2016

Optimal Rates of Convergence for Noisy Sparse Phase Retrieval via Thresholded Wirtinger Flow

Tony Cai
University of Pennsylvania

Xiadong Li
University of California, Davis

Zongming Ma
University of Pennsylvania

Recommended Citation

Cai, T., Li, X., & Ma, Z. (2016). Optimal Rates of Convergence for Noisy Sparse Phase Retrieval via Thresholded Wirtinger Flow. *The Annals of Statistics, 44* (5), 2221-2251. http://dx.doi.org/10.1214/16-AOS1443

Optimal Rates of Convergence for Noisy Sparse Phase Retrieval via Thresholded Wirtinger Flow

Abstract

This paper considers the noisy sparse phase retrieval problem: recovering a sparse signal $\mathbf{x} \in \mathbb{R}^p$ from noisy quadratic measurements $y_j = (a_j'\mathbf{x})^2 + \varepsilon_j$, $j=1,\ldots,m$, with independent sub-exponential noise ε_j. The goals are to understand the effect of the sparsity of $\mathbf{x}$ on the estimation precision and to construct a computationally feasible estimator to achieve the optimal rates adaptively. Inspired by the Wirtinger Flow [*IEEE Trans. Inform. Theory* **61** (2015) 1985–2007] proposed for non-sparse and noiseless phase retrieval, a novel thresholded gradient descent algorithm is proposed and it is shown to adaptively achieve the minimax optimal rates of convergence over a wide range of sparsity levels when the $\mathbf{a}_j$'s are independent standard Gaussian random vectors, provided that the sample size is sufficiently large compared to the sparsity of $\mathbf{x}$.

Keywords

Iterative adaptive thresholding, minimax rate, non-convex empirical risk, phase retrieval, sparse recovery, thresholded gradient method

Disciplines

Physical Sciences and Mathematics

The Annals of Statistics
2016, Vol. 44, No. 5, 2221–2251
DOI: 10.1214/16-AOS1443

OPTIMAL RATES OF CONVERGENCE FOR NOISY SPARSE PHASE RETRIEVAL VIA THRESHOLDED WIRTINGER FLOW

BY T. TONY CAI[*,1], XIAODONG LI[†,2] AND ZONGMING MA[*,3]

University of Pennsylvania[*] and University of California, Davis[†]

This paper considers the noisy sparse phase retrieval problem: recovering a sparse signal $\mathbf{x} \in \mathbb{R}^p$ from noisy quadratic measurements $y_j = (\mathbf{a}_j'\mathbf{x})^2 + \varepsilon_j$, $j = 1, \ldots, m$, with independent sub-exponential noise ε_j. The goals are to understand the effect of the sparsity of $\mathbf{x}$ on the estimation precision and to construct a computationally feasible estimator to achieve the optimal rates adaptively. Inspired by the *Wirtinger Flow* [*IEEE Trans. Inform. Theory* **61** (2015) 1985–2007] proposed for non-sparse and noiseless phase retrieval, a novel thresholded gradient descent algorithm is proposed and it is shown to adaptively achieve the minimax optimal rates of convergence over a wide range of sparsity levels when the $\mathbf{a}_j$'s are independent standard Gaussian random vectors, provided that the sample size is sufficiently large compared to the sparsity of $\mathbf{x}$.

1. Introduction. In a range of fields in science and engineering, researchers face the problem of recovering a p-dimensional signal of interest $\mathbf{x}$ by probing the signal via a set of p-dimensional sensing vectors $\mathbf{a}_j$ for $j = 1, \ldots, m$, and hence the observations are the $(\mathbf{a}_j'\mathbf{x})$'s contaminated with noise. This gives rise to the linear regression model in statistical terminology where $\mathbf{x}$ is the regression coefficient vector and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]'$ is the design matrix. There is an extensive literature on the theory and methods for the estimation/recovery of $\mathbf{x}$ under such

Tribute: Peter Hall, a great scholar, a mentor, and a friend, passed away on January 9, 2016 at the age of 64. We are all very saddened by the loss of Peter and we are honored to have been asked to contribute a paper to this special issue in appreciation of this legendary statistician. We thank the co-editor, Runze Li, for the opportunity to offer a token of our affection and immense respect for Peter. With over 600 journal papers, Peter's scholarly achievements and tremendous productivity are well known. His exceptional intelligence was matched by his extensive service to the community and great personal warmth and generosity. TTC had the good fortune to have known him for 20 years, to have benefited from his wisdom, and to have collaborated with him. Peter was a constant source of inspiration and support. We dedicate this paper to the memory of Peter, who we sorely miss.

Received June 2015; revised January 2016.

[1]Supported in part by NSF Grants DMS-12-08982 and DMS-14-03708, and NIH Grant R01 CA127334.

[2]Supported by NIH Grant R01 CA127334.

[3]Supported in part by NSF CAREER Grant DMS-13-52060.

MSC2010 subject classifications. Primary 62C20; secondary 62P35.

Key words and phrases. Iterative adaptive thresholding, minimax rate, non-convex empirical risk, phase retrieval, sparse recovery, thresholded gradient method.

a linear model. However, in many important applications, including X-ray crystallography, microscopy, astronomy, diffraction and array imaging, interferometry and quantum information, it is sometimes impossible to observe $\mathbf{a}_j'\mathbf{x}$ directly and the measurement that one is able to obtain is the magnitude/energy of $\mathbf{a}_j'\mathbf{x}$ contaminated with noise. In other words, the observations are generated by the following *phase retrieval* model:

$$y_j = |\mathbf{a}_j'\mathbf{x}|^2 + \varepsilon_j, \qquad j = 1, \ldots, m, \tag{1.1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)'$ is a vector of stochastic noise with $\mathbb{E}\boldsymbol{\varepsilon} = \mathbf{0}$. Note that $\mathbb{E}(y_j) = |\mathbf{a}_j'\mathbf{x}|^2$, so in the real case, (1.1) can be treated as a generalized linear model with the multi-value link function $g(z) := \pm\sqrt{z}$. We refer interested readers to [41] and the reference therein for more detailed discussions on the scientific and engineering background for this model.

In many applications, especially those related to imaging, the signal $\mathbf{x} \in \mathbb{R}^p$ admits a sparse representation under some known and deterministic linear transformation. Without loss of generality, we assume in the rest of the paper that such a linear transform has already taken place, and hence the signal $\mathbf{x}$ is sparse itself. In this case, model (1.1) is referred to as the *sparse phase retrieval* model. In addition, we consider the case where $\boldsymbol{\varepsilon}$ are independent centered sub-exponential random errors. This is motivated by the observation that in the application settings where model (1.1) is appropriate, especially in optics, heavy-tailed noise may arise due to photon counting.

Efficient computational methods for phase retrieval have been proposed in the community of optics, and they are mostly based on the seminal work by Gerchberg, Saxton, and Fienup [19, 21]. The effectiveness of these methods relies on careful exploration of prior information of the signal in the spatial domain. Moreover, these methods were revealed later as non-convex successive projection algorithms [4, 30]. This provides insight for the occasional observation of stagnation of iterates and failure of convergence.

Recently, in view of multiple illumination with masks, novel computational methods were proposed for phase retrieval without exploring or employing a priori information of the signal. These methods include semidefinite programming [9, 12–14, 44], polarization [2], alternating minimization [37], gradient methods [11], alternating projection [35], etc. More importantly, elegant and remarkable theoretical guarantees for these methods have also been established. As for noiseless sparse phase retrieval, semidefinite programming has been proven to be effective with theoretical guarantees [22, 31, 38]. Other empirical methods for sparse phase retrieval include belief propagation [39] and greedy methods [40].

Regarding noisy phase retrieval, some stability results have been established in the literature; see [10, 15, 42]. In particular, stability results have been obtained in [16] for noisy sparse phase retrieval by semidefinite programming, though the authors did not study the optimal estimation error rates with respect to either the

sparsity level of the signal or the sample size. Nearly minimax convergence rates for sparse phase retrieval with Gaussian noise have been given in [28] under sub-Gaussian design matrices. However, the optimal rates are achieved by empirical risk minimization under sparsity constraints, in which both the objective function and the constraint are non-convex, implying that the procedure is not computationally feasible.

In the present paper, we establish the minimax optimal rates of convergence for noisy sparse phase retrieval under sub-exponential noise, and propose a novel thresholded gradient descent method in order to estimate the signal $\mathbf{x}$ under the model (1.1). For conciseness, we focus on the case where the signal and the sensing vectors are all real-valued, and the key ideas extend naturally to the complex case. The theoretical analysis sheds light on the effects of the sparsity of the signal $\mathbf{x}$ and the presence of sub-exponential noise on the minimax rates for the estimation of $\mathbf{x}$ under the ℓ_2 loss, as long as the sensing vectors $\mathbf{a}_j$'s are independent standard Gaussian vectors. Combining the minimax upper and lower bounds given in Section 3, the optimal rate of convergence for estimating the signal $\mathbf{x}$ under the ℓ_2 loss is $\frac{\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}$, where k is the sparsity of $\mathbf{x}$, $\|\cdot\|_2$ is the usual Euclidean norm, and σ characterizes the noise level. Moreover, it is shown that the thresholded gradient descent procedure is both rate-optimal and computationally fast, and the sample size requirement matches the state-of-the-art result in computational sparse phase retrieval under structureless Gaussian design matrices.

We now explain some notation used throughout the paper. For any n-dimensional vector $\mathbf{v} = (v_1, \ldots, v_n)'$ and a subset $S \subset \{1, \ldots, n\}$, we denote by $\mathbf{v}_S$ the n-dimensional vector by keeping the coordinates of $\mathbf{v}$ with indices in S unchanged, while changing all other components to zero. We denote $\|\mathbf{v}\|_q := (v_1^q + \cdots + v_n^q)^{1/q}$ for $q \geq 1$, and $\|\mathbf{v}\|_\infty = \max_{1\leq k\leq n} |v_k|$. Also denote $\|\mathbf{v}\|_0$ as the number of nonzero components of $\mathbf{v}$. For any matrix $\mathbf{M} \in \mathbb{R}^{n_1\times n_2}$, and any subsets $S_1 \in \{1, \ldots, n_1\}$ and $S_2 \in \{1, \ldots, n_2\}$, $\mathbf{M}_{S_1S_2} \in \mathbb{R}^{n_1\times n_2}$ is defined by keeping the submatrix of $\mathbf{M}$ with row index set S_1 and column index set S_2, while changing all other entries to zero. For any $q_1 \geq 1$ and $q_2 \geq 1$, we denote $\|\mathbf{M}\|_{q_2\to q_1}$ the induced operator norm from the Banach space $(\mathbb{R}^{n_2}, \|\cdot\|_{q_2})$ to $(\mathbb{R}^{n_1}, \|\cdot\|_{q_1})$. For simplicity, denote $\|\mathbf{M}\| := \|\mathbf{M}\|_{2\to 2}$. We denote by $\mathbf{I}_n$ the $n \times n$ identity matrix. For any two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if a_n/b_n is uniformly bounded by some absolute constant. We further write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold.

The rest of the paper is organized as follows: In Section 2, we introduce in detail our sparse phase retrieval procedure, which consists of two steps. The first is an initialization step by applying a diagonal thresholding method to a matrix constructed with available data. The second step, that is, thresholded Wirtinger flow (TWF), applies iteratively thresholded gradient descent for the recovery of the sparse vector $\mathbf{x}$. Section 3 establishes the minimax optimal rates of convergence for noisy sparse phase retrieval under the ℓ_2 loss. The results show that the proposed thresholded gradient descent method is rate-optimal. In Section 4, numerical simulations

illustrate the effectiveness of thresholding in denoising, and demonstrate how the relative estimation error depends on the thresholding parameter β. In particular, by letting $\beta = 0$, TWF reduces to Wirtinger flow (WF) iterate proposed in [11], and we show numerically that TWF is much more accurate than WF in estimation. Moreover, we verify the minimax rate established in Section 3 by studying how the estimation error depends on the sample size m, sparsity k, and the noise-to-signal ratio $\sigma/\|\mathbf{x}\|_2^2$. In Section 5, we discuss the connections between our thresholded gradient method for noisy sparse phase retrieval and related methods proposed in the literature for high-dimensional regression. The proofs are given in Section 6 with some technical details deferred to the Appendix.

2. Methodology. The major component of the our method is a thresholded gradient descent algorithm to obtain a sparse solution to a given non-convex empirical risk minimization problem. Due to the non-convex and sparse nature of the problem, in order to avoid any local optimum or non-sparse solution that is far away from the truth; the initialization step is crucial. Thus, we also provide a candidate method which can be justified theoretically for yielding a good initializer. The methodology is proposed assuming that $\mathbf{A}$ has standard Gaussian entries, though it could potentially also be used when such an assumption does not necessarily hold.

2.1. *Thresholded Wirtinger flow.* Given the sensing vectors $\mathbf{a}_j$ and the noisy magnitude measurements y_j as in (1.1) for $j = 1, \ldots, m$, waiving the sparse assumption aside, one can consider estimating $\mathbf{x}$ by minimizing the following empirical risk function

$$f(\mathbf{z}) := \frac{1}{4m} \sum_{j=1}^{m} \big(|\mathbf{a}_j' \mathbf{z}|^2 - y_j\big)^2. \tag{2.1}$$

Statistically speaking, in the low-dimensional setup with fixed p and $m \to \infty$, if the additive noises are heavy-tailed, least-absolute-deviations (LAD) methods might be more robust than least-squares methods. However, recent progress in modern linear regression analysis shows that least-squares could be preferable to LAD when p and m are proportional, even the noises are sub-exponential [18]. Due to this surprising phenomenon, we simply take the least-squares empirical risk in (2.1), although phase retrieval is a nonlinear regression problem rather than linear regression discussed in [18]. More importantly, close-form gradient methods can be induced from the empirical risk function in (2.1), which is computationally convenient. To be specific, at any current value of $\mathbf{z}$, the estimator is updated by taking a step along the gradient direction

$$\nabla f(\mathbf{z}) = \frac{1}{m} \sum_{j=1}^{m} \big(|\mathbf{a}_j' \mathbf{z}|^2 - y_j\big)(\mathbf{a}_j' \mathbf{z})\mathbf{a}_j \tag{2.2}$$

until a stationary point is reached. Indeed, [11] showed that under appropriate conditions, initialized by an appropriate spectral method, a gradient method, referred to as Wirtinger flow, leads to accurate recovery of $\mathbf{x}$ up to a global phase in the complex domain and noiseless setting.

