Mathematical Programming (2019) 176:545–571
https://doi.org/10.1007/s10107-019-01364-5

FULL LENGTH PAPER

Series B

Misspecified nonconvex statistical optimization for sparse phase retrieval

Zhuoran Yang[1] · Lin F. Yang[1] · Ethan X. Fang[2,3] · Tuo Zhao[4,5] · Zhaoran Wang[6] · Matey Neykov[7]

Received: 22 November 2017 / Accepted: 12 January 2019 / Published online: 18 February 2019

Abstract

Existing nonconvex statistical optimization theory and methods crucially rely on the correct specification of the underlying "true" statistical models. To address this issue, we take a first step towards taming model misspecification by studying the high-dimensional sparse phase retrieval problem with misspecified link functions. In particular, we propose a simple variant of the thresholded Wirtinger flow algorithm that, given a proper initialization, linearly converges to an estimator with optimal statistical accuracy for a broad family of unknown link functions. We further provide extensive numerical experiments to support our theoretical findings.

Mathematics Subject Classification 94A12 · 90C30 · 90C90

1 Introduction

We consider nonconvex optimization for high-dimensional phase retrieval with model misspecification. Phase retrieval finds applications in numerous scientific problems, for example, X-ray crystallography [22], electron microscopy [40], and diffractive imaging [5]. See, for example, Candès et al. [8] and the references therein for a detailed survey. Most of existing work [6,8] casts high-dimensional phase retrieval as a nonconvex optimization problem: Given n data points $\{(y^{(i)}, \boldsymbol{x}^{(i)}) \in \mathbb{R} \times \mathbb{R}^p\}_{i\in[n]}$,[1] one aims to solve

[1] Here we use the shorthand $[n] = \{1, 2, \ldots, n\}$.

Zhuoran Yang and Lin F. Yang have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10107-019-01364-5) contains supplementary material, which is available to authorized users.

✉ Ethan X. Fang
xxf13@psu.edu

Extended author information available on the last page of the article

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{n}\sum_{i=1}^{n}\big[y^{(i)} - (\boldsymbol{x}^{(i)\top}\boldsymbol{\beta})^2\big]^2, \quad \text{subject to } \|\boldsymbol{\beta}\|_0 \le s, \tag{1.1}$$

where $\|\boldsymbol{\beta}\|_0$ denotes the number of nonzero entries in $\boldsymbol{\beta}$.

The nonconvex problem in (1.1) gives rise to two challenges in optimization and statistics. From the perspective of optimization, (1.1) is NP-hard in the worst case [48], that is, under computational hardness hypotheses, no algorithm can achieve the global minimum in polynomial time. Particularly, most existing general-purpose first-order or second-order optimization methods [3,15,16,19,23,37,60] are only guaranteed to converge to certain stationary points. Meanwhile, since (1.1) can also be cast as a polynomial optimization problem, we can leverage various semidefinite programming approaches [1,30,44,59]. However, in real applications the problem dimension of practical interest is often large, for example, p can be of the order of millions. To the best of our knowledge, existing polynomial optimization approaches do not scale up to such large dimensions. The difficulty in optimization further leads to more challenges in statistics. From the perspective of statistics, researchers are interested in characterizing the statistical properties of $\hat{\boldsymbol{\beta}}$ with respect to some underlying ground truth $\boldsymbol{\beta}^*$, for example, the estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. Nevertheless, due to the lack of global optimality in nonconvex optimization, the statistical properties of the solutions obtained by existing algorithms remain rather difficult to analyze.

