^{\perp}\right\|^2 \leq 1.01|\overline{\mathcal{I}}_0|/n \tag{55}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_K n}$, provided that $|\overline{\mathcal{I}}_0|/n$ exceeds some constant. Note that c_K depends on the maximum subgaussian norm of the rows of $\sqrt{n}\boldsymbol{F}$, and we assume without loss of generality $c_K \geq 1/2$. Hence, $\|\overline{\boldsymbol{S}}_0\boldsymbol{u}\|^2$ in (6) is upper bounded simply by letting $\boldsymbol{u} = \boldsymbol{x}^{\perp}$ in (55).

A.3 Proof of Lemma 3

We next pursue a meaningful lower bound for $\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\|^2$ in (9). When $\boldsymbol{x}=\boldsymbol{e}_1$, one has $\|\overline{\boldsymbol{S}}_0\boldsymbol{x}\|^2 = \|\overline{\boldsymbol{S}}_0\boldsymbol{e}_1\|^2 = \sum_{i=1}^{|\overline{\mathcal{I}}_0|} s_{i,1}^2$. It is further worth mentioning that all squared entries of any spherical random vector $\boldsymbol{s}_i$ obey the *Beta* distribution with parameters $\alpha=\frac{1}{2}$, and $\beta=\frac{n-1}{2}$, i.e., $s_{i,j}^2 \sim \text{Beta}\big(\frac{1}{2},\,\frac{n-1}{2}\big)$, $\forall i,\,j$, [11, Lemma 2]. Although they have closed-form probability density functions (pdfs) that may facilitate deriving a wanted lower bound, we shall take another easier route detailed as follows. A simple yet useful inequality is established first.

Lemma 7. *Given m fractions obeying $1>\frac{p_1}{q_1}\geq\frac{p_2}{q_2}\geq\cdots\geq\frac{p_m}{q_m}>0$, in which $p_i,\,q_i>0$, $\forall i\in[m]$, the following holds for all $1\leq k\leq m$*

$$\sum_{i=1}^{k}\frac{p_i}{q_i}\geq\sum_{i=1}^{k}\frac{p_{[i]}}{q_{[1]}} \tag{56}$$

where $p_{[i]}$ denotes the i-th largest one among $\{p_i\}_{i=1}^m$, and hence, $q_{[1]}$ is the maximum in $\{q_i\}_{i=1}^m$.

Proof. For any $k\in[m]$, according to the definition of $q_{[i]}$, it holds that $p_{[1]}\geq p_{[2]}\geq\cdots\geq p_{[k]}$, so $\frac{p_{[1]}}{q_{[1]}}\geq\frac{p_{[2]}}{q_{[1]}}\geq\cdots\geq\frac{p_{[k]}}{q_{[1]}}$. Considering $q_{[1]}\geq q_i$, $\forall i\in[m]$, and letting $j_i\in[m]$ be the index such that $p_{j_i}=p_{[i]}$, then $\frac{p_{j_i}}{q_{j_i}}=\frac{p_{[i]}}{q_{j_i}}\geq\frac{p_{[i]}}{q_{[1]}}$ holds for any $i\in[k]$. Therefore, $\sum_{i=1}^k\frac{p_{j_i}}{q_{j_i}}=\sum_{i=1}^k\frac{p_{[i]}}{q_{j_i}}\geq\sum_{i=1}^k\frac{p_{[i]}}{q_{[1]}}$. Note that $\left\{\frac{p_{[i]}}{q_{j_i}}\right\}_{i=1}^k$ comprise a subset of terms in $\left\{\frac{p_i}{q_i}\right\}_{i=1}^m$. On the other hand, according to our assumption, $\sum_{i=1}^k\frac{p_i}{q_i}$ is the largest among all sums of k summands; hence, $\sum_{i=1}^k\frac{p_i}{q_i}\geq\sum_{i=1}^k\frac{p_{[i]}}{q_{j_i}}$ yields $\sum_{i=1}^k\frac{p_i}{q_i}\geq\sum_{i=1}^k\frac{p_{[i]}}{q_{[1]}}$ concluding the proof. □

Without loss of generality and for simplicity of exposition, let us assume that indices of $\boldsymbol{a}_i$'s have been re-ordered such that

$$\frac{a_{1,1}^2}{\|\boldsymbol{a}_1\|^2}\geq\frac{a_{2,1}^2}{\|\boldsymbol{a}_2\|^2}\geq\cdots\geq\frac{a_{m,1}^2}{\|\boldsymbol{a}_m\|^2}, \tag{57}$$

where $a_{i,1}$ denotes the first element of $\boldsymbol{a}_i$. Therefore, writing $\|\overline{\boldsymbol{S}}_0\boldsymbol{e}_1\|^2=\sum_{i=1}^{|\overline{\mathcal{I}}_0|}a_{i,1}^2/\|\boldsymbol{a}_i\|^2$, the next task amounts to finding the sum of the $|\overline{\mathcal{I}}_0|$ largest out of all m entities in (57). Applying the result (56) in Lemma 7 gives

$$\sum_{i=1}^{|\overline{\mathcal{I}}_0|}\frac{a_{i,1}^2}{\|\boldsymbol{a}_i\|^2}\geq\sum_{i=1}^{|\overline{\mathcal{I}}_0|}\frac{a_{[i],1}^2}{\max_{i\in[m]}\|\boldsymbol{a}_i\|^2}, \tag{58}$$

in which $a_{[i],1}^2$ stands for the i-th largest entity in $\big\{a_{i,1}^2\big\}_{i=1}^m$.