However, the direct application of gradient descent is not ideal for noisy sparse phase retrieval since it does not utilize the knowledge that the true signal $\mathbf{x}$ is sparse in order to mitigate the contamination of the noise. To incorporate this a priori knowledge, it makes sense to seek a "sparse minimizer" of (2.1). To this end, suppose we have a sparse initial guess $\mathbf{x}^{(0)}$ for $\mathbf{x}$. To update $\mathbf{x}^{(0)}$ to another sparse vector, we may take a step along $\nabla f(\mathbf{x}^{(0)})$, and then sparsify the result by thresholding.

Indeed, if we were given the oracle knowledge of the support S of $\mathbf{x}$, then we can reduce the problem to recovering $\mathbf{x}_S$ based on the $\{y_j, a_{jS}\}_{j=1}^m$. By avoiding estimating any coordinate of $\mathbf{x}$ in S^c, we could greatly reduce variance of the resulting estimator of $\mathbf{x}$. In reality, we do not have such oracle knowledge and the additional thresholding step added on top of gradient descent is intended to mimic the oracle behavior by hopefully restricting all the updated coordinates on S.

Let $\mathcal{T}_\tau$ be any thresholding function satisfying

$$\mathcal{T}_\tau(x) = 0 \qquad \forall x \in [-\tau, \tau], \quad \text{and} \quad \big|\mathcal{T}_\tau(x) - x\big| \le \tau \qquad \forall x \in \mathbb{R}. \tag{2.3}$$

For any vector $\mathbf{b} = (b_1, \dots, b_p)'$, let $\mathcal{T}_\tau(\mathbf{b}) = (\mathcal{T}_\tau(b_1), \dots, \mathcal{T}_\tau(b_p))'$. With the foregoing definition, the proposed thresholded gradient descent method can be summarized as Algorithm 1. In view of the Wirtinger flow method for noiseless phase retrieval [11], we name our approach the "Thresholded Wirtinger Flow" (TWF) method. The data-driven choice of the threshold level in (2.5) is motivated by the following intuition. Assume that the sensing vectors $\{\mathbf{a}_j : j = 1, \dots, m\}$ are independent standard Gaussian vectors. For a fixed $\mathbf{z}$, if we act as if each $(|\mathbf{a}_j'\mathbf{z}|^2 - y_j)(\mathbf{a}_j'\mathbf{z})$ is a fixed constant, then the gradient in (2.2) is a linear combination of Gaussian vectors and hence has i.i.d. Gaussian entries with mean zero and variance $\frac{1}{m^2}\sum_{j=1}^m (|\mathbf{a}_j'\mathbf{z}|^2 - y_j)^2(\mathbf{a}_j'\mathbf{z})^2$. Therefore, the threshold $\tau(\mathbf{z})$ is simply $\sqrt{\beta \log(mp)}$ times the standard deviation of these Gaussian random variables, which is essentially the universal thresholding in the Gaussian sequence model literature [24]. The above intuition is not exactly true, since in each iterate $\mathbf{z} = \widehat{\mathbf{x}}^{(n)}$ depends on $\mathbf{A}$ hence not fixed. However, the resulting thresholds in (2.5) are indeed the right choices as justified by Theorem 3.1 in Section 3, and illustrated numerically in Section 4. Notice that there are two tuning parameters μ and β, which should be treated as absolute constants. We will validate some theoretical choices and also provide practical choices in Sections 3 and 4, respectively. Although the algorithm is motivated by assuming standard Gaussian sensing vectors, its applicability might extend to more general cases.

Algorithm 1: Thresholded Wirtinger flow for noisy sparse phase retrieval

Input: Data $\{\mathbf{a}_j, y_j\}_{j=1}^m$; initial estimator $\widehat{\mathbf{x}}_0$; thresholding function $\mathcal{T}$; gradient tuning parameter μ; thresholding tuning parameter β; number of iterations T.

Output: Final estimator $\widehat{\mathbf{x}}$.

Initialize $n \leftarrow 0$, $\widehat{\mathbf{x}}^{(0)} = \widehat{\mathbf{x}}_0$ and

$$\phi^2 = \frac{1}{m}\sum_{j=1}^m y_j \tag{2.4}$$

repeat

Compute threshold level

$$\tau\big(\widehat{\mathbf{x}}^{(n)}\big) = \sqrt{\frac{\beta \log(mp)}{m^2}\sum_{j=1}^m \big(|\mathbf{a}_j'\widehat{\mathbf{x}}^{(n)}|^2 - y_j\big)^2 |\mathbf{a}_j'\widehat{\mathbf{x}}^{(n)}|^2}; \tag{2.5}$$

Update

$$\widehat{\mathbf{x}}^{(n+1)} = \varphi\big(\widehat{\mathbf{x}}^{(n)}\big) := \mathcal{T}_{(\mu/\phi^2)\tau(\widehat{\mathbf{x}}^{(n)})}\Big(\widehat{\mathbf{x}}^{(n)} - \frac{\mu}{\phi^2}\nabla f\big(\widehat{\mathbf{x}}^{(n)}\big)\Big), \tag{2.6}$$

until $n = T$;

where ∇f is defined in (2.2);

Return $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}^{(T)}$.

2.2. *Initialization.* It is worth noting that the success of Algorithm 1 depends crucially on the initial estimator for two reasons. First, the empirical risk (2.1) is a non-convex function of $\mathbf{z}$, and hence the success of a gradient descent based approach depends naturally on the starting point. Moreover, an accurate initializer can reduce the required number of iterations in the thresholded Wirtinger flow algorithm. In view of its crucial rule, we propose in Algorithm 2 an initialization method which can be proven to yield a decent starting point for Algorithm 1 under our modeling assumption.

The motivation of the algorithm is similar to that of diagonal thresholding [25] for sparse PCA: we want to identify a small collection of coordinates with big marginal signals and then compute an estimator of $\mathbf{x}$ by focusing only on these coordinates. In particular, the quantity I_l in (2.7) captures the marginal signal strength of the lth coordinate and $\widehat{S}_0$ (2.8) selects all coordinates with big marginal signals. Last but not least, (2.9) and (2.10) computes the initial estimator by focusing only on the coordinates in $\widehat{S}_0$. There is a tuning parameter α needed as input of the algorithm, which can be treated as an absolute constant. We will provide some justified theoretical choice later.

Algorithm 2: Initialization for Algorithm 1

Input: Data $\{\mathbf{a}_j, y_j\}_{j=1}^m$; tuning parameter α.
Output: Initial estimator $\widehat{\mathbf{x}}_0$.

1 Compute

$$I_l = \frac{1}{m}\sum_{j=1}^{m} y_j a_{jl}^2, \qquad l = 1, \ldots, p. \tag{2.7}$$

2 Let ϕ^2 be defined as in (2.4). Select a set of coordinates

$$\widehat{S}_0 = \left\{ l \in [p] : I_l > \left(1 + \alpha\sqrt{\frac{\log(mp)}{m}}\right)\phi^2 \right\}. \tag{2.8}$$

3 Compute a $p \times p$ matrix

$$\mathbf{W}_{\widehat{S}_0\widehat{S}_0} := \frac{1}{m}\sum_{j=1}^{m} y_j \mathbf{a}_{j\widehat{S}_0}\mathbf{a}'_{j\widehat{S}_0}. \tag{2.9}$$

4 Return

$$\widehat{\mathbf{x}}_0 = \phi\widehat{\mathbf{v}}_1, \tag{2.10}$$

where $\widehat{\mathbf{v}}_1$ as the leading eigenvector of $\mathbf{W}_{\widehat{S}_0\widehat{S}_0}$.

3. Theory. We first establish the statistical convergence rate for the thresholded Wirtinger flow method under the case of "Gaussian design", that is, $\mathbf{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ for $j = 1, \ldots, m$ in (1.1). Moreover, we assume the signal $\mathbf{x}$ is k-sparse, that is, $\|\mathbf{x}\|_0 = k$, and the noises $\varepsilon_1, \ldots, \varepsilon_m$ are m independent centered sub-exponential random variables with maximum sub-exponential norm σ, that is, $\sigma := \max_{1\le i\le m} \|\varepsilon_i\|_{\psi_1}$. Here, for any random variable X, its sub-exponential norm is defined as $\|X\|_{\psi_1} := \sup_{p\ge 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$. This definition, as well as some fundamental properties of sub-exponential variables (such as Bernstein inequality), can be found in Section 5.2.4 of [43].

THEOREM 3.1. *Suppose $\beta = 4$ in (2.5) and $\mu \le \mu_0$ in (2.6) for some absolute constant μ_0. If*

$$\alpha \ge K\left(1 + \frac{\sigma}{\|\mathbf{x}\|_2^2}\right) \tag{3.1}$$

in (2.8) for some absolute constant $K > 0$ and

$$m \ge C\alpha^2 k^2 \log(mp) \tag{3.2}$$

for some absolute constant $C > 0$, then

$$\sup_{\|\mathbf{x}\|_0=k} \mathbb{P}_{(\mathbf{A},\mathbf{y}|\mathbf{x})}\Big(\min_{i=0,1}\big\|\widehat{\mathbf{x}}^{(t)} - (-1)^i\mathbf{x}\big\|_2 > \frac{1}{6}\Big(1-\frac{\mu}{16}\Big)^t\|\mathbf{x}\|_2 + \frac{C_0\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}\Big) \le \frac{46}{m} + \frac{10}{e^k} + \frac{t}{mp^2} \tag{3.3}$$

for some absolute constant $C_0 > 0$.

When $\sigma/\|\mathbf{x}\|_2^2 = o(\sqrt{m/\log m})$ and is unknown, and the noises are i.i.d. Gaussian with variance σ^2 in (1.1), we can estimate $\|\mathbf{x}\|_2^2$ by ϕ^2 in (2.4) and define $\hat{\sigma} = \sqrt{(\frac{1}{m}\sum_{j=1}^m y_j^2 - 3\phi^4)_+}$. Then with probability at least $1 - 1/m$, there holds $1 + \frac{\hat{\sigma}}{\phi^2} \asymp 1 + \frac{\sigma}{\|\mathbf{x}\|_2^2}$. Set $\alpha = K(1 + \hat{\sigma}/\phi^2)$ for some absolute constant $K > 0$. Then the claim (3.3) continues to hold with the first term on the right side $46/m$ replaced by $47/m$.

The proof is given in Section 6, where Lemma 6.3 guarantees the efficacy of the initialization step Algorithm 2, and Lemmas 6.4 and 6.5 explain why the thresholded Wirtinger flow method leads to accurate estimation.

There are three conditions in the theorem concerning the three tuning parameters: the thresholding parameter β, gradient step size parameter μ and the initialization thresholding parameter α. For β, although theoretically we let $\beta = 4$, Section 4 shows that choosing $\beta \le 1$ usually yields the smallest estimation error for $\mathbf{x}$ empirically. For μ, the condition $\mu \le \mu_0$ implies that μ should be chosen conservatively as a small constant in each iteration, and in Section 4 we follow this principle to choose $\mu = 0.01$. For the initialization parameter α, the condition (3.1) is relatively strong though it is essential for the effectiveness of Algorithm 2. However, as we will see in Section 4, the final TWF estimator $\widehat{\mathbf{x}}$ is not sensitive to the choice of α. When the noises are i.i.d. Gaussian with variance σ^2 in (1.1), the proposed estimators for $\|\mathbf{x}\|_2^2$ and σ in Theorem 3.1 are motivated by the observations that $\mathbb{E}[y_j] = \|\mathbf{x}\|_2^2$ and $\mathrm{Var}(y_j) = 2\|\mathbf{x}\|_2^4 + \sigma^2$.

In the noiseless case where $\sigma = 0$, with high probability, we obtain $\min_{i=0,1}\|\widehat{\mathbf{x}}^{(t)} - (-1)^i\mathbf{x}\|_2 \le \frac{1}{6}(1-\frac{\mu}{16})^t\|\mathbf{x}\|_2$. This implies that thresholded gradient descent method leads to linear convergence to the original signal up to a global sign. It shows explicitly that the smaller μ is, the more slowly thresholded Wirtinger flow converges.

In the noisy case, if $\mu > 0$ is an absolute constant, by letting $t \asymp \log(1/\delta)$ where $\delta = \frac{\sigma}{\|\mathbf{x}\|_2^2}\sqrt{\frac{k\log p}{m}}$, we obtain $\min_{i=0,1}\|\widehat{\mathbf{x}}^{(t)} - (-1)^i\mathbf{x}\|_2 \precsim \frac{\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}$ with high probability. If the knowledge of δ is not available, by choosing $t = O(\log p)$, we can obtain $\min_{i=0,1}\|\widehat{\mathbf{x}}^{(t)} - (-1)^i\mathbf{x}\|_2 \precsim \frac{\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}} + \frac{1}{p^c}$ for any predetermined $c > 0$. The convergence rate $\frac{\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}$ is better than the upper bound result es-

tablished in [28], which is achieved by minimizing (2.1) under the sparsity constraint. Notice that the procedure in [28] is almost minimax. Consequently, our thresholded Wirtinger flow method leads to an estimator close to the global minimizer of the non-convex objective function (2.1) over the set of k-sparse vectors.