Recently, Cai et al. [6] proposed a thresholded Wirtinger flow (TWF) algorithm to tackle the problem, which essentially employs proximal-type iterations. TWF starts from a carefully specified initial point, and iteratively performs gradient descent steps. In particular, at each iteration, TWF performs a thresholding step to preserve the sparsity of the solution. Cai et al. [6] further prove that TWF achieves a linear rate of convergence to an approximate global minimum that has optimal statistical accuracy. Note that Cai et al. [6] can establish such strong theoretical results because their algorithm and analysis exploit the underlying "true" data generating process of sparse phase retrieval, which enables bypassing the computational hardness barrier. In specific, they assume that the response $Y \in \mathbb{R}$ and the covariate $\boldsymbol{X} \in \mathbb{R}^p$ satisfy

$$\boldsymbol{X} \sim N(0, \mathbf{I}_p) \quad \text{and} \quad Y = (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2 + \epsilon, \tag{1.2}$$

in which $\epsilon \sim N(0, \sigma^2)$ and $\|\boldsymbol{\beta}^*\|_0 \le s$. The data points $\{(y^{(i)}, \boldsymbol{x}^{(i)}) \in \mathbb{R}\times\mathbb{R}^p\}_{i\in[n]}$ are n independent realizations of $(Y, \boldsymbol{X})$, and n needs to be sufficiently large, for example, $n = \Omega(s^2 \log p)$. Intuitively speaking, these distributional and scaling assumptions essentially rule out many worst-case instances of (1.1), and hence ease the theoretical analysis. Such a statistical view of nonconvex optimization, however, heavily relies on the correct model specification that the data points are indeed generated by the model defined in (1.2). To the best of our knowledge, we are not aware of any existing results on nonconvex optimization under misspecification of the underlying statistical models.

In this paper, we propose a simple variant of the TWF algorithm to tackle problem (1.1) under statistical model misspecification. In specific, we prove that the proposed algorithm is robust to certain model misspecifications, and achieves a linear convergence rate to an approximate global minimum with optimal statistical accuracy. To be more precise, we focus our discussion on the single index model (SIM), which assumes that the response variable $Y \in \mathbb{R}$ and the covariate $\boldsymbol{X} \sim N(0, \mathbf{I}_p)$ are linked through

$$Y = h(\boldsymbol{X}^\top \boldsymbol{\beta}^*, \epsilon), \tag{1.3}$$

in which $\boldsymbol{\beta}^*$ is the parameter of interest, ϵ is the random noise, and $h\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a link function. Existing work on phase retrieval requires h to be known and the entire model to be correctly specified. Interesting examples of h include $h(u, v) = u^2 + v$ as in (1.2), or $h(u, v) = |u| + v$ and $h(u, v) = |u + v|$, where, if the model is correctly specified, the data points are generated by

$$Y = |\boldsymbol{X}^\top \boldsymbol{\beta}^*| + \epsilon \;\; \text{and} \;\; Y = |\boldsymbol{X}^\top \boldsymbol{\beta}^* + \epsilon|, \tag{1.4}$$

respectively. In contrast with existing work, here we do not assume any specific form of h. The only assumption that we impose on h is that Y and $(\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2$ have non-null covariance, i.e.,

$$\mathrm{Cov}\big[Y, (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2\big] \neq 0, \tag{1.5}$$

The condition in (1.5) is quite mild and encompasses all aforementioned examples. Perhaps surprisingly, for this highly misspecified statistical model, we prove that the proposed variant of TWF still achieves a linear rate of convergence to an approximate global minimum that has optimal statistical accuracy. Thus, our results lead to new understandings of the robustness of nonconvex optimization algorithms.

Note that the model in (1.3) is not identifiable, because the norm of $\boldsymbol{\beta}^*$ can be absorbed into the unknown function h while preserving the same value of Y. We henceforth assume that $\|\boldsymbol{\beta}^*\|_2 = 1$ to make the model identifiable. Meanwhile, note that the moment condition in (1.5) remains the same if we replace $\boldsymbol{\beta}^*$ by $-\boldsymbol{\beta}^*$. Hence, even under the assumption that $\|\boldsymbol{\beta}^*\|_2 = 1$, the model in (1.2) is only identifiable up to the global sign of $\boldsymbol{\beta}^*$. Similar phenomenon also appears in the classical phase retrieval model in (1.2).