Observe that for i.i.d. random vectors $\boldsymbol{a}_i\sim\mathcal{N}\big(\boldsymbol{0},\boldsymbol{I}_n\big)$, the property $\mathbb{P}(\|\boldsymbol{a}_i\|^2\geq 2.3n)\leq\mathrm{e}^{-n/2}$ holds for large enough n (e.g., $n\geq 20$), which can be understood upon substituting $\xi:=n/2$ into the following standard result [12, Lemma 1]

$$\mathbb{P}\left(\|\boldsymbol{a}_i\|^2-n\geq 2\sqrt{\xi}+2\xi\right)\leq\mathrm{e}^{-\xi}. \tag{59}$$

In addition, one readily concludes that $\mathbb{P}\left(\max_{i\in[m]}\|\boldsymbol{a}_i\|\leq\sqrt{2.3n}\right)\geq 1-m\mathrm{e}^{-n/2}$. We will henceforth build our subsequent proofs on this event without stating this explicitly each time encountering it.

Therefore, (58) can be lower bounded by

$$\left\|\overline{\boldsymbol{S}}\boldsymbol{x}\right\|^2 = \sum_{i=1}^{|\overline{\mathcal{I}}_0|} \frac{a_{i,1}^2}{\left\|\boldsymbol{a}_i\right\|^2} \geq \sum_{i=1}^{|\overline{\mathcal{I}}_0|} \frac{a_{[i],1}^2}{\max_{i\in[m]}\left\|\boldsymbol{a}_i\right\|^2} \geq \frac{1}{2.3n}\sum_{i=1}^{|\overline{\mathcal{I}}_0|}\left|a_{[i],1}\right|^2, \tag{60}$$

which holds with probability at least $1-m\mathrm{e}^{-n/2}$. The task left for bounding $\|\overline{\boldsymbol{S}}\boldsymbol{x}\|^2$ is to derive a meaningful lower bound for $\sum_{i=1}^{|\overline{\mathcal{I}}_0|} a_{[i],1}^2$. Roughly speaking, because the ratio $|\overline{\mathcal{I}}_0|/m$ is small, e.g., $|\overline{\mathcal{I}}_0|/m \leq 1/5$, a trivial result consists of bounding $(1/|\overline{\mathcal{I}}_0|)\sum_{i=1}^{|\overline{\mathcal{I}}_0|} a_{[i],1}^2$ by its sample average $(1/m)\sum_{i=1}^{m} a_{[i],1}^2$. The latter can be bounded using its ensemble mean, i.e., $\mathbb{E}[a_{i,1}^2]=1$, $\forall i \in [\overline{\mathcal{I}}_0]$, to yield $(1/m)\sum_{i=1}^{m} a_{[i],1}^2 \geq (1-\epsilon)\mathbb{E}[a_{i,1}^2] = 1-\epsilon$, which holds with high probability for some numerical constant $\epsilon > 0$ [10, Lemma 3.1]. Therefore, one has a candidate lower bound $\sum_{i=1}^{|\overline{\mathcal{I}}_0|} a_{[i],1}^2 \geq (1-\epsilon)|\overline{\mathcal{I}}_0|$. Nonetheless, this lower bound is in general too loose, and it contributes to a relatively large upper bound on the wanted term in (6).

To obtain an alternative bound, let us examine first the typical size of the maximum in $\left\{a_{i,1}^2\right\}_{i=1}^{m}$. Observe obviously that the modulus $|a_{i,1}|$ follows the half-normal distribution having the pdf $p(r) = \sqrt{2/\pi}\cdot \mathrm{e}^{-r^2/2}$, $r>0$, and it is easy to verify that

$$\mathbb{E}[|a_{i,1}|] = \sqrt{2/\pi}. \tag{61}$$

Then integrating the pdf from 0 to $+\infty$ yields the corresponding accumulative distribution function (cdf) expressible in terms of the error function $\mathbb{P}\left(|a_{i,1}|>\xi\right) = 1-\mathrm{erf}\left(\xi/2\right)$, i.e., $\mathrm{erf}\left(\xi\right) := 2/\sqrt{\pi}\cdot\int_0^{\xi}\mathrm{e}^{-r^2}\mathrm{d}r$. Appealing to a lower bound on the complimentary error function $\mathrm{erfc}\left(\xi\right) := 1-\mathrm{erf}\left(\xi\right)$ from [13, Theorem 2], one establishes that $\mathbb{P}\left(|a_{i,1}|>\xi\right) = 1-\mathrm{erf}\left(\xi/2\right) \geq (3/5)\mathrm{e}^{-\xi^2/2}$. Additionally, direct application of probability theory and Taylor expansion confirms that

$$\begin{aligned}\mathbb{P}\big(\max_{i\in[m]}|a_{i,1}|\geq\xi\big) &= 1-\left[\mathbb{P}\left(|a_{i,1}|\leq\xi\right)\right]^m\\ &\geq 1-\left(1-0.6\mathrm{e}^{-\xi^2/2}\right)^m\\ &\geq 1-\mathrm{e}^{-0.6m\mathrm{e}^{-\xi^2/2}}.\end{aligned} \tag{62}$$

Choosing now $\xi := \sqrt{2\log n}$ leads to

$$\mathbb{P}\big(\max_{i\in[m]}|a_{i,1}|\geq\sqrt{2\log n}\big) \geq 1-\mathrm{e}^{-0.6m/n} \geq 1-o(1) \tag{63}$$

which holds with the proviso that m/n is large enough, and the symbol $o(1)$ represents a small constant probability. Thus, provided that m/n exceeds some large constant, the event $\max_{i\in[m]} a_{i,1}^2 \geq 2\log n$ occurs with high probability. Hence, one may expect a tighter lower bound than $(1-\epsilon_0)|\overline{\mathcal{I}}_0|$, which is on the same order of m under the assumption that $|\overline{\mathcal{I}}_0|/m$ is about a constant.