Ignoring any polylog factor, the above convenient properties of thresholded Wirtinger flow are guaranteed by the sample size condition $m \gtrsim k^2$. When $m \ll p$, this condition is crucial for the effectiveness of initialization Algorithm 2. An immediate question is whether such a minimum sample size condition is in some sense necessary for any computationally efficient algorithm, if the sensing matrix is random and structureless?[4] A similar phenomenon has been previously observed in the related but different problem of sparse principal component analysis. Assuming the hardness of the planted clique problem [3], a series of papers [6, 20, 45] have shown that a comparable minimum sample size condition is necessary for any estimator computable in polynomial time complexity to achieve consistency and optimal convergence rates uniformly over a parameter space of interest. In particular, it was shown in [20] that this is the case even for the most restrictive parameter space in sparse principal component analysis—(discretized) Gaussian single spiked model with a sparse leading eigenvector. Establishing comparable computational lower bounds for sparse phase retrieval, especially under the Gaussian design, is an interesting project for future research.

In the case when $m \gtrsim p$ (ignoring polylog factor), to obtain an appropriate initialization, Algorithm 2 can be replaced by the standard spectral initialization for phase retrieval shown in [11, 37], in which the signal does not need to be sparse. In other words, under the scenario $m \gtrsim p$, the sample size condition $m \gtrsim k^2$ is likely not essential for the minimaxity of the thresholded Wirtinger flow method. It is interesting to study in the future, say, whether the condition $m \gtrsim k^2$ in Theorem 3.1 can be relaxed into $m \gtrsim \min(k^2, p)$ with the same level of estimation accuracy.

The convergence rate $\frac{\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k \log p}{m}}$ is essentially optimal. The following lower bound result has been essentially proven in [28].

THEOREM 3.2 ([28]). *Let $\Theta(k, p, R) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_2 = R, \|\mathbf{x}\|_0 = k\}$. Suppose the $\mathbf{a}_j$'s are i.i.d. $\mathcal{N}(0, \mathbf{I}_p)$, and the ε_j's are i.i.d. $\mathcal{N}(0, \sigma_0^2)$ with sub-exponential norm $\sigma = C_0'\sigma_0$ for some absolute constant C_0'. Moreover, assume $\mathbf{A}$ and $\boldsymbol{\varepsilon}$ are mutually independent. There holds under model (1.1),*

$$\inf_{\widehat{\mathbf{x}}} \sup_{\mathbf{x} \in \Theta(k,p,R)} \mathbb{P}_{(\mathbf{A},\mathbf{y}|\mathbf{x})}\left(\min_{i=0,1} \left\|\widehat{\mathbf{x}} - (-1)^i \mathbf{x}\right\|_2 \geq C_0 \frac{\sigma}{R}\sqrt{\frac{k \log(ep/k)}{m}}\right) \geq \frac{1}{5},$$

provided $m \geq C(\frac{\sigma^2}{\|\mathbf{x}\|_2^4} + 1)k \log(ep/k)$, where both C and C_0 are some absolute constants.

[4] With intractable methods, a slightly larger upper bound than (3.3) can be obtained the weaker sample size condition $m \gtrsim k$; see, for example, [28].

4. Numerical simulation. In this section, we report numerical simulation results to demonstrate how the relative estimation error depends on the thresholding parameter β, the noise-to-signal ratio (NSR) $\sigma/\|\mathbf{x}\|_2^2$, the sample size m, and the sparsity k. The main purpose of the section is two-folded. First, we demonstrate the theoretical bound in Theorem 3.1 is informative to the actual performance of the TWF algorithm. In addition, we provide numerical evidence that the performance of the proposed algorithm is not overly sensitive to the choices of the tuning parameters.

To guarantee fair comparison, we always fix the dimension of the signal $p = 1000$ and the initialization parameter $\alpha = 0.1$ (except for the first case on thresholding effect). Moreover, in each numerical experiment, we conservatively choose gradient descent step size parameter $\mu = 0.01$, and the number of iterations $T = 1000$ for thresholded Wirtinger flow. The resulting estimator is denoted as $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}^{(1000)}$.

As discussed before, the design matrix $\mathbf{A}$ consists of independent standard Gaussian random variables. With each fixed k, the support of $\mathbf{x}$ is uniformly distributed at random. The nonzero entries of $\mathbf{x}$ are i.i.d. $\sim \mathcal{N}(0, 1)$. The noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m)$, where σ is determined by $\|\mathbf{x}\|_2$ and the choice of NSR $\sigma/\|\mathbf{x}\|_2^2$. Notice that we simply denote by σ the common standard deviation of the noise rather than their sub-exponential norm, since there is an absolute constant that can absorb the scaling between the variance of a Gaussian variable and its sub-exponential norm.

1. *Thresholding effect*: Fix $\alpha = 0.1$, $m = 7000$, $k = 100$, and $\sigma/\|\mathbf{x}\|_2^2 = 1$. For each $\beta = 0, 0.25, 0.5, \ldots, 3$, we implement the algorithm for 10 times with independently generated $\mathbf{A}$, $\mathbf{x}$, and $\boldsymbol{\varepsilon}$. and then take the average of the 10 independent relative errors $\min(\|\widehat{\mathbf{x}}-\mathbf{x}\|_2, \|\widehat{\mathbf{x}}+\mathbf{x}\|_2)/\|\mathbf{x}\|_2$. The relation between the average relative error and β is plotted as the dashed curve in Figure 1. The result shows that the average relative error essentially decreases from 0.2365 to 0.1151 as the thresholding parameter increases from 0 to 0.75, and then increases slowly up to 0.1684 as β continues to increase to 3. Since in the case $\beta = 0$, the thresholded Wirtinger flow is essentially reduced to Wirtinger flow without thresholding. Therefore, our simulation illustrates the improvement of TWF over WF in terms of estimation accuracy in the noisy sparse phase retrieval setting.

We implement the above experiments again with the only difference $\alpha = 0.5$. The relation curve between the relative estimation error and β is plotted as the solid curve in Figure 1. It is clear that the performance of the algorithm is very close to the case $\alpha = 0.1$. This indicates that the estimation accuracy of TWF is relatively insensitive to the initialization parameter α.

2. *Sparsity effect*: Fix $m = 7000$, $\sigma/\|\mathbf{x}\|_2^2 = 1$, and $\beta = 1$. In each choice of sparsity $k = 25, 50, \ldots, 200$, the average of the relative error $\min(\|\widehat{\mathbf{x}} - \mathbf{x}\|_2, \|\widehat{\mathbf{x}} + \mathbf{x}\|_2)/\|\mathbf{x}\|_2$ is taken over 10 independent instances of $(\mathbf{A}, \mathbf{x}, \boldsymbol{\varepsilon})$. Figure 2 demonstrates that the average relative error essentially increases from 0.1059 to 0.1666 as

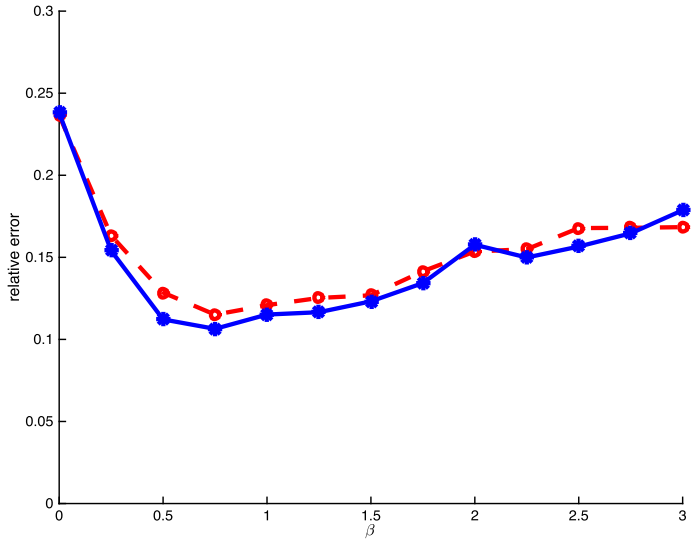

FIG. 1. *The relation between the average relative error and the thresholding parameter* β. *Setup of parameters*: $p = 1000$, $m = 7000$, $k = 100$, $\sigma/\|\mathbf{x}\|_2^2 = 1$, $\mu = 0.01$, *and* $T = 1000$. *Dashed curve with* $\alpha = 0.1$, *while solid curve with* $\alpha = 0.5$.

the sparsity increases from 25 to 200. This verifies our prediction by Theorem 3.1 that due to the denoising effect of iterative thresholding, the relative estimation accuracy of TWF improves as the signal becomes sparser.

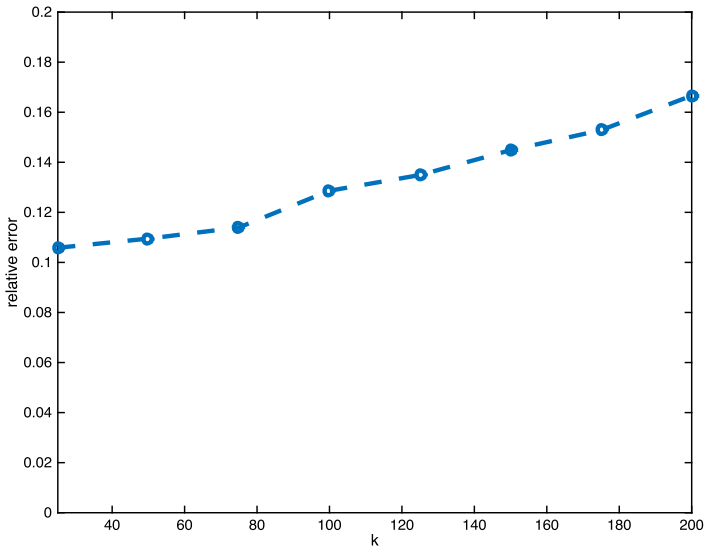

FIG. 2. *The relation between the average relative error and the sparsity* k. *Setup of parameters*: $p = 1000$, $\sigma/\|\mathbf{x}\|_2^2 = 1$, $m = 7000$, $\beta = 1$, $\alpha = 0.1$, $\mu = 0.01$ *and* $T = 1000$.

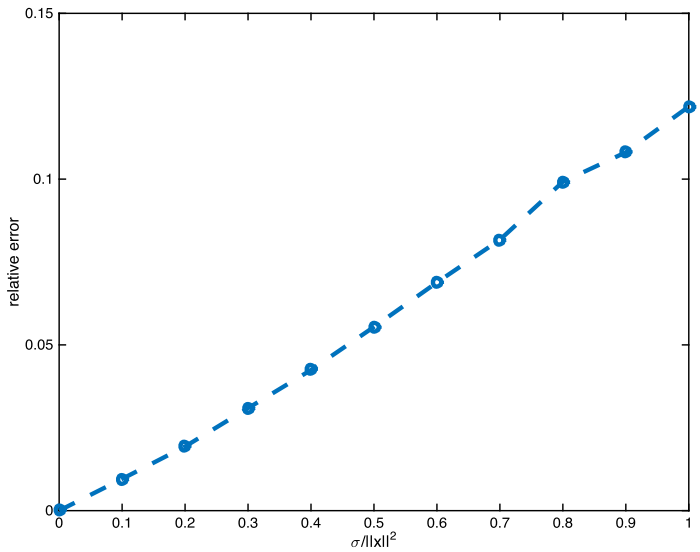

FIG. 3. *The relation between the average relative error and the noise-to-signal-ratio* $\sigma/\|\mathbf{x}\|_2^2$. *Setup of parameters*: $p = 1000$, $m = 7000$, $k = 100$, $\beta = 1$, $\alpha = 0.1$, $\mu = 0.01$ *and* $T = 1000$.

3. *Noise effect*: Fix $m = 7000$, $k = 100$, and $\beta = 1$. In each choice of NSR $\sigma/\|\mathbf{x}\|_2^2 = 0, 0.1, \ldots, 1$, with 5 instances of $(\mathbf{A}, \mathbf{x}, \boldsymbol{\varepsilon})$ generated independently, we take the average of the relative error $\min(\|\widehat{\mathbf{x}} - \mathbf{x}\|_2, \|\widehat{\mathbf{x}} + \mathbf{x}\|_2)/\|\mathbf{x}\|_2$. In Figure 3, it shows how the average relative error depends on NSR. The average relative error strictly increases from 0.0000 to 0.1219, as the NSR increases from 0 to 1. The figure perfectly verifies Theorem 3.1 in terms of the linear relationship between estimation relative error and NSR $\sigma/\|\mathbf{x}\|_2^2$.

4. *Sample size effect*: Fix $k = 100$, $\sigma/\|\mathbf{x}\|_2^2 = 1$, and $\beta = 1$. In each choice of $m = 2000, 3000, \ldots, 11{,}000$, with 5 instances of $(\mathbf{A}, \mathbf{x}, \boldsymbol{\varepsilon})$ generated independently, the average of the relative error $\min(\|\widehat{\mathbf{x}} - \mathbf{x}\|_2, \|\widehat{\mathbf{x}} + \mathbf{x}\|_2)/\|\mathbf{x}\|_2$ is calculated. In Figure 4, it shows how the average relative error depends on the sample size. When the sample sizes are 2000 and 3000, that is, twice and three times as large as p, the average relative errors are 0.8444 and 0.3651 respectively. In these cases, the thresholded gradient descent method leads to poor recovery of the original signal. When the sample size increases from 4000 to 11,000, the average relative error decreases steadily from 0.1692 to 0.0956. This also validates Theorem 3.1 in that the relative estimation error of TWF decreases as the sample size increases.

5. Discussion. In this paper, we established the optimal rates of convergence for noisy sparse phase retrieval under the Gaussian design in the presence of sub-exponential noise, provided that the sample size is sufficiently large. Furthermore, a thresholded gradient descent method called "Thresholded Wirtinger Flow" was introduced and shown to achieve the optimal rates.