Major contributions Our contributions include:

- We propose a variant of the thresholded Wirtinger flow algorithm for misspecified phase retrieval, which reduces to the TWF algorithm in Cai et al. [6] when the true model is correctly specified as in (1.2).
- We establish unified computational and statistical results for the proposed algorithm. In specific, initialized with a thresholded spectral estimator, the proposed algorithm converges linearly to an estimator with optimal statistical accuracy. Our theory illustrates the robustness of TWF, which sheds new light on the wide empirical success of nonconvex optimization algorithms on real problems with possible model misspecification [33].

Recent advance on nonconvex statistical optimization largely depends on the correct specification of statistical models [20,34,36,52,53,57,58,63,64,66]. To the best of our knowledge, this paper establishes the first nonconvex optimization algorithm and theory that incorporate model misspecification.

Related work The problem of recovering a signal from the magnitude of its linear measurements has broad applications, including electron microscopy [11], optical imaging [49], and X-ray crystallography [38,41]. There exists a large body of literature on the statistical and computational aspects of phase retrieval. In specific, Candès et al. [7,9] and Waldspurger et al. [56] establish efficient algorithms based upon semidefinite programs. For nonconvex methods, Netrapalli et al. [42] and Waldspurger [55] propose alternating minimization algorithms and Sun et al. [51] develop a trust-region algorithm. Also, Candès et al. [8] establish the Wirtinger flow algorithm, which performs gradient descent on the nonconvex loss function. Such an algorithm is then further extended by Chen and Candès [10] and Zhang et al. [65] to even more general settings. Furthermore, for sparse phase retrieval in high dimensions, Cai et al. [6] propose the truncated Wirtinger flow algorithm and prove that it achieves the optimal statistical rate of convergence under correct model specification. For a more detailed survey, we refer interested readers to Jaganathan et al. [25] and the references therein.

Our work is also closely related to SIM, which is extensively studied in statistics. See, for example, Han [21], McCullagh and Nelder [39], Horowitz [24] and the references therein. Most of this line of work estimates the parameter $\boldsymbol{\beta}^*$ and the unknown link function jointly through M-estimation. However, these M-estimators are often formulated as the global minima of nonconvex optimization problems, and hence are often computationally intractable to obtain. A more related line of research is sufficient dimension reduction, which assumes that Y only depends on $\boldsymbol{X}$ through the projection of $\boldsymbol{X}$ onto a subspace $\mathcal{U}$. SIM falls into such a framework with $\mathcal{U}$ being the subspace spanned by $\boldsymbol{\beta}^*$. See Li and Duan [33], Li [31], Cook and Ni [12] and the references therein for details. Most work in this direction proposes spectral methods based on the conditional covariance of $\boldsymbol{X}$ given Y, which is difficult to estimate consistently in high dimensions. In fact, the moment condition in (1.5) is inspired by Li [32], which studies the case with $\mathbb{E}(Y\boldsymbol{X}^\top\boldsymbol{\beta}^*) = 0$. Note that this condition also holds for phase retrieval models.

Moreover, the recent success of high-dimensional regression [4] also sparks the study of SIMs in high dimensions. Plan et al. [46], Plan and Vershynin [45], Thrampoulidis et al. [54], Goldstein et al. [17] and Genzel [14] propose to estimate the direction of $\boldsymbol{\beta}^*$ using least squares regression with ℓ_1-regularization. They show that the ℓ_1-regularized estimator enjoys optimal statistical rate of convergence. However, their method depends on the pivotal condition that $\mathrm{Cov}(Y, \boldsymbol{X}^\top\beta^*) \neq 0$, which is not satisfied by the phase retrieval model. A more related work is Neykov et al. [43], which propose a two-step estimator based on convex optimization. Specifically, their estimator is constructed via refining the solution of a semidefinite program by ℓ_1-regularized regression. Compared with our method, their estimator incurs higher computational cost. In addition, Yang et al. [61,62] propose efficient estimators for high-dimensional SIMs with non-Gaussian covariates based on convex optimization. Furthermore, Jiang and Liu [26], Lin et al. [35] and Zhu et al. [67] extend sliced inverse regression [31] to the high-dimensional setting. However, they mainly focus on support recovery and consistency properties. Moreover, there is a line of work focussing on estimating both the parametric and the nonparametric component [2,28,29,47]. These methods are either not applicable to our model or have suboptimal rates.