Although $a_{i,1}^2$ obeys the *Chi-square* distribution with $k=1$ degrees of freedom, its cdf is rather complicated and does not admit a nice closed-form expression. A small trick is hence taken in the sequel. Postulate without loss of generality that both m and $|\overline{\mathcal{I}}_0|$ are even. Grouping two consecutive $a_{[i],1}^2$'s together, introduce a new variable $\vartheta[i] := a_{[2k-1],1}^2 + a_{[2k],1}^2$, $\forall k \in [m/2]$, hence yielding a sequence of ordered numbers, i.e., $\vartheta_{[1]} \geq \vartheta_{[2]} \geq \cdots \geq \vartheta_{[m/2]} > 0$. Then, one can equivalently write the wanted sum as

$$\sum_{i=1}^{|\overline{\mathcal{I}}_0|} a_{[i],1}^2 = \sum_{i=1}^{|\overline{\mathcal{I}}_0|/2} \vartheta_{[i]}. \tag{64}$$

On the other hand, for i.i.d. standard normal random variables $\{a_{i,1}\}_{i=1}^{m}$, let us consider grouping randomly two of them and denote the corresponding sum of their squares by $\chi_k := a_{k_i,1}^2 + a_{k_j,1}^2$, where $k_i \neq k_j \in [m]$, and $k \in [m/2]$. It is self-evident that the χ_k's are identically distributed obeying the *Chi-square* distribution with $k = 2$ degrees of freedom, having the pdf

$$p\left(r\right) = \frac{1}{2}\mathrm{e}^{-\frac{r}{2}}, \quad r \geq 0, \tag{65}$$

and the following complementary cdf (ccdf)

$$\mathbb{P}(\chi_k \geq \xi) := \int_{\xi}^{\infty} \frac{1}{2}\mathrm{e}^{-\frac{r}{2}}\mathrm{d}r = \mathrm{e}^{-\frac{\xi}{2}}, \quad \forall \xi \geq 0. \tag{66}$$

Ordering all χ_k's, summing the $|\overline{\mathcal{I}}_0|/2$ largest ones, and comparing the resultant sum with the one in (64) confirm that

$$\sum_{i=1}^{|\overline{\mathcal{I}}_0|/2} \chi_{[i]} \leq \sum_{i=1}^{|\overline{\mathcal{I}}_0|/2} \vartheta_{[i]} = \sum_{i=1}^{|\overline{\mathcal{I}}_0|} a_{[i],1}^2, \quad \forall |\overline{\mathcal{I}}_0| \in [m]. \tag{67}$$

Upon setting $\mathbb{P}(\chi_k \geq \xi) = |\overline{\mathcal{I}}_0|/m$, one obtains an estimate of $\chi_{|\overline{\mathcal{I}}_0|/2}$, the $(|\overline{\mathcal{I}}_0|/2)$-th largest value in $\{\chi_k\}_{k=1}^{m/2}$ as follows

$$\hat{\chi}_{|\overline{\mathcal{I}}_0|/2} := 2\log\left(m/|\overline{\mathcal{I}}_0|\right). \tag{68}$$

Furthermore, applying the Hoeffding-type inequality [9, Proposition 5.10] and leveraging the convexity of the ccdf in (66), one readily establishes that

$$\mathbb{P}\Big(\hat{\chi}_{|\overline{\mathcal{I}}_0|/2} - \chi_{|\overline{\mathcal{I}}_0|/2} > \xi\Big) \leq \mathrm{e}^{-\frac{1}{4}m\xi^2\mathrm{e}^{-\xi}(|\overline{\mathcal{I}}_0|/m)^2}, \quad \forall \xi > 0. \tag{69}$$

Taking without loss of generality $\xi := 0.05\hat{\chi}_{|\overline{\mathcal{I}}_0|/2} = 0.1\log\left(m/|\overline{\mathcal{I}}_0|\right)$ gives

$$\mathbb{P}\Big(\chi_{|\overline{\mathcal{I}}_0|/2} < 0.95\hat{\chi}_{|\overline{\mathcal{I}}_0|/2}\Big) \leq \mathrm{e}^{-c_0 m} \tag{70}$$

for some universal constants c_0, $c_\chi > 0$, and sufficiently large n such that $|\overline{\mathcal{I}}_0|/m \gtrsim c_\chi > 0$. The remaining part in this section assumes that this event occurs.

Choosing $\xi := 4\log n$ and substituting this into the ccdf in (66) leads to

$$\mathbb{P}\left(\chi \leq 4\log n\right) = 1 - 1/n^2. \tag{71}$$

Notice that each summand in $\sum_{i=1}^{|\overline{\mathcal{I}}_0|/2}\chi_{[i]} \geq \sum_{i=1}^{m/2}\chi_i\mathbb{1}_{\tilde{\mathcal{E}}_i}$ is Chi-square distributed, and hence could be unbounded, so we choose to work with the truncation $\sum_{i=1}^{m/2}\chi_i\mathbb{1}_{\tilde{\mathcal{E}}_i}$, where the $\mathbb{1}_{\tilde{\mathcal{E}}_i}$'s are independent copies of $\mathbb{1}_{\tilde{\mathcal{E}}}$, and $\mathbb{1}_{\tilde{\mathcal{E}}}$ denotes the indicator function for the ensuing events

$$\tilde{\mathcal{E}} := \Big\{\chi \geq \hat{\chi}_{|\overline{\mathcal{I}}_0|/2}\Big\} \cap \left\{\chi \leq 4\log n\right\}. \tag{72}$$