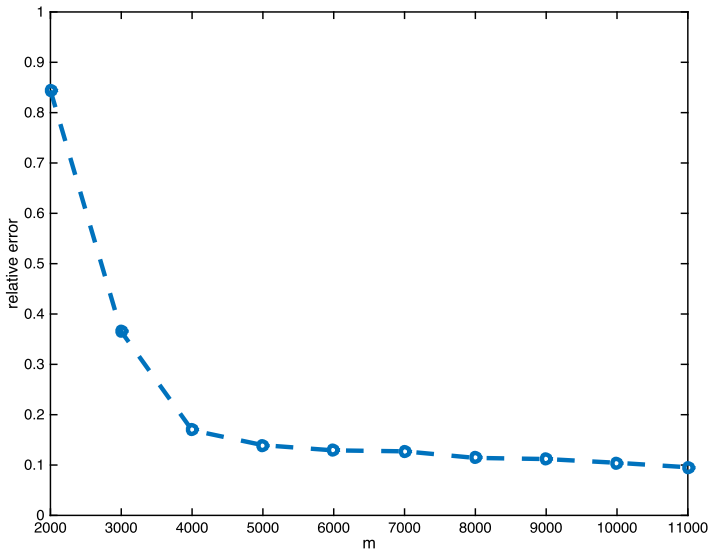

FIG. 4. *The relation between the average relative error and the sample size* m. *Setup of parameters*: $p = 1000$, $\sigma/\|\mathbf{x}\|_2^2 = 1$, $k = 100$, $\beta = 1$, $\alpha = 0.1$, $\mu = 0.01$ *and* $T = 1000$.

Iterative thresholding has been employed in a variety of problems in high-dimensional statistics, machine learning and signal processing, under the assumption that the signal or parameter vector/matrix satisfies a sparse or low-rank constraint. Examples include compressed sensing/sparse approximation [7, 17, 34, 36], sparse principal component analysis [33, 48], high-dimensional regression [1, 23, 47] and low-rank matrix recovery [8, 26, 29]. There are some connections between the present paper and [33]. The initialization method in Algorithm 2 is similar to the initialization method in [33], both of which are originated from [25]. However, the method in [33] was motivated by the fact that the sparsity constrained leading eigenvector of the sample covariance matrix is a minimax rate-optimal estimator and the thresholded power iteration serves as a heuristic algorithm for computing the sparsity constrained sample eigenvector. Thus, it is an unsupervised learning problem. In contrast, the thresholded gradient descent approach in the current paper is designed to seek a sparsity constrained solution to some least square problem, which is a supervised learning problem.

Regarding the application of iterative thresholding and projected gradient methods in high-dimensional M-estimation, their statistical optimality has been established when the empirical risk function satisfies certain properties, such as restrictive strong convexity and smoothness (RSC and RSS) [1, 23, 47]. Although our thresholded gradient method aims to solve (2.1) for a sparse solution, the existing analytical framework for high-dimensional M-estimation does not apply to the sparse phase retrieval problem, since the empirical risk function in (2.1) does not satisfy RSC in general, no matter how large the sample size is. Instead, we

have shown that thresholded gradient methods can achieve optimal statistical precision for signal recovery, even when the empirical risk function does not satisfy the common assumption of RSC.

Besides thresholded gradient methods, convexly and non-convexly regularized methods are also widely-used for high-dimensional M-estimation. In fact, some iterative thresholding methods are induced by regularizations; see, for example, [17]. Therefore, an alternative candidate method for solving the noisy sparse phase retrieval problem is to penalize the empirical risk function in (2.1) before taking the minimum, in order to promote a sparse solution. The major difficulty is apparently the non-convexity of the empirical risk function. An interesting result in [32] guarantees the statistical precision of all local optima, as long as the nonconvex penalty satisfies certain regularity conditions, and the empirical risk function, possibly non-convex, satisfies the restricted strong convexity. A similar result appeared in [46], in which the empirical risk function is required to satisfy a sparse eigenvalue (SE) condition. However, back to noisy sparse phase retrieval, the empirical risk function in (2.1) satisfies neither RSC nor SE in general, so there is no guarantee that all local optima are consistent. A natural question is whether some penalized version of (2.1) is strongly convex in a sufficiently large neighborhood of its global minimum, such that a tractable initializer lies in this neighborhood provided the sample size is sufficiently large. Another interesting question is whether the global minimizer of such penalized version of (2.1) is a rate-optimal estimator of the original sparse signal. We leave these questions for future research.

6. Proof of Theorem 3.1. We focus on the case where (3.1) holds. To prove the second part of theorem, we simply observe that Lemmas 6.2 and A.1 ensure that (3.1) holds with high probability under the extra conditions.

In model (1.1), denote $S = \operatorname{supp}(\mathbf{x})$, which implies $|S| = k$. Without loss of generality, we assume $S = \{1, \dots, k\}$. As to the Gaussian design matrix $\mathbf{A} \in \mathbb{R}^{m\times p}$, denote

$$\mathbf{A}_S := \begin{pmatrix} \mathbf{a}'_{1\,S} \\ \vdots \\ \mathbf{a}'_{m\,S} \end{pmatrix}, \qquad \mathbf{A}_{S^c} := \begin{pmatrix} \mathbf{a}'_{1\,S^c} \\ \vdots \\ \mathbf{a}'_{m\,S^c} \end{pmatrix}, \tag{6.1}$$

both of which are in $\mathbb{R}^{m\times p}$.

For any two random variables/vectors/matrices/sets X and Y, we denote by $X \perp\!\!\!\perp Y$ if X and Y are independent.

LEMMA 6.1. *From the model* (1.1), *we have* $\mathbf{y} \perp\!\!\!\perp \mathbf{A}_{S^c}$. *Moreover, we have* $\{I_1, \dots, I_k\} \perp\!\!\!\perp \mathbf{A}_{S^c}$ *and* $\phi \perp\!\!\!\perp \mathbf{A}_{S^c}$, *where* ϕ *and* $\{I_1, \dots, I_k\}$ *are defined in* (2.4) *and* (2.7), *respectively.*

PROOF. The fact $\mathbf{y} = |\mathbf{A}\mathbf{x}|^2 + \boldsymbol{\varepsilon} = |\mathbf{A}_S\mathbf{x}_S|^2 + \boldsymbol{\varepsilon}$ implies straightforwardly that $\mathbf{y} \perp\!\!\!\perp \mathbf{A}_{S^c}$. By (2.7), we know for all $l = 1, \dots, k$, I_l are defined by $\mathbf{y}$ and $\mathbf{A}_S$,

which implies that $I_l \perp\!\!\!\perp \mathbf{A}_{S^c}$ for all $l = 1, \ldots, k$. Finally, by (2.4), we know ϕ is determined uniquely by $\mathbf{y}$, which implies that $\phi \perp\!\!\!\perp \mathbf{A}_{S^c}$. $\square$

LEMMA 6.2. *On an event $\widetilde{E}_0$ with probability at least $1 - \frac{3}{m}$,*

$$1 - \left(2 + C_0 \frac{\sigma}{\|\mathbf{x}\|_2^2}\right)\sqrt{\frac{\log m}{m}} \le \frac{\phi^2}{\|\mathbf{x}\|_2^2} \le 1 + \left(2 + C_0 \frac{\sigma}{\|\mathbf{x}\|_2^2}\right)\sqrt{\frac{\log m}{m}} + \frac{2\log m}{m}$$

for some numerical constant $C_0 > 0$. As a consequence, for any $\delta < 1/10$, as long as $\frac{m}{\log m} \ge C(\delta)(1 + \frac{\sigma^2}{\|\mathbf{x}\|_2^4})$, there holds

$$\frac{9}{10} \le 1 - \delta \le \frac{\phi^2}{\|\mathbf{x}\|_2^2} \le 1 + \delta \le \frac{11}{10}.$$

PROOF. By the definition of ϕ^2 and y_j, $j = 1, \ldots, m$, we have

$$\phi^2 = \frac{1}{m}\sum_{j=1}^{m} (\mathbf{a}_j'\mathbf{x})^2 + \frac{1}{m}\sum_{j=1}^{m} \varepsilon_j.$$

As shown in Lemma A.7, with probability at least $1 - \frac{1}{m}$,

$$\left|\frac{1}{m}\sum_{j=1}^{m} \varepsilon_j\right| \le C_0\sigma\sqrt{\frac{\log m}{m}}$$

for some numerical constant $C_0 > 0$. Moreover, since $\mathbf{x}$ is fixed, there holds

$$\frac{\sum_{j=1}^{m} (\mathbf{a}_j'\mathbf{x})^2}{\|\mathbf{x}\|_2^2} \sim \chi^2(m).$$

By Lemma 4.1 of [27], with probability at least $1 - \frac{2}{m}$, we have

$$1 - 2\sqrt{\frac{\log m}{m}} \le \frac{\sum_{j=1}^{m} (\mathbf{a}_j'\mathbf{x})^2}{m\|\mathbf{x}\|_2^2} \le 1 + 2\sqrt{\frac{\log m}{m}} + \frac{2\log m}{m}.$$

The proof is complete. $\square$

LEMMA 6.3. *Let $\alpha \ge K(1 + \frac{\sigma}{\|\mathbf{x}\|_2^2})$ for some large enough absolute constant K, and $\widehat{\mathbf{x}}^{(0)}$ be defined in Algorithm 2. There exists a random vector $\mathbf{x}^{(0)}$ satisfying $\mathbf{x}^{(0)} \perp\!\!\!\perp \mathbf{A}_{S^c}$ and $\operatorname{supp}(\mathbf{x}^{(0)}) \subset S$, such that on an event E_{01} with probability at least $1 - \frac{16}{m} - 2e^{-k}$, we have*

$$\mathbf{x}^{(0)} = \widehat{\mathbf{x}}^{(0)}, \quad \text{and} \quad \min\left(\left\|\mathbf{x}^{(0)} - \mathbf{x}\right\|_2, \left\|\mathbf{x}^{(0)} + \mathbf{x}\right\|_2\right) \le \tfrac{1}{6}\|\mathbf{x}\|_2,$$

provided $m \ge C\alpha^2 k^2 \log(mp)$. Here, C is an absolute constant.

PROOF. Recall that $S = \{1, \ldots, k\}$ and $I_l = \frac{1}{m}\sum_{j=1}^m y_j a_{jl}^2$ for $l = 1, \ldots, p$. Define

$$S_0 = \left\{ l \in S : I_l > \left(1 + \alpha\sqrt{\frac{\log(mp)}{m}}\right)\phi^2 \right\} \subset S. \tag{6.2}$$

Since $\{I_1, \ldots, I_k, \phi\} \perp\!\!\!\perp \mathbf{A}_{S^c}$, we have $S_0 \perp\!\!\!\perp \mathbf{A}_{S^c}$. Define $\mathbf{x}^{(0)} \in \mathbb{R}^p$ as the leading eigenvector of

$$\mathbf{W}_{S_0 S_0} := \frac{1}{m}\sum_{j=1}^m y_j \mathbf{a}_{jS_0}\mathbf{a}_{jS_0}' \in \mathbb{R}^{p\times p}$$

with 2-norm ϕ. The fact $\operatorname{supp}(\mathbf{W}_{S_0\times S_0}) \subset S_0 \times S_0$ implies $\operatorname{supp}(\mathbf{x}^{(0)}) \subset S_0 \subset S$. Since $\{\mathbf{W}_{S_0S_0}, \phi\} \perp\!\!\!\perp \mathbf{A}_{S^c}$, we also have $\mathbf{x}^{(0)} \perp\!\!\!\perp \mathbf{A}_{S^c}$.

To simplify notation, let us write for any $j \in [m]$, $\widetilde{y}_j := (\mathbf{a}_j'\mathbf{x})^2 = \mathbf{a}_{jS}{}'\mathbf{x})^2$, which implies $y_j = \widetilde{y}_j + \varepsilon_j$. Notice that

$$I_l - \phi^2 = \frac{1}{m}\sum_{j=1}^m \widetilde{y}_j(a_{jl}^2 - 1) + \frac{1}{m}\sum_{j=1}^m \varepsilon_j(a_{jl}^2 - 1), \tag{6.3}$$

in which we will first control the second term. For a given $l \in [p]$, we know $a_{1l}^2 - 1, \ldots, a_{ml}^2 - 1$ are i.i.d. centered sub-exponential random variables with sub-exponential norms being an absolute constant. Then, by Bernstein inequality (see, e.g., Proposition 16 in [43]), conditionally on the ε_j's, we have with probability at least $1 - \frac{2}{mp}$,

$$\left|\sum_{j=1}^m \varepsilon_j(a_{jl}^2 - 1)\right| \le C_0\big(\|\boldsymbol{\varepsilon}\|_2\sqrt{\log(mp)} + \|\boldsymbol{\varepsilon}\|_\infty \log(mp)\big)$$

for some absolute constant C_0. Then by Lemma A.7, with probability at least $1 - 4/m$, we have

$$\begin{aligned}\max_{1\le l\le p}\left|\frac{1}{m}\sum_{j=1}^m \varepsilon_j(a_{jl}^2 - 1)\right| &\le C_0\sigma\left(\sqrt{\frac{\log(mp)}{m}} + \frac{\log^2(mp)}{m}\right)\\ &\le C_0\sigma\sqrt{\frac{\log(mp)}{m}},\end{aligned} \tag{6.4}$$

provided $m \ge C(\log p)$ for some absolute constant C.