In summary, although there is a large body of related literature, the success of existing nonconvex optimization algorithms for high-dimensional phase retrieval largely relies on the correct specification of the underlying generative models. By establishing connections with SIM, we propose and analyze a simple variant of TWF, which is provable robust to misspecified models.

Notation For an integer m, we use $[m]$ to denote the set $\{1, 2, \ldots, m\}$. Let $S \subseteq [m]$ be a set, we define $S^c = [m]\backslash S$ as its complement. Let $\boldsymbol{x} \in \mathbb{R}^p$ be a vector, we denote by $\boldsymbol{x}_i$ the i-th coordinate of $\boldsymbol{x}$. We also take $\boldsymbol{x}_S$ as the projection of $\boldsymbol{x}$ onto the index set S, i.e., $[\boldsymbol{x}_S]_i = \boldsymbol{x}_i \cdot \mathbb{1}(i \in S)$ for all $i \in [p]$. Here $[\boldsymbol{x}_S]_i$ is the i-th entry of $\boldsymbol{x}_S$. Furthermore, we denote by $\|\boldsymbol{x}\|_0$ the number of non-zero entries of $\boldsymbol{x}$, that is, $\|\boldsymbol{x}\|_0 = \sum_{j\in[p]} \mathbb{1}\{\boldsymbol{x}_j \neq 0\}$. We denote the support of $\boldsymbol{x}$ to be $\text{supp}(\boldsymbol{x})$, which is defined as $\{j : \boldsymbol{x}_j \neq 0\}$. Hence, we have that $\|\boldsymbol{x}\|_0 = |\text{supp}(\boldsymbol{x})|$. We denote by $\|\cdot\|$ the ℓ_2-norm of a vector, by $\|\cdot\|_2$ the spectral (operator) norm of a matrix, by $\|\cdot\|_{\max}$ the elementwise ℓ_∞-norm of a matrix, and by $\|\cdot\|_{\text{fro}}$ the Frobenius norm of a matrix. We also employ $\|\cdot\|_\alpha$ to denote the ℓ_α-norm of a vector for any $\alpha \geq 1$. Furthermore, let $\alpha, \beta \in [1, \infty)$, we use $\|\cdot\|_{\alpha\to\beta}$ to denote the induced operator norm of a matrix, i.e., let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, then $\|\boldsymbol{A}\|_{\alpha\to\beta} = \sup_{\boldsymbol{x}\neq\boldsymbol{0}}\{\|\boldsymbol{A}\boldsymbol{x}\|_\beta / \|\boldsymbol{x}\|_\alpha\}$. For two random variables X and Y, we write $X \perp\!\!\!\perp Y$ when X is independent of Y. We use $\text{Var}(X)$ to denote the variance of X, and use $\text{Cov}(X, Y)$ to denote the covariance between random variables X and Y. Also, let $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$ be two positive sequences. If there exists some constant C such that $\limsup_{n\to\infty} a_n/b_n \leq C$, we write $a_n = \mathcal{O}(b_n)$, or equivalently, $b_n = \Omega(a_n)$. We write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$.

Definition 1.1 (*Sub-exponential variable and ψ_1-norm*) A random variable $X \in \mathbb{R}$ is called sub-exponential if its ψ_1-norm is bounded. The ψ_1-norm of X is defined as $\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}$.

2 Nonconvex optimization for misspecified phase retrieval

We consider the phase retrieval problem under model misspecification. Specifically, given the covariate $\boldsymbol{X} \in \mathbb{R}^p$ and the signal parameter $\boldsymbol{\beta}^* \in \mathbb{R}^p$, we assume that the response variable Y is given by the single index model in (1.3) for some unknown link function $h\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, where ϵ is the random noise which is assumed to be independent of $\boldsymbol{X}$. Throughout our discussion, for simplicity, we assume each covariate is sampled from $N(0, \mathbf{I}_p)$. Since the norm of $\boldsymbol{\beta}^*$ can be incorporated into h, we impose an additional constraint $\|\boldsymbol{\beta}^*\| = 1$, in order to make the model partially identifiable. Our goal is to estimate $\boldsymbol{\beta}^*$ using n i.i.d. realizations of $(\boldsymbol{X}, Y)$, where we denote the set of n samples by $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i\in[n]}$, and we do not have the knowledge of h in advance. Throughout this paper, we consider a high-dimensional and sparse regime where $\boldsymbol{\beta}^*$ has at most s non-zero entries for some $s < n$, and the dimensionality p can be much larger than the sample size n.