Apparently, it holds that $\sum_{i=1}^{|\overline{\mathcal{I}}_0|/2}\chi_{[i]} \geq \sum_{i=1}^{m/2}\chi_i\mathbb{1}_{\tilde{\mathcal{E}}_i}$. One further establishes that

$$\begin{aligned}
\mathbb{E}\left[\chi_i\mathbb{1}_{\tilde{\mathcal{E}}_i}\right] :&= \int_{\hat{\chi}_{|\overline{\mathcal{I}}_0|/2}}^{4\log n} \frac{1}{2}r\mathrm{e}^{-r/2}\mathrm{d}r \\
&= \left(\hat{\chi}_{|\overline{\mathcal{I}}_0|/2} + 2\right)\mathrm{e}^{-\hat{\chi}_{|\overline{\mathcal{I}}_0|/2}/2} - \left(4\log n + 2\right)\mathrm{e}^{-2\log n} \\
&= \frac{2|\overline{\mathcal{I}}_0|}{m}\Big[1 + \log\left(m/|\overline{\mathcal{I}}_0|\right)\Big] - \frac{(4\log n + 2)}{n^2}.
\end{aligned} \tag{73}$$

The task of bounding $\sum_{i=1}^{|\bar{\mathcal{I}}_0|} a_{[i],1}^2$ in (67) now boils down to bounding $\sum_{i=1}^{m/2} \chi_i \mathbb{1}_{\bar{\mathcal{E}}_i}$ from its expectation in (73). A convenient way to accomplish this is using the Bernstein inequality [9, Proposition 5.16], that deals with bounded random variables. That also justifies the reason of introducing the upper-bound truncation on χ in (72). Specifically, let us define

$$\vartheta_i := \chi_i \mathbb{1}_{\bar{\mathcal{E}}_i} - \mathbb{E}\left[\chi_i \mathbb{1}_{\bar{\mathcal{E}}_i}\right], \quad \forall i \in [m/2]. \tag{74}$$

Thus, $\{\vartheta_i\}_{i=1}^{m/2}$ are i.i.d. centered and bounded random variables following from the mean-subtraction and the upper-bound truncation. Further, according to the ccdf (66) and the definition of sub-exponential random variables [9, Definition 5.13], the terms $\{\vartheta_i\}_{i=1}^{m/2}$ are sub-exponential. Then, the following

$$\left|\sum_{i=1}^{m/2} \vartheta_i\right| \geq \tau \tag{75}$$

holds with probability at least $1 - 2\mathrm{e}^{-c_s \min\left(\tau/K_s, \tau^2/K_s^2\right)}$, in which $c_s > 0$ is a universal constant, and $K_s := \max_{i\in[m/2]} \|\vartheta_i\|_{\psi_1}$ represents the maximum subexponential norm of the ϑ_i's. Indeed, K_s can be found as follows [9, Definition 5.13]

$$\begin{aligned} K_s &:= \sup_{p\geq 1} p^{-1} \left(\mathbb{E}\left[|\vartheta_i|^p\right]\right)^{1/p} \\ &\leq \left(4\log n - 2\log\left(m/|\bar{\mathcal{I}}_0|\right)\right)\left[|\bar{\mathcal{I}}_0|/m - 1/n^2\right] \\ &\leq \frac{2|\bar{\mathcal{I}}_0|}{m} \log\left(n^2|\bar{\mathcal{I}}_0|/m\right) \\ &\leq \frac{4|\bar{\mathcal{I}}_0|}{m} \log n. \end{aligned} \tag{76}$$

Choosing $\tau := 8|\bar{\mathcal{I}}_0|/(c_s m) \cdot \log^2 n$ in (75) yields

$$\begin{aligned} \sum_{i=1}^{m/2} \chi_i \mathbb{1}_{\bar{\mathcal{E}}_i} &\geq |\bar{\mathcal{I}}_0|\left[1 + \log\left(m/|\bar{\mathcal{I}}_0|\right)\right] - 8|\bar{\mathcal{I}}_0|/(c_s m) \cdot \log^2 n - m\left(2\log n + 1\right)/n^2 \\ &\geq (1-\epsilon_s)|\bar{\mathcal{I}}_0|\left[1 + \log\left(m/|\bar{\mathcal{I}}_0|\right)\right] \end{aligned} \tag{77}$$

for some small constant $\epsilon_s > 0$, which holds with probability at least $1 - m\mathrm{e}^{-n/2} - \mathrm{e}^{-c_0 m} - 3/n^2$ as long as m/n exceeds some numerical constant and n is sufficiently large. Therefore, combining (60), (67), and (77), one concludes that the following holds with high probability

$$\left\|\bar{\boldsymbol{S}}_0 \boldsymbol{x}\right\|^2 = \sum_{i=1}^{|\bar{\mathcal{I}}_0|} \frac{a_{i,1}^2}{\|\boldsymbol{a}_i\|^2} \geq (1-\epsilon_s)\frac{|\bar{\mathcal{I}}_0|}{2.3n}\left[1 + \log\left(m/|\bar{\mathcal{I}}_0|\right)\right]. \tag{78}$$

Taking $\epsilon_s := 0.01$ without loss of generality concludes the proof of Lemma 3.