Next, we prove that with high probability $\mathbf{x}^{(0)} = \widehat{\mathbf{x}}^{(0)}$. It suffices to prove $\widehat{S}_0 = S_0$, that is, $\widehat{S}_0 \subset S$. For any $l \in S^c$, a_{jl} and $\widetilde{y}_j$ are independent, and so conditional on $\{\widetilde{y}_j, j \in [m]\}$, $\sum_{j=1}^m \widetilde{y}_j a_{jl}^2$ is a weighted sum of χ_1^2 variables. By Lemma 4.1 of [27],

$$\mathbb{P}\left\{\sum_{j=1}^m \widetilde{y}_j(a_{jl}^2 - 1) > 2\sqrt{t}\left(\sum_{j=1}^m \widetilde{y}_j^2\right)^{1/2} + 2\Big(\max_j \widetilde{y}_j\Big)t\right\} \le \exp(-t).$$

Moreover, Chebyshev's inequality, the Gaussian tail bound and the union bound lead to

$$\mathbb{P}\left\{\sum_{j=1}^{m}\widetilde{y}_j^2/\|\mathbf{x}\|_2^4 > 3m+\sqrt{96mt}\right\} \le t^{-2},$$

$$\mathbb{P}\left\{\max_j \widetilde{y}_j/\|\mathbf{x}\|_2^2 > t\right\} \le 2m\exp(-t/2).$$

Thus, with probability at least $1-\frac{4}{m}$, for all $l\in S^c$,

$$\frac{1}{m}\sum_{j=1}^{m}\widetilde{y}_j\big(a_{jl}^2-1\big) \le 2\sqrt{3+\sqrt{96}}\|\mathbf{x}\|_2^2\sqrt{\frac{\log(mp)}{m}}+8\|\mathbf{x}\|_2^2\frac{(\log(mp))^2}{m} \tag{6.5}$$

$$\le 8\|\mathbf{x}\|_2^2\sqrt{\frac{\log(mp)}{m}}. \tag{6.6}$$

Here, the last inequality holds when $m\ge C$ for some absolute constant C.

Since $\alpha \ge K(1+\frac{\sigma}{\|\mathbf{x}\|_2^2})$ with large enough K, by (6.3), (6.5), (6.4) and Lemma 6.2, we obtain that with probability at least $1-\frac{11}{m}$, for all $l\in S^c$,

$$I_l-\phi^2 \le \big(8\|\mathbf{x}\|_2^2+C_0\sigma\big)\sqrt{\frac{\log(mp)}{m}} \le \alpha\phi^2\sqrt{\frac{\log(mp)}{m}},$$

which implies that $\widehat{S}_0\subset S$.

Next, we prove that $\|\mathbf{x}^{(0)}-\mathbf{x}\|_2/\|\mathbf{x}\|_2\le\frac{1}{6}$ with high probability. For any fixed $l\in S$, straightforward calculation yields $\mathbb{E}\widetilde{y}_j a_{jl}^2=\|\mathbf{x}\|_2^2+2x_l^2$. On the other hand,

$$\mathbb{E}\widetilde{y}_j^2 a_{jl}^4 = 105x_l^4+90x_l^2\big(\|\mathbf{x}\|_2^2-x_l^2\big)+9\big(\|\mathbf{x}\|_2^2-x_l^2\big)^2.$$

So for $X_j=\|\mathbf{x}\|_2^2+2x_l^2-\widetilde{y}_j a_{jl}^2$, we have $X_j\le\|\mathbf{x}\|_2^2+2x_l^2\le 3\|\mathbf{x}\|_2^2$, $\mathbb{E}X_i=0$ and $\mathbb{E}X_i^2=20x_j^4+68\|\mathbf{x}\|_2^2x_l^2+8\|\mathbf{x}\|_2^4\le 96\|\mathbf{x}\|_2^4$. By Lemma A.1,

$$\mathbb{P}\left\{\sum_{j=1}^{m}\widetilde{y}_j a_{jl}^2-m\big(\|\mathbf{x}\|_2^2+2x_l^2\big)\le -t\right\} \le \exp\left(-\frac{t^2}{192\|\mathbf{x}\|_2^4 m}\right).$$

Next, Lemma 4.1 of [27] leads to with probability at least $1-\frac{1}{m}$,

$$\frac{1}{m}\sum_{j=1}^{m}\widetilde{y}_j-\|\mathbf{x}\|_2^2 \le \left(2\sqrt{\frac{\log m}{m}}+\frac{2\log m}{m}\right)\|\mathbf{x}\|_2^2 \le 2.1\|\mathbf{x}\|_2^2\sqrt{\frac{\log m}{m}}.$$

The last two inequalities, together with (6.4) and (6.3), imply that with probability at least $1-\frac{6}{m}$, for all $l\in S$,

$$I_l-\phi^2 \ge 2x_l^2-\big(16\|\mathbf{x}\|_2^2+C_0\sigma\big)\sqrt{\frac{\log(mp)}{m}}.$$

Define $S_- = \{l \in S : x_l^2 \geq (11 + \frac{3}{5}\alpha)\|\mathbf{x}\|_2^2\sqrt{\frac{\log(mp)}{m}}\}$. Then, for all $l \in S_-$ we have

$$I_l - \phi^2 \geq \left(\frac{6}{5}\alpha\|\mathbf{x}\|_2^2 + 6\|\mathbf{x}\|_2^2 - C_0\sigma\right)\sqrt{\frac{\log(mp)}{m}}.$$

Since $\alpha \geq K(1 + \frac{\sigma}{\|\mathbf{x}\|_2^2})$ with sufficiently large absolute constant K, by Lemma 6.2, we have $I_l - \phi^2 \geq \alpha\phi^2\sqrt{\frac{\log(mp)}{m}}$ for all $l \in S_-$ on an event with probability at least $1 - 9/m$. This implies $S_- \subset S_0$.

Therefore, we have

$$\|\mathbf{x} - \mathbf{x}_{S_0}\|_2^2 \leq \|\mathbf{x} - \mathbf{x}_{S_-}\|_2^2 \leq \left(11 + \frac{3\alpha}{5}\right)\|\mathbf{x}\|_2^2\sqrt{\frac{k^2\log(mp)}{m}} \leq \delta^2\|\mathbf{x}\|_2^2, \tag{6.7}$$

provided that $m \geq C(\delta)\alpha^2k^2\log(mp)$. Notice that $\mathbb{E}\mathbf{W} = \|\mathbf{x}\|_2^2\mathbf{I}_p + 2\mathbf{x}\mathbf{x}'$, which implies that $(\mathbb{E}\mathbf{W})_{SS} = \|\mathbf{x}\|_2^2(\mathbf{I}_p)_{SS} + 2\mathbf{x}\mathbf{x}'$. Furthermore, by the definition of $\mathbf{W}$, we have

$$\mathbf{W}_{SS} = \frac{1}{m}\sum_{j=1}^m |\mathbf{a}_{jS}'\mathbf{x}|^2\mathbf{a}_{jS}\mathbf{a}_{jS}' + \frac{1}{m}\sum_{j=1}^m \varepsilon_j\mathbf{a}_{jS}\mathbf{a}_{jS}'.$$

By Lemma A.6, with probability at least $1 - 1/m$, we have

$$\left\|\frac{1}{m}\sum_{j=1}^m |\mathbf{a}_{jS}'\mathbf{x}|^2\mathbf{a}_{jS}\mathbf{a}_{jS}' - \left(\|\mathbf{x}\|_2^2(\mathbf{I}_p)_{SS} + 2\mathbf{x}\mathbf{x}'\right)\right\| \leq \frac{\delta}{2}\|\mathbf{x}\|_2^2,$$

provided $m \geq C(\delta)k\log p$. Moreover, by Lemma A.7 and Lemma A.8, with probability at least $1 - 2/m - 2e^{-k}$, we have $\|\sum_{j=1}^m \varepsilon_j\mathbf{a}_{jS}\mathbf{a}_{jS}'\| \leq C_0\sigma\sqrt{m(k + \log m)}$. Since $m \geq C(\delta)\frac{\sigma^2}{\|\mathbf{x}\|_2^4}k\log(mp)$, we have $\frac{1}{m}\|\sum_{j=1}^m \varepsilon_j\mathbf{a}_{jS}\mathbf{a}_{jS}'\| \leq \frac{\delta}{2}\|\mathbf{x}\|_2^2$. This implies that

$$\|\mathbf{W}_{S_0S_0} - (\mathbb{E}\mathbf{W})_{S_0S_0}\| \leq \|\mathbf{W}_{SS} - (\mathbb{E}\mathbf{W})_{SS}\| \leq \delta\|\mathbf{x}\|_2^2.$$

It is noteworthy that the leading eigenvector of $(\mathbb{E}\mathbf{W})_{S_0S_0}$ with unit norm is $\mathbf{x}_{S_0}/\|\mathbf{x}_{S_0}\|_2$, and the eigengap between the leading two eigenvalues of $(\mathbb{E}\mathbf{W})_{S_0S_0}$ is $2\|\mathbf{x}_{S_0}\|_2^2$. Recall that $\mathbf{x}^{(0)}$ is the leading eigenvector $\mathbf{W}_{S_0S_0}$ with norm ϕ. Then by the sin-theta theorem,

$$\left\|\frac{\mathbf{x}^{(0)}(\mathbf{x}^{(0)})^T}{\phi^2} - \frac{\mathbf{x}_{S_0}\mathbf{x}_{S_0}^T}{\|\mathbf{x}_{S_0}\|_2^2}\right\| \leq \frac{\delta\|\mathbf{x}\|_2^2}{2\|\mathbf{x}_{S_0}\|_2^2 - \delta\|\mathbf{x}\|_2^2} \leq \frac{\delta}{2 - 5\delta}.$$

By Lemma 6.2, we have $1 + \delta \geq \phi/\|\mathbf{x}\|_2 \geq 1 - \delta$. Together with $1 \geq \|\mathbf{x}_{S_0}\|_2/\|\mathbf{x}\|_2 \geq 1 - \delta$, we can easily obtain that $\min(\|\mathbf{x}^{(0)} - \mathbf{x}\|_2, \|\mathbf{x}^{(0)} + \mathbf{x}\|_2) \leq C_0\delta\|\mathbf{x}\|_2$ for some absolute constant C_0. By letting δ be small enough, we have $\min(\|\mathbf{x}^{(0)} - \mathbf{x}\|_2, \|\mathbf{x}^{(0)} + \mathbf{x}\|_2) \leq 1/6\|\mathbf{x}\|_2$. □

LEMMA 6.4. *Define* $\eta(\mathbf{z}) = \mathcal{T}_{(\mu/\phi^2)\tau(\mathbf{z})}(\mathbf{z} - \frac{\mu}{\phi^2}\nabla f(\mathbf{z})_S)$. *With probability at least* $1 - \frac{15}{m} - 4e^{-k}$, *for all* $\mathbf{z} \in \mathbb{R}^p$ *satisfying* $\|\mathbf{z} - \mathbf{x}\|_2 \le \frac{1}{6}\|\mathbf{x}\|_2$ *and* $\operatorname{supp}(\mathbf{z}) \subset S$, *we have*

$$\frac{\|\eta(\mathbf{z}) - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \le \left(1 - \frac{\mu}{8}\right)\frac{\|\mathbf{z} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} + C_0\frac{\mu\sigma}{\|\mathbf{x}\|_2^2}\sqrt{\frac{k\log p}{m}},$$

provided $\mu \le \mu_0$ *and* $m \ge Ck^2 \log p$. *Here,* C_0, C, *and* μ_0 *are numerical constants. This implies that, on an event* E_{02} *with probability at least* $1 - \frac{30}{m} - 8e^{-k}$, *for all* $\mathbf{z} \in \mathbb{R}^p$ *satisfying* $\min(\|\mathbf{z} - \mathbf{x}\|_2, \|\mathbf{z} + \mathbf{x}\|_2) \le \frac{1}{6}\|\mathbf{x}\|_2$ *and* $\operatorname{supp}(\mathbf{z}) \subset S$, *we have*

$$\min_{i=0,1}\|\eta(\mathbf{z}) - (-1)^i\mathbf{x}\|_2 \le \left(1 - \frac{\mu}{8}\right)\min_{i=0,1}\|\mathbf{z} - (-1)^i\mathbf{x}\|_2 + C_0\frac{\mu\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}.$$

PROOF. Recall that $\|\mathbf{z}-\mathbf{x}\|_2 \le \|\mathbf{x}\|_2/6$ and $\mathbf{z}$ is supported on S. Define $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^p$ by

$$\mathbf{u} = \eta(\mathbf{z}) = \mathcal{T}_{(\mu/\phi^2)\tau(\mathbf{z})}\left(\mathbf{z} - \frac{\mu}{\phi^2}\nabla f(\mathbf{z})_S\right) = \mathbf{z} - \frac{\mu}{\phi^2}\nabla f(\mathbf{z})_S + \frac{\mu}{\phi^2}\tau(\mathbf{z})\mathbf{v},$$

such that $\operatorname{supp}(\mathbf{v}) \subset S$ and $\|\mathbf{v}\|_\infty \le 1$.

Since $\operatorname{supp}(\mathbf{z}) \subset S = \{1, \dots, k\}$, we have

$$\nabla f(\mathbf{z})_S = \frac{1}{m}\sum_{j=1}^m (|\mathbf{a}'_{j\,S}\mathbf{z}|^2 - y_j)(\mathbf{a}'_{j\,S}\mathbf{z})\mathbf{a}_{j\,S}. \tag{6.8}$$

For convenience, let

$$\widetilde{\nabla f(\mathbf{z})}_S = \frac{1}{m}\sum_{j=1}^m (|\mathbf{a}'_{j\,S}\mathbf{z}|^2 - |\mathbf{a}'_{j\,S}\mathbf{x}|^2)(\mathbf{a}'_{j\,S}\mathbf{z})\mathbf{a}_{j\,S}, \tag{6.9}$$

and so

$$\nabla f(\mathbf{z})_S - \widetilde{\nabla f(\mathbf{z})}_S = -\frac{1}{m}\sum_{j=1}^m \varepsilon_j(\mathbf{a}'_{j\,S}\mathbf{z})\mathbf{a}_{j\,S}. \tag{6.10}$$

Denote $\mathbf{h} = \mathbf{z} - \mathbf{x} \in \mathbb{R}^p$, which implies $\operatorname{supp}(\mathbf{h}) \subset S$ and $\|\mathbf{h}\|_2 \le \|\mathbf{x}\|_2/6$. Straightforward calculation yields

$$\begin{aligned}\|\mathbf{u} - \mathbf{x}\|_2 &\le \left\|\mathbf{h} - \frac{\mu}{\phi^2}\widetilde{\nabla f(\mathbf{z})}_S\right\|_2 + \frac{\mu}{\phi^2}\|\nabla f(\mathbf{z})_S - \widetilde{\nabla f(\mathbf{z})}_S\|_2 + \frac{\mu\sqrt{k}}{\phi^2}\tau(\mathbf{z}) \\ &:= T_1 + \frac{\mu}{\phi^2}T_2 + \frac{\mu\sqrt{k}}{\phi^2}\tau(\mathbf{z}).\end{aligned} \tag{6.11}$$

It suffices to bound T_1, T_2 and $\tau(\mathbf{z})$.