2.1 Motivation

Our modified TWF algorithm is inspired by the following optimization problem

$$\overline{\boldsymbol{\beta}} = \mathop{\mathrm{argmin}}_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathrm{Var}\big[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2\big], \tag{2.1}$$

which aims at estimating $\boldsymbol{\beta}^*$ by minimizing the variability of $Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2$. Intuitively, the variability would be minimized by a vector $\boldsymbol{\beta}$ which is parallel to $\boldsymbol{\beta}^*$. To see this, first note that Y depends on $\boldsymbol{X}$ only through $\boldsymbol{X}^\top \boldsymbol{\beta}^*$. If $\boldsymbol{\beta}$ is not parallel to $\boldsymbol{\beta}^*$, $\boldsymbol{X}^\top \boldsymbol{\beta}$ would contain a component which is independent of $\boldsymbol{X}^\top \boldsymbol{\beta}^*$ and Y, which increases the variability of $Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2$. The following proposition justifies this intuition by showing that (2.1) has two global minima which are both parallel to $\boldsymbol{\beta}^*$. This proposition forms the basis for our modified version of the TWF algorithm.

Proposition 2.1 *Suppose* $\mathrm{Cov}[Y, (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2] = \rho > 0$. *Let* $\overline{\boldsymbol{\beta}}$ *be defined in* (2.1). *Then we have* $\overline{\boldsymbol{\beta}} = \boldsymbol{\beta}^* \cdot \sqrt{\rho/2}$ *or* $\overline{\boldsymbol{\beta}} = -\boldsymbol{\beta}^* \cdot \sqrt{\rho/2}$.

Sketch of the Proof We first decompose $\boldsymbol{\beta}$ into $\boldsymbol{\beta} = \zeta \boldsymbol{\beta}^* + \boldsymbol{\beta}^\perp$, where $\zeta = \boldsymbol{\beta}^\top \boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^\perp$ is perpendicular to $\boldsymbol{\beta}^*$. Since $\boldsymbol{X} \sim N(0, \mathbf{I}_p)$, we have $\boldsymbol{X}^\top \boldsymbol{\beta}^* \perp\!\!\!\perp \boldsymbol{X}^\top \boldsymbol{\beta}^\perp$ and $Y \perp\!\!\!\perp \boldsymbol{X}^\top \boldsymbol{\beta}^\perp$. In addition, the norm of $\boldsymbol{\beta}^\perp$ is given by $\|\boldsymbol{\beta}^\perp\|_2^2 = \|\boldsymbol{\beta}\|_2^2 - \zeta^2$. By expanding the variance term in (2.1) and some simple manipulations, we obtain

$$\mathrm{Var}\big[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2\big] = \mathrm{Var}(Y) - 2\zeta^2 \rho + 2\|\boldsymbol{\beta}\|^4. \tag{2.2}$$

To find the minimizer of the right-hand side in (2.2), first we note that $\boldsymbol{\beta} = \mathbf{0}$ cannot be a minimizer. To see this, by direct computation, if we let $\tilde{\boldsymbol{\beta}} = c\boldsymbol{\beta}^*$ with $0 < c < \sqrt{\rho}$, it can be shown that

$$\mathrm{Var}\big[Y - (\boldsymbol{X}^\top \tilde{\boldsymbol{\beta}})^2\big] < \mathrm{Var}\big[Y - (\boldsymbol{X}^\top \mathbf{0})^2\big].$$

Now we fix $\|\boldsymbol{\beta}\| > 0$ as a constant. Then $|\zeta| \le \|\boldsymbol{\beta}\|$ by definition. By (2.2), $\mathrm{Var}[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2]$ can be viewed as a function of ζ. The minimum is achieved only when $\zeta^2 = \|\boldsymbol{\beta}\|^2$, which implies $\boldsymbol{\beta}^\perp = \mathbf{0}$. In other words, any minimizer $\overline{\boldsymbol{\beta}}$ given by (2.2) is parallel to $\boldsymbol{\beta}^*$.