A.4 Proof of Lemma 5

Let us first prove the argument for a fixed pair $\boldsymbol{h}$ and $\boldsymbol{x}$, so $\boldsymbol{h}$ and $\boldsymbol{z}$ are independent of $\{\boldsymbol{a}_i\}_{i=1}^m$, and then apply a covering argument. To start, introduce a Lipschitz-continuous counterpart for the

discontinuous indicator function [3, A.2]

$$\chi_E(\theta) := \begin{cases} 1, & |\theta| \geq \frac{\sqrt{1.01}}{1+\gamma}, \\ 100(1+\gamma)^2\theta^2 - 100, & \frac{1}{1+\gamma} \leq |\theta| < \frac{\sqrt{1.01}}{1+\gamma}, \\ 0, & |\theta| < \frac{1}{1+\gamma} \end{cases} \tag{79}$$

with Lipschitz constant $\mathcal{O}(1)$. Recall that $\mathcal{E}_i := \left\{ \left| \frac{\boldsymbol{a}_i^\mathcal{T} \boldsymbol{z}}{\boldsymbol{a}_i^\mathcal{T} \boldsymbol{x}} \right| \geq \frac{1}{1+\gamma} \right\}$, so it holds that $0 \leq \chi_E\left(\left| \frac{\boldsymbol{a}_i^\mathcal{T} \boldsymbol{z}}{\boldsymbol{a}_i^\mathcal{T} \boldsymbol{x}} \right| \right) \leq \mathbb{1}_{\mathcal{E}_i}$ for any $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^n$, thus yielding

$$\frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}\right)^2 \mathbb{1}_{\mathcal{E}_i} \geq \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}\right)^2 \chi_E\left(\left| \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{z}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right) = \frac{1}{m}\sum_{i=1}^m \left(\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}\right)^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right). \tag{80}$$

By homogeneity and rotational invariance property of normal distributions, it suffices to prove the case where $\boldsymbol{x} = \boldsymbol{e}_1$ and $\|\boldsymbol{h}\|/\|\boldsymbol{x}\| = \|\boldsymbol{h}\| \leq \rho$. According to (80), lower bounding the first term in (31) can be achieved by lower bounding $\sum_{i=1}^m (\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h})^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right)$ instead. To that end, let us find the mean of $\left(\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}\right)^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right)$. Note that $\left(\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}\right)^2$ and $\chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right)$ are dependent. Introduce an orthonormal matrix $\boldsymbol{U_h}$ that contains $\boldsymbol{h}^\mathcal{T}/\|\boldsymbol{h}\|$ as its first row, i.e.,

$$\boldsymbol{U_h} := \begin{bmatrix} \boldsymbol{h}^\mathcal{T}/\|\boldsymbol{h}\| \\ \widetilde{\boldsymbol{U}}_{\boldsymbol{h}} \end{bmatrix} \tag{81}$$

for some orthogonal matrix $\widetilde{\boldsymbol{U}}_{\boldsymbol{h}} \in \mathbb{R}^{(n-1)\times n}$ such that $\boldsymbol{U_h}$ is orthonormal. Moreover, define $\tilde{\boldsymbol{h}} := \boldsymbol{U_h}\boldsymbol{h}$, and $\tilde{\boldsymbol{a}}_i := \boldsymbol{U_h}\boldsymbol{a}_i$; and let $\tilde{a}_{i,1}$ and $\tilde{\boldsymbol{a}}_{i,\setminus 1}$ denote the first entry and the remaining entries in vector $\tilde{\boldsymbol{a}}_i$; likewise for vector $\tilde{\boldsymbol{h}}$. Then, for any $\boldsymbol{h}$ such that $\|\boldsymbol{h}\| \leq \rho$, the next holds

$$\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}\right)^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right) \right] &= \mathbb{E}\left[\left(\tilde{a}_{i,1}\tilde{h}_1\right)^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right) \right] + \mathbb{E}\left[\left(\tilde{\boldsymbol{a}}_{i,\setminus 1}^\mathcal{T}\tilde{\boldsymbol{h}}_{\setminus 1}\right)^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right) \right] \\
&= \tilde{h}_1^2 \mathbb{E}\left[\tilde{a}_{i,1}^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{a_{i,1}} \right| \right) \right] + \mathbb{E}\left[\left(\tilde{\boldsymbol{a}}_{i,\setminus 1}^\mathcal{T}\tilde{\boldsymbol{h}}_{\setminus 1}\right)^2 \right] \mathbb{E}\left[\chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{a_{i,1}} \right| \right) \right] \\
&= \tilde{h}_1^2 \mathbb{E}\left[\tilde{a}_{i,1}^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{a_{i,1}} \right| \right) \right] + \left\|\tilde{\boldsymbol{h}}_{\setminus 1}\right\|^2 \mathbb{E}\left[\chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{a_{i,1}} \right| \right) \right] \\
&\geq \left(\tilde{h}_1^2 + \left\|\tilde{\boldsymbol{h}}_{\setminus 1}\right\|^2 \right) \min\left\{ \mathbb{E}\left[a_{i,1}^2 \chi_E\left(\left| 1 + h_1 + \frac{\boldsymbol{a}_{i,\setminus 1}^\mathcal{T}\boldsymbol{h}_{\setminus 1}}{a_{i,1}} \right| \right) \right], \right. \\
&\qquad \left. \mathbb{E}\left[\chi_E\left(\left| 1 + h_1 + \frac{\boldsymbol{a}_{i,\setminus 1}^\mathcal{T}\boldsymbol{h}_{\setminus 1}}{a_{i,1}} \right| \right) \right] \right\} \\
&\geq \|\boldsymbol{h}\|^2 \min\left\{ \mathbb{E}\left[a_{i,1}^2 \chi_E\left(\left| 1 - \rho + \frac{a_{i,2}}{a_{i,1}}\rho \right| \right) \right], \mathbb{E}\left[\chi_E\left(1 - \rho + \frac{a_{i,2}}{a_{i,1}}\rho \right) \right] \right\} \\
&= (1 - \zeta_1)\|\boldsymbol{h}\|^2
\end{aligned} \tag{82}$$