Bound for T_1. By simple algebra, we have

$$T_1^2 = \|\mathbf{h}\|_2^2 - \frac{\mu}{m\phi^2}\sum_{j=1}^{m}\big(2|\mathbf{a}'_{j\,S}\mathbf{x}|^2|\mathbf{a}'_{j\,S}\mathbf{h}|^2 + 3(\mathbf{a}'_{j\,S}\mathbf{x})(\mathbf{a}'_{j\,S}\mathbf{h})^3 + |\mathbf{a}'_{j\,S}\mathbf{h}|^4\big)$$

$$+ \frac{\mu^2}{\phi^4}\big\|\widetilde{\nabla f(\mathbf{z})}_S\big\|_2^2 \tag{6.12}$$

$$:= \|\mathbf{h}\|_2^2 - \frac{\mu}{\phi^2}T_{11} + \frac{\mu^2}{\phi^4}T_{12}.$$

In what follows, we derive lower bound for T_{11} and upper bound for T_{12} separately.

Notice that

$$T_{11} = \frac{1}{m}\sum_{j=1}^{m}\big(2(\mathbf{a}'_{j\,S}\mathbf{x})^2(\mathbf{a}'_{j\,S}\mathbf{h})^2 + 3(\mathbf{a}'_{j\,S}\mathbf{x})(\mathbf{a}'_{j\,S}\mathbf{h})^3 + (\mathbf{a}'_{j\,S}\mathbf{h})^4\big).$$

First, by Lemma A.6 with probability at least $1 - 1/m$, we have

$$\frac{1}{m}\sum_{j=1}^{m}2(\mathbf{a}'_{j\,S}\mathbf{x})^2(\mathbf{a}'_{j\,S}\mathbf{h})^2 \geq (2-2\delta)\big(2(\mathbf{x}'\mathbf{h})^2 + \|\mathbf{x}\|_2^2\|\mathbf{h}\|_2^2\big).$$

By Lemma A.5, with probability at least $1 - 2/m$, we have

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m}3(\mathbf{a}'_{j\,S}\mathbf{x})(\mathbf{a}'_{j\,S}\mathbf{h})^3 &\leq \frac{3}{m}\left(\sum_{j=1}^{m}(\mathbf{a}'_{j\,S}\mathbf{x})^4\right)^{1/4}\left(\sum_{j=1}^{m}(\mathbf{a}'_{j\,S}\mathbf{h})^4\right)^{3/4}\\ &\leq \frac{3}{m}\big((3m)^{1/4} + k^{1/2} + \sqrt{2\log m}\big)^4\|\mathbf{x}\|_2\|\mathbf{h}\|_2^3\\ &\leq 10\|\mathbf{x}\|_2\|\mathbf{h}\|_2^3,\end{aligned}$$

provided $m \geq Ck^2$ for some sufficiently large numerical constant C. This implies

$$T_{11} \geq (2-2\delta)\|\mathbf{x}\|_2^2\|\mathbf{h}\|_2^2 - 10\|\mathbf{x}\|_2\|\mathbf{h}\|_2^3 \geq (1/3-2\delta)\|\mathbf{x}\|_2^2\|\mathbf{h}\|_2^2.$$

As to the upper bound for T_{12}, we can find $\|\mathbf{w}\|_2 = 1$, such that

$$T_{12} = \big\|\widetilde{\nabla f(\mathbf{z})}_S\big\|_2^2 \leq \frac{2}{m^2}\left|\sum_{j=1}^{m}|\mathbf{a}'_{j\,S}\mathbf{h}||\mathbf{a}'_{j\,S}(2\mathbf{x}+\mathbf{h})||\mathbf{a}'_{j\,S}(\mathbf{x}+\mathbf{h})||\mathbf{a}'_{j\,S}\mathbf{w}|\right|^2.$$

By Hölder's inequality and Lemma A.5, we have

$$\begin{aligned}T_{12} &\leq \frac{2}{m^2}\left(\sum_{j=1}^{m}|\mathbf{a}'_{j\,S}\mathbf{h}|^4\sum_{j=1}^{m}|\mathbf{a}'_{j\,S}(2\mathbf{x}+\mathbf{h})|^4\sum_{j=1}^{m}|\mathbf{a}'_{j\,S}(\mathbf{x}+\mathbf{h})|^4\sum_{j=1}^{m}|\mathbf{a}'_{j\,S}\mathbf{w}|^4\right)^{1/2}\\ &\leq \frac{2}{m^2}\big((3m)^{1/4} + k^{1/2} + \sqrt{2\log m}\big)^8\|\mathbf{h}\|_2^2\|2\mathbf{x}+\mathbf{h}\|_2^2\|\mathbf{x}+\mathbf{h}\|_2^2\|\mathbf{w}\|_2^2\\ &\leq C_0\|\mathbf{h}\|_2^2\|\mathbf{x}\|_2^4,\end{aligned}$$

provided $m \geq Ck^2$, with sufficiently large constants C_0 and C. To summarize, with probability at least $1 - 3/m$,

$$T_1^2 \leq \|\mathbf{h}\|_2^2 - \frac{\mu}{\phi^2}(1/3 - 2\delta)\|\mathbf{h}\|_2^2\|\mathbf{x}\|_2^2 + C_0\frac{\mu^2}{\phi^4}\|\mathbf{x}\|_2^4\|\mathbf{h}\|_2^2. \tag{6.13}$$

By Lemma 6.2, letting δ small enough, we have with probability at least $1 - 6/m$,

$$T_1 \leq (1 - \mu/8)\|\mathbf{h}\|_2,$$

provided $\mu \leq \mu_0$ with sufficiently small absolute constant $\mu_0 > 0$.

Bound for T_2. Note that

$$T_2 \leq \frac{7}{6m}\|\mathbf{x}\|_2\left\|\sum_{j=1}^m \varepsilon_j \mathbf{a}_{jS}\mathbf{a}'_{jS}\right\|.$$

By Lemma A.7 and Lemma A.8, with probability at least $1 - 2/m - 2e^{-k}$, we have

$$\left\|\sum_{j=1}^m \varepsilon_j \mathbf{a}_{jS}\mathbf{a}'_{jS}\right\| \leq C_0\sigma\sqrt{m(k + \log m)}$$

provided $m/\log m \geq k$. In summary, by Lemma 6.2, we have that with probability at least $1 - 5/m - 2e^{-k}$,

$$\frac{\mu}{\phi^2}T_2 \leq C_0\mu\frac{\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k + \log m}{m}}.$$

Bound for $\tau(\mathbf{z})$. By simple algebra,

$$\begin{aligned}
\tau^2(\mathbf{z}) &= \frac{\beta \log p}{m^2}\sum_{j=1}^m \big((\mathbf{a}'_{jS}\mathbf{h})\mathbf{a}'_{jS}(2\mathbf{x} + \mathbf{h}) - \varepsilon_j\big)^2|\mathbf{a}'_{jS}(\mathbf{x} + \mathbf{h})|^2 \\
&\leq \frac{2\beta \log p}{m^2}\left\{\sum_{j=1}^m |\mathbf{a}'_{jS}\mathbf{h}|^2|\mathbf{a}'_{jS}(2\mathbf{x} + \mathbf{h})|^2|\mathbf{a}'_{jS}(\mathbf{x} + \mathbf{h})|^2\right. \\
&\quad \left. + \sum_{j=1}^m \varepsilon_j^2|\mathbf{a}'_{jS}(\mathbf{x} + \mathbf{h})|^2\right\} \\
&:= \frac{2\beta \log p}{m^2}(\mathcal{T}_1 + \mathcal{T}_2).
\end{aligned}$$

By Hölder's inequality and Lemma A.5, with probability at least $1 - 2/m$, we have

$$\begin{aligned}
\mathcal{T}_1 &\leq \left(\sum_{j=1}^m |\mathbf{a}'_{jS}\mathbf{h}|^6\right)^{1/3}\left(\sum_{j=1}^m |\mathbf{a}'_{jS}(2\mathbf{x} + \mathbf{h})|^6\right)^{1/3}\left(\sum_{j=1}^m |\mathbf{a}'_{jS}(\mathbf{x} + \mathbf{h})|^6\right)^{1/3} \\
&\leq C_0\|\mathbf{A}_S\|_{2\to 6}^6\|\mathbf{h}\|_2^2\|\mathbf{x}\|_2^4 \leq C_0(m + k^3)\|\mathbf{h}\|_2^2\|\mathbf{x}\|_2^4
\end{aligned}$$

for some numerical constant C_0. By Lemma A.7 and Lemma A.8, with probability at least $1 - 2/m - 2e^{-k}$, we have

$$\mathcal{T}_2 \le \frac{49}{36}\|\mathbf{x}\|_2^2 \left\| \sum_{j=1}^{m} \varepsilon_j^2 \mathbf{a}_{j_S} \mathbf{a}_{j_S}' \right\| \le C_0 m\sigma^2 \|\mathbf{x}\|_2^2,$$

for some numerical constant C_0, provided $\frac{m}{\log^2 m} \ge k$. In summary,

$$\begin{aligned} \frac{\mu}{\phi^2}\sqrt{k}\tau &\le C_0\mu\left(\frac{\sqrt{(mk+k^4)\log p}}{m}\|\mathbf{h}\|_2 + \frac{\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}\right) \\ &\le \frac{\mu\|\mathbf{h}\|_2}{16} + C_0\frac{\mu\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}, \end{aligned} \tag{6.14}$$

provided $m \ge C\max(k\log p, k^2\sqrt{\log p})$.

Summary. We can guarantee that, with probability at least $1 - \frac{15}{m} - 4e^{-k}$,

$$\frac{\|\mathbf{u}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \le \left(1 - \frac{\mu}{16}\right)\frac{\|\mathbf{z}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} + C_0\mu\sqrt{\frac{k\log p}{m}}\frac{\sigma}{\|\mathbf{x}\|_2^2}, \tag{6.15}$$

for some absolute constant $C_0 > 0$, provided $m \ge Ck^2\log(mp)$ and $\mu \le \mu_0$. □

Suppose E_0 is the intersection of the events E_{01} and E_{02} described by Lemmas 6.3 and 6.4, respectively. Then we have

$$\mathbb{P}(E_0) \ge 1 - \frac{46}{m} - 10e^{-k}.$$

The effectiveness of thresholded Wirtinger flow is guaranteed by the following induction argument.

LEMMA 6.5. *Let $\beta = 4$ and $\widehat{\mathbf{x}}^{(n)}, n = 0, 1, 2, \ldots$ are defined iteratively by (2.10) and (2.6). For fixed $n \ge 0$, assume that there exists a random vector $\mathbf{x}^{(n)}$ satisfying $\mathbf{x}^{(n)} \perp\!\!\!\perp \mathbf{A}_{S^c}$ and $\operatorname{supp}(\mathbf{x}^{(n)}) \subset S$, and that on an event $E_n \subset E_0$ we have $\widehat{\mathbf{x}}^{(n)} = \mathbf{x}^{(n)}$ and $\min_{i=0,1}\|\widehat{\mathbf{x}}^{(n)} - (-1)^i\mathbf{x}\|_2 \le \frac{1}{6}\|\mathbf{x}\|_2$. Then there exists a random vector $\mathbf{x}^{(n+1)}$ satisfying $\mathbf{x}^{(n+1)} \perp\!\!\!\perp \mathbf{A}_{S^c}$ and $\operatorname{supp}(\mathbf{x}^{(n+1)}) \subset S$, and on an event $E_{n+1} \subset E_n$ satisfying $\mathbb{P}(E_n/E_{n+1}) \le 1 - \frac{1}{m^2 p}$, we have $\widehat{\mathbf{x}}^{(n+1)} = \mathbf{x}^{(n+1)}$ and*

$$\begin{aligned} &\min_{i=0,1}\|\widehat{\mathbf{x}}^{(n+1)} - (-1)^i\mathbf{x}\|_2 \\ &\quad\le \left(1 - \frac{\mu}{16}\right)\min_{i=0,1}\|\widehat{\mathbf{x}}^{(n)} - (-1)^i\mathbf{x}\|_2 + C_0\frac{\mu\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}} \le \frac{1}{6}\|\mathbf{x}\|_2. \end{aligned}$$

PROOF. The improved estimation is defined as

$$\widehat{\mathbf{x}}^{(n+1)} = \mathcal{T}_{(\mu/\phi^2)\tau(\widehat{\mathbf{x}}^{(n)})}\Big(\widehat{\mathbf{x}}^{(n)} - \frac{\mu}{\phi^2}\nabla f\big(\widehat{\mathbf{x}}^{(n)}\big)\Big),$$

where $\mathcal{T}_\tau$ is the soft-thresholding operator. We now define

$$\mathbf{x}^{(n+1)} := \eta\big(\mathbf{x}^{(n)}\big) = \mathcal{T}_{(\mu/\phi^2)\tau(\mathbf{x}^{(n)})}\Big(\mathbf{x}^{(n)} - \frac{\mu}{\phi^2}\nabla f\big(\mathbf{x}^{(n)}\big)_S\Big).$$

By the definition of ∇f, τ and ϕ, as well as the assumption that $\mathbf{x}^{(n)} \perp\!\!\!\perp \mathbf{A}_{S^c}$ and $\operatorname{supp}(\mathbf{x}^{(n)}) \subset S$, we can prove $\operatorname{supp}(\mathbf{x}^{(n+1)}) \subset S$ as well as $\mathbf{x}^{(n+1)} \perp\!\!\!\perp \mathbf{A}_{S^c}$. In fact, by the definition (2.5), we know if $\mathbf{x}^{(n)}$ is supported on S and independent of $\mathbf{A}_{S^c}$, then $\tau(\mathbf{x}^{(n)})$ is independent of $\mathbf{A}_{S^c}$. Moreover, by the definition of the gradient (2.2), we know $(\nabla f(\mathbf{x}^{(n)}))_S$ is supported on S and independent of $\mathbf{A}_{S^c}$. The assertion is established by the obvious fact $\phi \perp\!\!\!\perp \mathbf{A}_{S^c}$ shown in Lemma 6.1.