Now we set $|\zeta| = \|\boldsymbol{\beta}\|$ in the right-hand side of (2.2), which now becomes a function of $\|\boldsymbol{\beta}\|$. We minimize it with respect to $\|\boldsymbol{\beta}\|$ and the minimizer is given by $\|\boldsymbol{\beta}\| = \sqrt{\rho/2}$. Therefore, we prove that the minima of $\mathrm{Var}[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2]$ are $\boldsymbol{\beta}^* \sqrt{\rho/2}$ and $-\boldsymbol{\beta}^* \sqrt{\rho/2}$, which concludes the proof. See §B.1 for a detailed proof. □

Proposition 2.1 suggests that estimating $\boldsymbol{\beta}^*$ up to a proportionality constant is feasible by minimizing (2.1) even without the knowledge of h. Specifically, it suggests that $\mathrm{Var}[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2]$ is a reasonable loss function on the "population level" (i.e., with infinite samples) for the class of SIMs satisfying the moment condition in (1.5). In addition, we note that the assumption that $\rho > 0$ is not essential. The reason is, if $\rho < 0$, one could apply the same logic to $-Y$. That is, estimate $\boldsymbol{\beta}^*$ by minimizing

$\mathrm{Var}[Y + (\boldsymbol{X}^\top \boldsymbol{\beta})^2]$. In the following, we explain the reason why it is preferable to use $\mathrm{Var}[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2]$ over $\mathbb{E}[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2]^2$ as a population loss function.

As we discussed in the introduction, the original motivation of the TWF algorithm is based on the likelihood perspective, which leads to the least squares loss function, i.e., $\mathbb{E}[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2]^2$. When model (1.2) is correctly specified, and the random noise ϵ has mean zero, the least squares loss function coincides with the variance loss in (2.1). However, in cases where the expectation of ϵ is negative, the minimizer of the least squares loss function could be $\boldsymbol{\beta} = 0$. In this case, minimizing the least squares loss function will lead to uninformative results. In addition, it is not hard to see that the original TWF algorithm of Cai et al. [6] is not well-defined in cases where $\mathbb{E}(Y) \le 0$. Cai et al. [6] do not exhibit this problem since they assume that model (1.2) is correctly specified, and that ϵ is centered at 0. These results implies that $\mathbb{E}(Y)$ is positive. In summary, the variance loss in (2.1) is more appropriate for robust estimation over the misspecified phase retrieval models. Note that the work by Chen and Candès [10] gives a Truncated Wirtinger Flow algorithm, which is similar to ours. However, instead of doing thresholding, they truncate terms in the summation formula. For instance, they only keep the terms that satisfying some properties. From the theoretical perspective, they do not achieve the robustness to link-function misspecification as ours.

In addition to relating $\boldsymbol{\beta}^*$ to the minimizer of the variance loss function, in the following proposition, we show that the moment condition in (1.5) implies spectral method can be used to estimate $\boldsymbol{\beta}^*$ and ρ simultaneously.

Proposition 2.2 *Assuming* $\mathrm{Cov}[Y, (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2] = \rho$, *we have*

$$\mathbb{E}[(Y - \mu)\boldsymbol{X}\boldsymbol{X}^\top] = \mathbb{E}[Y(\boldsymbol{X}\boldsymbol{X}^\top - \mathbf{I}_p)] = \rho \cdot \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\top}, \tag{2.3}$$

where $\mu = \mathbb{E}(Y)$. *Thus* $\pm \boldsymbol{\beta}^*$ *are the eigenvector of* $\mathbb{E}[(Y - \mu)\boldsymbol{X}\boldsymbol{X}^\top]$ *corresponding to eigenvalue* ρ, *and* ρ *is the largest eigenvalue in magnitude.*