where the second equality follows from the independence between $\tilde{\boldsymbol{a}}_{i,\setminus 1}^\mathcal{T}\tilde{\boldsymbol{h}}_{\setminus 1}$ and $\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}$, the second inequality holds for $\rho \leq 1/10$ and $\gamma > 1/2$, and the last equality comes from the definition of ζ_1 in (74). Notice that $\varrho := (\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h})^2 \chi_E\left(\left| 1 + \frac{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h}}{\boldsymbol{a}_i^\mathcal{T}\boldsymbol{x}} \right| \right) \leq (\boldsymbol{a}_i^\mathcal{T}\boldsymbol{h})^2 \overset{d}{=} \|\boldsymbol{h}\|^2 a_{i,1}^2$ is a subexponential variable, and thus its subexponential norm $\|\varrho\|_{\psi_1} := \sup_{p \geq 1} \left[\mathbb{E}(|\varrho|^p)\right]^{1/p}$ is finite.

Direct application of the Berstein-type inequality [9, Proposition 5.16] confirms that for any $\epsilon > 0$, the following

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right) &\geq \mathbb{E}\left[\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)\right]-\epsilon\|\boldsymbol{h}\|^2\\ &\geq \left(1-\zeta_1-\epsilon\right)\|\boldsymbol{h}\|^2 \end{aligned} \tag{83}$$

holds with probability at least $1-\mathrm{e}^{-c_5 m\epsilon^2}$ for some numerical constant $c_5 > 0$ provided that $\epsilon \leq \|\varrho\|_{\psi_1}$ by assumption.

To obtain uniform control over all vectors $\boldsymbol{z}$ and $\boldsymbol{x}$ such that $\|\boldsymbol{z}-\boldsymbol{x}\| \leq \rho$, the net covering argument is applied [9, Definition 5.1]. Let $\mathcal{S}_\epsilon$ be an ϵ-net of the unit sphere, $\mathcal{L}_\epsilon$ be an ϵ-net of $[0,\, \rho]$, and define

$$\mathcal{N}_\epsilon := \left\{(\boldsymbol{z},\, \boldsymbol{h},\, t) : (\boldsymbol{z}_0,\, \boldsymbol{h}_0,\, t_0) \in \mathcal{S}_\epsilon \times \mathcal{S}_\epsilon \times \mathcal{L}_\epsilon\right\}. \tag{84}$$

Since the cardinality $|\mathcal{S}_\epsilon| \leq (1+2/\epsilon)^n$ [9, Lemma 5.2], then

$$|\mathcal{N}_\epsilon| \leq (1+2/\epsilon)^{2n}\,\rho/\epsilon \leq (1+2/\epsilon)^{2n+1} \tag{85}$$

due to the fact that $\rho/\epsilon < 2/\epsilon < 1+2/\epsilon$ for $0 \leq \rho < 1$.

Consider now any $(\boldsymbol{z},\, \boldsymbol{h},\, t)$ obeying $\|\boldsymbol{h}\| = t \leq \rho$. There exists a pair $(\boldsymbol{z}_0,\, \boldsymbol{h}_0,\, t_0) \in \mathcal{N}_\epsilon$ such that $\|\boldsymbol{z}-\boldsymbol{z}_0\|$, $\|\boldsymbol{h}-\boldsymbol{h}_0\|$, and $|t-t_0|$ are each at most ϵ. Taking the union bound yields

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right) &\geq \frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0\right)^2\chi_E\left(\left|1-t_0+\frac{a_{i,2}}{a_{i,1}}t_0\right|\right)\\ &\geq \left(1-\zeta_1-\epsilon\right)\|\boldsymbol{h}_0\|^2, \quad \forall\,(\boldsymbol{z}_0,\, \boldsymbol{h}_0,\, t_0) \in \mathcal{N}_\epsilon \end{aligned} \tag{86}$$

with probability at least $1-(1+2/\epsilon)^{2n+1}\,\mathrm{e}^{-c_5\epsilon^2 m} \geq 1-\mathrm{e}^{-c_0 m}$, which follows by choosing m such that $m \geq \left(c_6\cdot\epsilon^{-2}\log\epsilon^{-1}\right)n$ for some constant $c_6 > 0$.

Recall that $\chi_E\left(\tau\right)$ is Lipschitz continuous, thus

$$\begin{aligned}&\left|\frac{1}{m}\sum_{i=1}^{m}\left\{\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)-\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}^{\mathcal{T}}\boldsymbol{h}_0}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)\right\}\right|\\ &\quad\lesssim \frac{1}{m}\sum_{i=1}^{m}\left|\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2-\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}_0\right)^2\right|\\ &\quad= \frac{1}{m}\sum_{i=1}^{m}\left|\boldsymbol{a}_i^{\mathcal{T}}\left(\boldsymbol{h}\boldsymbol{h}^{\mathcal{T}}-\boldsymbol{h}_0\boldsymbol{h}_0^{\mathcal{T}}\right)\boldsymbol{a}_i\right|\\ &\quad\lesssim c_7\sum_{i=1}^{m}\left|\boldsymbol{h}\boldsymbol{h}^{\mathcal{T}}-\boldsymbol{h}_0\boldsymbol{h}_0^{\mathcal{T}}\right|\\ &\quad\leq 2.5c_7\,\|\boldsymbol{h}-\boldsymbol{h}_0\|\,\|\boldsymbol{h}\|\\ &\quad\leq 2.5c_7\rho\epsilon \end{aligned} \tag{87}$$

for some numerical constant c_7 and provided that $\epsilon < 1/2$ and $m \geq \left(c_6\cdot\epsilon^{-2}\log\epsilon^{-1}\right)n$, where the first inequality arises from the Lipschitz property of $\chi_E(\tau)$, the second uses the results in Lemma 1 in [3], and the third from Lemma 2 in [3].