In the following, we will construct $E_{n+1} \subset E_n$ such that $\widehat{\mathbf{x}}^{(n+1)} = \mathbf{x}^{(n+1)}$ on E_{n+1}. For any $i = k+1, k+2, \ldots, p$, with probability $1 - \frac{1}{m^2p^2}$,

$$\begin{aligned}\left|\frac{\partial}{\partial z_i} f\big(\mathbf{x}^{(n)}\big)\right| &= \left|\frac{1}{m}\sum_{j=1}^m \big(\big|\mathbf{a}_j'\mathbf{x}^{(n)}\big|^2 - y_j\big)\big(\mathbf{a}_j'\mathbf{x}^{(n)}\big)(\mathbf{a}_j)_i\right| \\ &\le \frac{\sqrt{4\log(mp)}}{m}\sqrt{\sum_{j=1}^m \big(\big|\mathbf{a}_j'\mathbf{x}^{(n)}\big|^2 - y_j\big)^2\big|\mathbf{a}_j'\mathbf{x}^{(n)}\big|^2} \\ &\le \tau\big(\mathbf{x}^{(n)}\big).\end{aligned}$$

The first inequality is due to $\operatorname{supp}(\mathbf{x}^{(n)}) \subset S$ and $\mathbf{x}^{(n)} \perp\!\!\!\perp \mathbf{A}_{S^c}$, and the second inequality is due to $\beta = 4$. Then with probability at least $1 - \frac{1}{m^2p}$,

$$\max_{k+1\le i\le p}\left|\frac{\partial}{\partial z_i} f\big(\mathbf{x}^{(n)}\big)\right| \le \tau\big(\mathbf{x}^{(n)}\big),$$

which implies

$$\mathcal{T}_{(\mu/\phi^2)\tau(\mathbf{x}^{(n)})}\Big(\mathbf{x}^{(n)} - \frac{\mu}{\phi^2}\nabla f\big(\mathbf{x}^{(n)}\big)\Big) = \mathcal{T}_{(\mu/\phi^2)\tau(\mathbf{x}^{(n)})}\Big(\mathbf{x}^{(n)} - \frac{\mu}{\phi^2}\nabla f\big(\mathbf{x}^{(n)}\big)_S\Big).$$

Notice that on the event E_n, we have $\widehat{\mathbf{x}}^{(n)} = \mathbf{x}^{(n)}$, and hence

$$\widehat{\mathbf{x}}^{(n+1)} = \mathcal{T}_{(\mu/\phi^2)\tau(\mathbf{x}^{(n)})}\Big(\mathbf{x}^{(n)} - \frac{\mu}{\phi^2}\nabla f\big(\mathbf{x}^{(n)}\big)\Big).$$

Then there exists $E_{n+1} \subset E_n$, such that $\mathbb{P}(E_n/E_{n+1}) \le \frac{1}{m^2p}$, and

$$\widehat{\mathbf{x}}^{(n+1)} = \mathcal{T}_{(\mu/\phi^2)\tau(\mathbf{x}^{(n)})}\Big(\mathbf{x}^{(n)} - \frac{\mu}{\phi^2}\nabla f\big(\mathbf{x}^{(n)}\big)_S\Big) = \mathbf{x}^{(n+1)}.$$

By the assumption, we have

$$\min\bigl(\|\mathbf{x}^{(n)}-\mathbf{x}\|_2,\|\mathbf{x}^{(n)}+\mathbf{x}\|_2\bigr)\le \tfrac{1}{6}\|\mathbf{x}\|_2 \qquad \text{on } E_n.$$

Since $E_n\subset E_0$ and $\mathbf{x}^{(n+1)}=\eta(\mathbf{x}^{(n)})$, by Lemma 6.4, we have

$$\begin{aligned}&\min\bigl(\|\mathbf{x}^{(n+1)}-\mathbf{x}\|_2,\|\mathbf{x}^{(n+1)}+\mathbf{x}\|_2\bigr)\\&\quad\le\Bigl(1-\frac{\mu}{16}\Bigr)\min\bigl(\|\mathbf{x}^{(n)}-\mathbf{x}\|_2,\|\mathbf{x}^{(n)}+\mathbf{x}\|_2\bigr)\\&\qquad+C_0\frac{\mu\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}\le\frac{1}{6}\|\mathbf{x}\|_2 \qquad \text{on } E_n,\end{aligned}$$

provided $m\ge C(\sigma^2/\|\mathbf{x}\|_2^4)k\log p$ for a sufficiently large absolute constant C. Since $E_{n+1}\subset E_n$, and $\widehat{\mathbf{x}}^{(n+1)}=\mathbf{x}^{(n+1)}$ on E_{n+1}, we have on E_{n+1}

$$\begin{aligned}&\min_{i=0,1}\bigl\|\widehat{\mathbf{x}}^{(n+1)}-(-1)^i\mathbf{x}\bigr\|_2\\&\quad\le\Bigl(1-\frac{\mu}{16}\Bigr)\min_{i=0,1}\bigl\|\widehat{\mathbf{x}}^{(n)}-(-1)^i\mathbf{x}\bigr\|_2+C_0\frac{\mu\sigma}{\|\mathbf{x}\|_2}\sqrt{\frac{k\log p}{m}}\le\frac{1}{6}\|\mathbf{x}\|_2.\end{aligned}$$

□

Theorem 3.1 can be directly implied by Lemma 6.5. In fact, by Lemma 6.3, we know the initial condition in Lemma 6.5 holds. For all $t=1,2,3,\dots$, straight forward calculation yields

$$\frac{\min(\|\widehat{\mathbf{x}}^{(t)}-\mathbf{x}\|_2,\|\widehat{\mathbf{x}}^{(t)}+\mathbf{x}\|_2)}{\|\mathbf{x}\|_2}\le\frac{1}{6}\Bigl(1-\frac{\mu}{16}\Bigr)^t+C_0\frac{\sigma}{\|\mathbf{x}\|_2^2}\sqrt{\frac{k\log p}{m}} \qquad \text{on } E_t$$

for some universal constant C_0, where $\mathbb{P}(E_t)\ge 1-\frac{46}{m}-10e^{-k}-\frac{t}{mp^2}$.

APPENDIX A: PRELIMINARIES AND SUPPORTING LEMMAS

LEMMA A.1 ([5]). *Suppose $X_1,\dots,X_m$ are i.i.d. real-valued random variables obeying $X_i\le b$ for some absolute constant $b>0$, $\mathbb{E}X_i=0$ and $\mathbb{E}X_i^2=v^2$. Setting $\sigma^2=m(b^2\vee v^2)$,*

$$\mathbb{P}\{X_1+\cdots+X_m\ge y\}\le\exp\Bigl(-\frac{y^2}{2\sigma^2}\Bigr)\wedge c_0\bigl(1-\Phi(y/\sigma)\bigr)$$

where one can take $c_0=25$.

LEMMA A.2 (Proposition 34 [43]). *Suppose that $\mathbf{x}\sim\mathcal{N}(0,\mathbf{I}_n)$ is a standard normal random vector, and $f:\mathbb{R}^n\to\mathbb{R}$ is a 1-Lipschitz function. Then*

$$\mathbb{P}\bigl(f(\mathbf{x})-\mathbb{E}f(\mathbf{x})\ge t\bigr)\le e^{-t^2/2}.$$

LEMMA A.3 (Proposition 33 [43]). *Consider two centered Gaussian processes* $(X_t)_{t\in T}$ *and* $(Y_t)_{t\in T}$ *whose increments satisfy the inequality*

$$\mathbb{E}|X_s - X_t|^2 \le \mathbb{E}|Y_s - Y_t|^2$$

for all $s, t \in T$. *Then*

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t.$$

LEMMA A.4 (Proposition 35 [43]). *Let* $\mathbf{A}_S \in \mathbb{R}^{m\times p}$ *be defined in* (6.1). *Then, with probability at least* $1 - 2\exp(-t^2/2)$, *we have the following inequality*:

$$\|\mathbf{A}_S\| \le \sqrt{m} + \sqrt{k} + t. \tag{A.1}$$

LEMMA A.5. *Let* $\mathbf{A}_S \in \mathbb{R}^{m\times p}$ *be defined in* (6.1). *Then, with probability at least* $1 - 4\exp(-t^2/2)$, *the following inequalities hold*:

$$\|\mathbf{A}_S\|_{2\to 6} \le (15m)^{1/6} + \sqrt{k} + t, \tag{A.2}$$

and

$$\|\mathbf{A}_S\|_{2\to 4} \le (3m)^{1/4} + \sqrt{k} + t. \tag{A.3}$$

PROOF. The proof follows that of Theorem 32 in [43] step by step. Define $X_{\mathbf{u},\mathbf{v}} = \langle \mathbf{A}_S\mathbf{u}, \mathbf{v}\rangle$ on

$$T = \big\{(\mathbf{u},\mathbf{v}) : \mathbf{u}\in\mathbb{R}^p, \operatorname{supp}(U)\subset S, \|\mathbf{u}\|_2 = 1, \mathbf{v}\in\mathbb{R}^m, \|\mathbf{v}\|_{6/5} = 1\big\}.$$

Then $\|\mathbf{A}_S\|_{2\to 6} = \max_{(\mathbf{u},\mathbf{v})\in T} X_{\mathbf{u},\mathbf{v}}$. Define

$$Y_{\mathbf{u},\mathbf{v}} = \langle \mathbf{g}_S, \mathbf{u}\rangle + \langle \mathbf{h}, \mathbf{v}\rangle,$$

where $\mathbf{g}_S \in \mathbb{R}^p$ with $\operatorname{supp}(\mathbf{g}_S) = S$ and $\mathbf{h}\in\mathbb{R}^m$ are independent standard Gaussian random vectors.

For any $(\mathbf{u},\mathbf{v}), (\mathbf{u}',\mathbf{v}') \in T$, we have

$$\mathbb{E}|X_{\mathbf{u},\mathbf{v}} - X_{\mathbf{u}',\mathbf{v}'}| = \|\mathbf{v}\|_2^2 + \|\mathbf{v}'\|_2^2 - 2\langle \mathbf{u},\mathbf{u}'\rangle\langle \mathbf{v},\mathbf{v}'\rangle$$

and

$$\mathbb{E}|Y_{\mathbf{u},\mathbf{v}} - Y_{\mathbf{u}',\mathbf{v}'}| = 2 + \|\mathbf{v}\|_2^2 + \|\mathbf{v}'\|_2^2 - 2\langle \mathbf{u},\mathbf{u}'\rangle - \langle \mathbf{v},\mathbf{v}'\rangle.$$

Therefore,

$$\mathbb{E}|X_{\mathbf{u},\mathbf{v}} - X_{\mathbf{u}',\mathbf{v}'}| - \mathbb{E}|Y_{\mathbf{u},\mathbf{v}} - Y_{\mathbf{u}',\mathbf{v}'}| = 2\big(1 - \langle \mathbf{u},\mathbf{u}'\rangle\big)\big(1 - \langle \mathbf{v},\mathbf{v}'\rangle\big) \ge 0,$$

due to $\|\mathbf{u}\|_2 = \|\mathbf{u}'\|_2 = 1$, $\|\mathbf{v}\|_2 \le \|\mathbf{v}\|_{6/5} = 1$, and $\|\mathbf{v}'\|_2 \le \|\mathbf{v}'\|_{6/5} = 1$. Then by Lemma A.3, we have

$$\begin{aligned}\mathbb{E}\|\mathbf{A}_S\|_{2\to 6} &\le \mathbb{E}\max_{(\mathbf{u},\mathbf{v})\in T} Y_{\mathbf{u},\mathbf{v}} = \mathbb{E}\|\mathbf{g}_S\|_2 + \mathbb{E}\|\mathbf{h}\|_6 \\ &\le \sqrt{\mathbb{E}\|\mathbf{g}_S\|_2^2} + \big(\mathbb{E}\|\mathbf{h}\|_6^6\big)^{1/6} = \sqrt{k} + (15m)^{1/6}.\end{aligned}$$

Since $\|\cdot\|_{2\to 6}$ is a 1-Lipschitz function, by Lemma A.2, there holds with probability at least $1-2\exp(-t^2/2)$

$$\|\mathbf{A}_S\|_{2\to 6} \le \sqrt{k} + (15m)^{1/6} + t.$$

Similarly, with probability at least $1-2\exp(-t^2/2)$

$$\|\mathbf{A}_S\|_{2\to 4} \le \sqrt{k} + (3m)^{1/4} + t. \qquad \square$$

LEMMA A.6. *On an event with probability at least* $1-1/m$, *we have*

$$\left\| \frac{1}{m}\sum_{j=1}^{m} |\mathbf{a}'_{j\,S}\mathbf{x}|^2 \mathbf{a}_{j\,S}\mathbf{a}'_{j\,S} - \big(\|\mathbf{x}\|_2^2(\mathbf{I}_p)_S + 2\mathbf{x}\mathbf{x}'\big) \right\| \le \delta\|\mathbf{x}\|_2^2$$

provided $m \ge C(\delta)k\log k$, *where* $C(\delta)$ *is constant only depending on* δ. *Here*, $(\mathbf{I}_p)_S$ *by definition is a diagonal matrix with first* k *diagonal entries equal to* 1, *whereas other entries being* 0. *Furthermore*, *it implies that*

$$\frac{1}{m}\sum_{j=1}^{m} \big(\mathbf{a}'_{j\,S}\mathbf{x}\big)^2\big(\mathbf{a}'_{j\,S}\mathbf{h}\big)^2 \ge 2\big(\mathbf{x}'\mathbf{h}\big)^2 + (1-\delta)\|\mathbf{x}\|_2^2\|\mathbf{h}\|_2^2$$

for any $\mathbf{h}\in\mathbb{R}^p$ *that satisfies* $\operatorname{supp}(\mathbf{h})\subset S$.