Putting all results together confirms that with probability exceeding $1-2\mathrm{e}^{-c_0 m}$, the following holds

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2\chi_E\left(\left|1+\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)\geq\left[1-\zeta_1-\left(1+2.5c_7\rho\right)\epsilon\right]\|\boldsymbol{h}\|^2 \tag{88}$$

for all vectors $\|\boldsymbol{h}\|/\|\boldsymbol{x}\|\leq\rho$, concluding the proof.

A.5 Proof of Lemma 6

Similar to the proof in Section A.4, it is convenient to work with the following auxiliary function instead of the discontinuous indicator function

$$\chi_D(\theta):=\begin{cases}1, & |\theta|\geq\frac{2+\gamma}{1+\gamma}\\ -100\left(\frac{1+\gamma}{2+\gamma}\right)^2\theta^2+100, & \sqrt{0.99}\cdot\frac{2+\gamma}{1+\gamma}\leq|\theta|<\frac{2+\gamma}{1+\gamma}\\ 0, & |\theta|<\sqrt{0.99}\cdot\frac{2+\gamma}{1+\gamma}\end{cases} \tag{89}$$

which is Lipschitz continuous in θ with Lipschitz constant $\mathcal{O}(1)$. For $\mathcal{D}_i=\left\{\left|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\geq\frac{2+\gamma}{1+\gamma}\right\}$, it holds that $0\leq\mathbb{1}_{\mathcal{D}_i}\leq\chi_D\left(\left|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)$ for any $\boldsymbol{x}\in\mathbb{R}^n$ and $\boldsymbol{h}\in\mathbb{R}^n$. Assume without loss of generality $\boldsymbol{x}=\boldsymbol{e}_1$. Then for $\gamma>0$ and $\rho\leq 1/10$, it holds that

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\geq\frac{2+\gamma}{1+\gamma}\right\}}&\leq\frac{1}{m}\sum_{i=1}^{m}\chi_D\left(\left|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}}\right|\right)=\frac{1}{m}\sum_{i=1}^{m}\chi_D\left(\left|\frac{\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}}{a_{i,1}}\right|\right)\\&=\frac{1}{m}\sum_{i=1}^{m}\chi_D\left(\left|h_1+\frac{\boldsymbol{a}_{i,\setminus 1}^{\mathcal{T}}\boldsymbol{h}_{\setminus 1}}{a_{i,1}}\right|\right)\\&=\frac{1}{m}\sum_{i=1}^{m}\chi_D\left(\left|h_1+\frac{a_{i,2}}{a_{i,1}}\left\|\boldsymbol{h}_{\setminus 1}\right\|\right|\right)\\&\overset{\text{(i)}}{\leq}\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\left|h_1+\frac{a_{i,2}}{a_{i,1}}\|\boldsymbol{h}_{\setminus 1}\|\right|\geq\sqrt{0.99}\cdot\frac{2+\gamma}{1+\gamma}\right\}}\end{aligned} \tag{90}$$

where the last inequality arises from the definition of χ_D. Noting that $a_{i,2}/a_{i,1}$ obeys the standard Cauchy distribution, i.e., $a_{i,2}/a_{i,1}\sim\text{Cauchy}(0,1)$ [14], and particularly, transformation properties of Cauchy distributions assert that $h_1+\frac{a_{i,2}}{a_{i,1}}\|\boldsymbol{h}_{\setminus 1}\|\sim\text{Cauchy}(h_1,\|\boldsymbol{h}_{\setminus 1}\|)$ [15]. Recall that the cdf of a Cauchy distributed random variable $w\sim\text{Cauchy}\left(\mu_0,\alpha\right)$ is given by [14]

$$F(w;\mu_0,\alpha)=\frac{1}{\pi}\arctan\left(\frac{w-\mu_0}{\alpha}\right)+\frac{1}{2}. \tag{91}$$

It is easy to check that when $\|\boldsymbol{h}_{\setminus 1}\|=0$, the indicator function $\mathbb{1}_{\mathcal{D}_i}=0$ due to $|h_1|\leq\rho<\sqrt{0.99}(2+\gamma)/(1+\gamma)$. Consider only $\|\boldsymbol{h}_{\setminus 1}\|\neq 0$ next. Define for notational brevity $w:=a_{i,2}/a_{i,1}$,

$\alpha := \|\boldsymbol{h}_{\setminus 1}\|$, as well as $\mu_0 := h_1/\alpha$ and $w_0 := \sqrt{0.99}\frac{2+\gamma}{\alpha(1+\gamma)}$ to yield

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\{|\mu_0+w|\geq w_0\}}\right] &= 1 - \big[F(w_0;\mu_0,1) - F(-w_0;\mu_0,1)\big] \\
&= 1 - \frac{1}{\pi}\big[\arctan(w_0-\mu_0) - \arctan(-w_0-\mu_0)\big] \\
&\overset{\text{(i)}}{=} \frac{1}{\pi}\arctan\left(\frac{2w_0}{w_0^2-\mu_0^2-1}\right) \\
&\overset{\text{(ii)}}{\leq} \frac{1}{\pi}\cdot\frac{2w_0}{w_0^2-\mu_0^2-1} \\
&\overset{\text{(iii)}}{\leq} \frac{1}{\pi}\cdot\frac{2\sqrt{0.99}\rho(2+\gamma)/(1+\gamma)}{0.99(2+\gamma)^2/(1+\gamma)^2-\rho^2} \\
&\leq 0.0646
\end{aligned} \tag{92}$$

for all $\gamma > 0$ and $\rho \leq 1/10$. In deriving (i), the property $\arctan(u)+\arctan(v) = \arctan\left(\frac{u+v}{1-uv}\right) \pmod{\pi}$ for any $uv \neq 1$. Concerning (ii), the inequality $\arctan(x) \leq x$ for $x \geq 0$ is employed. Plugging given parameter values and using $\|\boldsymbol{h}_{\setminus 1}\| \leq \|\boldsymbol{h}\| \leq \rho$ confirms (iii). Apparently, $\mathbb{1}_{\{|\mu_0+w|\geq w_0\}}$ is bounded; and it is known that all bounded random variables are subexponential. Thus, upon applying the Bernstein-type inequality [9, Corollary 5.17], the next holds with probability at least $1-\mathrm{e}^{-c_5 m\epsilon^2}$ for some numerical constant $c_5 > 0$ and any sufficiently small $\epsilon > 0$

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\geq\frac{2+\gamma}{1+\gamma}\right\}} &\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\left|h_1+\frac{a_{i,2}}{a_{i,1}}\|\boldsymbol{h}_{\setminus 1}\|\right|\geq\sqrt{0.99}\frac{2+\gamma}{1+\gamma}\right\}} \leq (1+\epsilon)\mathbb{E}\left[\mathbb{1}_{\left\{\left|h_1+\frac{a_{i,2}}{a_{i,1}}\|\boldsymbol{h}_{\setminus 1}\|\right|\geq\sqrt{0.99}\frac{2+\gamma}{1+\gamma}\right\}}\right] \\
&\leq \frac{1+\epsilon}{\pi}\cdot\frac{2\sqrt{0.99}\rho(2+\gamma)/(1+\gamma)}{0.99(2+\gamma)^2/(1+\gamma)^2-\rho^2}.
\end{aligned} \tag{93}$$

On the other hand, one can easily establish that the following holds true for all $\boldsymbol{h}$

$$\mathbb{E}\left[\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^4\right] = \mathbb{E}\left[a_{i,1}^4\right]\|\boldsymbol{h}\|^4 = 3\|\boldsymbol{h}\|^4 \tag{94}$$

which has also been established in Lemma 1 [3] and Lemma 6.1 [16]. Further recalling our working assumption $\|\boldsymbol{a}_i\| \leq 2.3n$, then random variables $\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^4$ are bounded, and thus they are subexponential [9]. Appealing again to the Bernstein-type inequality for subexponential random variables and provided that $m/n > c_6\cdot\epsilon^{-2}\log\epsilon^{-1}$ for some numerical constant $c_6 > 0$, then

$$\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^4 \leq 3(1+\epsilon)\|\boldsymbol{h}\|^4 \tag{95}$$

which holds with probability exceeding $1-\mathrm{e}^{-c_5 m\epsilon^2}$ for some universal constant $c_5 > 0$ and any sufficiently small $\epsilon > 0$.

Collecting together results in (93) and (95) and leveraging the Cauchy-Schwartz inequality, one

establishes that for any $\rho \leq 1/10$ and $\gamma > 0$, the following

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^2 \mathbb{1}_{\mathcal{D}_i} &\leq \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}\right)^4}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\left\{\frac{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{h}|}{|\boldsymbol{a}_i^{\mathcal{T}}\boldsymbol{x}|}\geq\frac{2+\gamma}{1+\gamma}\right\}}} \\ &\leq \sqrt{3(1+\epsilon)\left\|\boldsymbol{h}\right\|^4}\sqrt{\frac{1+\epsilon}{\pi}\cdot\frac{2\sqrt{0.99}\rho(2+\gamma)/(1+\gamma)}{0.99(2+\gamma)^2/(1+\gamma)^2-\rho^2}} \\ &\triangleq (\zeta_2'+\epsilon')\left\|\boldsymbol{h}\right\|^2 \end{aligned} \tag{96}$$

where $\zeta_2' := 1.3785\sqrt{\rho\tau/(0.99\tau^2-\rho^2)}$ with $\tau := (2+\gamma)/(1+\gamma)$, which holds with probability at least $1-2\mathrm{e}^{-c_0 m}$. The latter arises if choosing $c_0 \leq c_5\epsilon^2$ in $1-2\mathrm{e}^{-c_5 m\epsilon^2}$, which can be accomplished by taking m/n sufficiently large.

References

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

[2] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, Apr. 2015.

[3] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Comm. Pure Appl. Math.*, 2016 (to appear).

[4] G. Wang and G. B. Giannakis, "Solving random systems of quadratic equations via truncated generalized gradient flow," in *Adv. Neural Inf. Process. Syst.*, Barcelona, Spain, 2016.

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

[6] P. Chen, A. Fannjiang, and G.-R. Liu, "Phase retrieval with one or two diffraction patterns by alternating projections of the null vector," *arXiv:1510.07379v2*, 2015.

[7] M. Soltanolkotabi, "Algorithms and theory for clustering and nonconvex quadratic programming," Ph.D. dissertation, Stanford University, 2014.

[8] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *arXiv:1605.08285*, 2016.

[9] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv:1011.3027*, 2010.

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

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

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

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

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

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

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