The proof of this lemma is the same as that of Lemma 7.4 in [11].

LEMMA A.7. *Suppose* $\varepsilon_1,\dots,\varepsilon_m$ *are independent zero-mean sub-exponential random variables with*

$$\sigma := \max_{1\le i\le m} \|\varepsilon_i\|_{\psi_1}.$$

Then with probability at least $1-\frac{3}{m}$, *we have*

$$\left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j\right| \le C_0\sigma\sqrt{\frac{\log m}{m}}, \qquad \|\boldsymbol{\varepsilon}\|_\infty \le C_0\sigma\log m,$$

$$\left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j^2\right| \le C_0\sigma^2, \qquad \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j^4\right| \le C_0\sigma^4,$$

provided $m\ge m_0$ *for some numerical constants* C_0 *and* m_0.

PROOF. By Proposition 16 in [43], we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}\varepsilon_i\right| \ge t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{m\sigma^2}, \frac{t}{\sigma}\right)\right].$$

This implies that with probability at least $1 - \frac{2}{m^{10}}$, we have

$$\left|\sum_{i=1}^{m} \varepsilon_i\right| \le C_0\sigma \max(\sqrt{m\log m}, \log m) \le C_0\sigma\sqrt{m\log m}$$

provided $m \ge m_0$. This implies that

$$\left|\frac{1}{m}\sum_{j=1}^{m} \varepsilon_j\right| \le C_0\sigma\sqrt{\frac{\log m}{m}}.$$

By the basic properties of sub-exponential random variables, for each $j = 1, \dots, m$, we have

$$\mathbb{P}\big(|\varepsilon_j| \ge t\big) \le \exp\left(1 - c\frac{t}{\sigma}\right),$$

which implies that $|\varepsilon_j| \le C_0\sigma \log m$ with probability at least $1 - e/m^{11}$. This implies that

$$\|\boldsymbol{\varepsilon}\|_\infty \le C_0\sigma \log m$$

with probability at least $1 - e/m^{10}$.

Since

$$\sigma \ge \|\varepsilon_j\|_{\Psi_1} = \sup_{p\ge 1} p^{-1}\big(\mathbb{E}|\varepsilon_j|^p\big)^{1/p},$$

we have $\mathbb{E}\varepsilon_j^2 \le (2\sigma)^2$ and $\mathbb{E}\varepsilon_j^4 \le (4\sigma)^4$. Define

$$X = \frac{1}{m}\sum_{j=1}^{m} \varepsilon_j^2.$$

Then we have $\mathbb{E}X \le (2\sigma)^2$, and

$$\mathrm{Var}(X) \le (4\sigma)^4/m.$$

By Chebyshev's inequality, we have

$$\mathbb{P}\big(|X - \mathbb{E}X| \ge t\big) \le \frac{\mathrm{Var}(X)}{t^2}.$$

By letting $t = (4\sigma)^2$, we obtain that with probability at least $1 - 1/m$, we have $|X| \le 20\sigma^2$.

Similarly, with probability at least $1 - 1/m$, we have $|\frac{1}{m}\sum_{j=1}^{m} \varepsilon_j^4| \le C_0\sigma^4$ for some absolute constant C_0. $\square$

LEMMA A.8. *Suppose* $\mathbf{z}_j \in \mathbb{R}^k$, $j = 1, \ldots, m$ *are i.i.d. standard normal random vectors. For fixed* $\mathbf{a} \in \mathbb{R}^m$, *with probability at least* $1 - 2e^{-k}$, *we have*

$$\left\| \sum_{j=1}^m a_j \mathbf{z}_j \mathbf{z}_j' - \left(\sum_{j=1}^m a_j \right) \mathbf{I}_k \right\| \le C_0 \left(\sqrt{k \|\mathbf{a}\|_2^2} + k \|\mathbf{a}\|_\infty \right)$$

for some absolute constant C_0.

PROOF. Define $\mathbf{A} := \sum_{j=1}^m a_j \mathbf{z}_j \mathbf{z}_j' - (\sum_{j=1}^m a_j) \mathbf{I}_k$. By Lemma 4 in [43], we have

$$\|\mathbf{A}\| \le 2 \sup_{\mathbf{x} \in \mathcal{N}_{1/4}} |\mathbf{x}' \mathbf{A} \mathbf{x}|,$$

where $\mathcal{N}_{1/4}$ is the 1/4-net of the unit sphere $\mathcal{T}^{k-1}$.

For fixed $\mathbf{x} \in \mathcal{N}_{1/4}$, let $y_j = |\mathbf{z}_j' \mathbf{x}|^2 - 1$. Then

$$\mathbf{x}' \mathbf{A} \mathbf{x} = \sum_{j=1}^m a_j y_j.$$

Notice that y_j, $j = 1, \ldots, m$ are i.i.d. sub-exponential variables with $\|y_j\|_{\psi_1} \le K$ where K is an absolute constant. By Bernstein inequality (see, e.g., Proposition 16 in [43]), we have with probability at least $1 - 2\exp(-4k)$,

$$\left| \sum_{j=1}^m a_j y_j \right| \le (C_0/2) \left(\sqrt{k \|\mathbf{a}\|_2^2} + k \|\mathbf{a}\|_\infty \right)$$

for some absolute constant C_0.

Since $|\mathcal{N}_{1/4}| \le 9^k$, we know with probability at least $1 - 2e^{-k}$, we have

$$\|\mathbf{A}\| \le 2 \sup_{\mathbf{x} \in \mathcal{N}_{1/4}} |\mathbf{x}' \mathbf{A} \mathbf{x}| \le C_0 \left(\sqrt{k \|\mathbf{a}\|_2^2} + k \|\mathbf{a}\|_\infty \right).$$

□

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609

[2] ALEXEEV, B., BANDEIRA, A. S., FICKUS, M. and MIXON, D. G. (2014). Phase retrieval with polarization. *SIAM J. Imaging Sci.* **7** 35–66.

[3] ALON, N., KRIVELEVICH, M. and SUDAKOV, B. (1998). Finding a large hidden clique in a random graph. In *Proceedings of the Eighth International Conference "Random Structures and Algorithms"* (*Poznan*, 1997) **13** 457–466. MR1662795

[4] BAUSCHKE, H. H., COMBETTES, P. L. and LUKE, D. R. (2002). Phase retrieval, error reduction algorithm, and Fienup variants: A view from convex optimization. *J. Opt. Soc. Amer. A* **19** 1334–1345. MR1914365

[5] BENTKUS, V. (2003). An inequality for tail probabilities of martingales with differences bounded from one side. *J. Theoret. Probab.* **16** 161–173. MR1956826

[6] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res.* (*COLT*) **30** 1046–1066.

[7] BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27** 265–274. MR2559726

[8] CAI, J., CANDÈS, E. J. and SHEN, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* **20** 1956–1982. MR2600248

[9] CANDÈS, E. J., ELDAR, Y. C., STROHMER, T. and VORONINSKI, V. (2013). Phase retrieval via matrix completion. *SIAM J. Imaging Sci.* **6** 199–225. MR3032952

[10] CANDÈS, E. J. and LI, X. (2014). Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. *Found. Comput. Math.* **14** 1017–1026. MR3260258

[11] CANDÈS, E. J., LI, X. and SOLTANOLKOTABI, M. (2015). Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory* **61** 1985–2007. MR3332993

[12] CANDÈS, E. J., LI, X. and SOLTANOLKOTABI, M. (2015). Phase retrieval from coded diffraction patterns. *Appl. Comput. Harmon. Anal.* **39** 277–299. MR3352016

[13] CANDÈS, E. J., STROHMER, T. and VORONINSKI, V. (2013). PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.* **66** 1241–1274. MR3069958

[14] CHAI, A., MOSCOSO, M. and PAPANICOLAOU, G. (2011). Array imaging using intensity-only measurements. *Inverse Probl.* **27** 015005, 16. MR2746408

[15] CHEN, Y. and CANDÈS, E. C. (2015). *Solving Random Quadratic Systems of Equations is Nearly as Easy as Solving Linear Systems. Advances in Neural Information Processing Systems* **34** 739–747. Curran Associates, Red Hook, NY.

[16] CHEN, Y., CHI, Y. and GOLDSMITH, A. J. (2015). Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inform. Theory* **61** 4034–4059. MR3367819

[17] DAUBECHIES, I., DEFRISE, M. and DE MOL, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.* **57** 1413–1457. MR2077704

[18] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *PNAS* **110** 14557–14562.

[19] FIENUP, J. R. (1982). Phase retrieval algorithms: A comparison. *Appl. Opt.* **21** 2758–2769.

[20] GAO, C., MA, Z. and ZHOU, H. H. (2014) Sparse CCA: Adaptive estimation and computational barriers. Preprint. Available at arXiv:1409.8565.

[21] GERCHBERG, R. W. and SAXTON, W. O. (1972). A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* **35** 237–246.

[22] JAGANATHAN, K., OYMAK, S. and HASSIBI, B. (2012). On robust phase retrieval for sparse signals. In 50*th Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*), *Monticello* 794–799. IEEE, Piscataway, NJ.

[23] JAIN, P., TEWARIY, A. and KAR, P. (2014). On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems* **27** 685–693. Curran Associates, Red Hook, NY.

[24] JOHNSTONE, I. M. (2013). Gaussian estimation: Sequence and wavelet models. Available at http://www-stat.stanford.edu/~imj/.

[25] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

[26] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56** 2980–2998. MR2683452

[27] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785

[28] LECUÉ, G. and MENDELSON, S. (2013). Minimax rates of convergence and the performance of ERM in phase recovery. *Electron. J. Probab.* **20** 1–29.

[29] LEE, K., WU, Y. and BRESLER, Y. (2013). Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization. Preprint. Available at arXiv:1312.0525.

[30] LEVI, A. and STARK, H. (1984). Image restoration by the method of generalized projections with application to restoration from magnitude. *J. Opt. Soc. Amer. A* **1** 932–943. MR0758183

[31] LI, X. and VORONINSKI, V. (2013). Sparse signal recovery from quadratic measurements via convex programming. *SIAM J. Math. Anal.* **45** 3019–3033. MR3106479

[32] LOH, P. and WAINWRIGHT, M. J. (2015), Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

[33] MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

[34] MALEKI, A. and DONOHO, D. L. (2010). Optimally tuned iterative reconstruction algorithms for compressed sensing. *IEEE J. Sel. Top. Signal Process.* **4** 330–341.

[35] MARCHESINI, S., TU, Y. C. and WU, H. (2014). Alternating projection, ptychographic imaging and phase synchronization. *Appl. Comput. Harmon. Anal.* To appear. Available at arXiv:1402.0550.

[36] NEEDELL, D. and TROPP, J. A. (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **26** 301–321. MR2502366

[37] NETRAPALLI, P., JAIN, P. and SANGHAVI, S. (2015). Phase retrieval using alternating minimization. *IEEE Trans. Signal Process.* **63** 4814–4826.

[38] OYMAK, S., JALALI, A., FAZEL, M., ELDAR, Y. C. and HASSIBI, B. (2015). Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Trans. Inform. Theory* **61** 2886–2908. MR3342310

[39] SCHNITER, P. and RANGAN, S. (2012). Compressive phase retrieval via generalized approximate message passing. In 50*th Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*), *Monticello* 815–822. IEEE, Piscataway, NJ.

[40] SHECHTMAN, Y., BECK, A. and ELDAR, Y. C. (2014). GESPAR: Efficient phase retrieval of sparse signals. *IEEE Trans. Signal Process.* **62** 928–938. MR3160324

[41] SHECHTMAN, Y., ELDAR, Y. C., COHEN, O., CHAPMAN, H. N., MIAO, J. and SEGEV, M. (2014). Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Process. Mag.* **32**. To appear. Available at arXiv:1402.7350.

[42] SOLTANOLKOTABI, M. (2014). Algorithms and theory for clustering and nonconvex quadratic programming. PhD dissertatin, Stanford Univ., Stanford, CA.

[43] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[44] WALDSPURGER, I., D'ASPREMONT, A. and MALLAT, S. (2015). Phase recovery, maxcut and complex semidefinite programming. *Math. Program.* **149** 47–81. MR3300456

[45] WANG, T., BERTHET, Q. and SAMWORTH, R. J. (2014). Statistical and computational trade-offs in estimation of sparse principal components. *Ann. Statist.*. To appear. Available at arXiv:1408.5369.

[46] WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201. MR3269977

[47] YUAN, X., LI, P. and ZHANG, T. (2014). Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning* (*ICML* 2014), Beijing, China.

[48] YUAN, X. and ZHANG, T. (2013). Truncated power method for sparse eigenvalue problems. *J. Mach. Learn. Res.* **14** 899–925. MR3063614

T. T. CAI
Z. MA
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
400 JON M. HUNTSMAN HALL
3730 WALNUT STREET
PHILADELPHIA, PENNSYLVANIA 19104-6340
USA
E-MAIL: tcai@wharton.upenn.edu
zongming@wharton.upenn.edu

X. LI
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, DAVIS
4109 MATHEMATICAL SCIENCES
DAVIS, CALIFORNIA 95616
USA
E-MAIL: xdgli@ucdavis.edu