Proof of Proposition 2.2 The first identity follows from direct calculation. For any $\mathbf{v} \in \mathbb{R}^p$, we decompose $\mathbf{v}$ into $\mathbf{v} = \zeta \boldsymbol{\beta}^* + \boldsymbol{\beta}^\perp$, where $\zeta = \mathbf{v}^\top \boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^\perp$ is perpendicular to $\boldsymbol{\beta}^*$. Since $\boldsymbol{X} \sim N(0, \mathbf{I}_p)$, $\boldsymbol{X}^\top \boldsymbol{\beta}^\perp$ is independent of $\boldsymbol{X}^\top \boldsymbol{\beta}^*$ and Y. By direct computation, we have

$$\mathbb{E}[(Y - \mu) \cdot \mathbf{v}^\top \boldsymbol{X}\boldsymbol{X}^\top \mathbf{v}] = \zeta^2 \rho = \rho \mathbf{v}^\top \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\top} \mathbf{v}, \tag{2.4}$$

which establishes (2.3). In addition, (2.3) implies that ρ is the eigenvalue of $\mathbb{E}[(Y - \mu)\boldsymbol{X}\boldsymbol{X}^\top]$ with the largest magnitude, and $\pm \boldsymbol{\beta}^*$ are the eigenvector corresponding to eigenvalue ρ. This concludes the proof of this proposition. □

As shown in Proposition 2.2, an alternative way of estimating ρ and $\boldsymbol{\beta}^*$ is to estimate the largest eigenvalue of $\mathbb{E}[(Y - \mu)\boldsymbol{X}\boldsymbol{X}^\top]$ in magnitude and the corresponding eigenvector. In the next section, we establish the sample versions of the optimization problems outlined by Propositions 2.1 and 2.2. These two problems correspond to the two stages of our modified version of the TWF algorithm, respectively.

2.2 Thresholded Wirtinger flow revisited

In this section, we present the modified TWF algorithm for misspecified phase retrieval. Note that In Sect. 2.1, we propose to use $\mathrm{Var}[Y-(\boldsymbol{X}^\top\boldsymbol{\beta})^2]$ as a more robust population loss function in comparison with the least squares loss function $\mathbb{E}[Y-(\boldsymbol{X}^\top\boldsymbol{\beta})^2]^2$ used by the classical Wirtinger Flow algorithm in [8]. In the following, we first establish the sample version of the minimization problem in (2.1).

Recall that we assume that $\boldsymbol{\beta}^*$ has at most s non-zero entries. For the time being, suppose that $\rho > 0$, which will be immediately relaxed in Sect. 2.3. Given $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I}_p)$ and $\mu = \mathbb{E}(Y)$, we have

$$\mathbb{E}\big[Y-(\boldsymbol{X}^\top\boldsymbol{\beta})^2\big] = \mu - \boldsymbol{\beta}^\top\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{\beta} = \mu - \|\boldsymbol{\beta}\|^2. \tag{2.5}$$

The sample version of $\mu - \|\boldsymbol{\beta}\|^2$ is defined by $\xi_n(\boldsymbol{\beta}) = \mu_n - \|\boldsymbol{\beta}\|^2$, where $\mu_n = n^{-1}\sum_{i=1}^n y^{(i)}$ is the sample mean of the response variables. We consider the following nonconvex optimization problem

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\|\boldsymbol{\beta}\|_0 \le s} \ell_n(\boldsymbol{\beta}), \quad \text{where} \quad \ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \big[y^{(i)} - (\boldsymbol{x}^{(i)\top}\boldsymbol{\beta})^2 - \xi_n(\boldsymbol{\beta})\big]^2. \tag{2.6}$$

Here the loss function $\ell_n(\boldsymbol{\beta})$ is the sample version of the variance loss function $\mathrm{Var}[Y-(\boldsymbol{X}^\top\boldsymbol{\beta})^2]$. In order to solve (2.6) efficiently, similar to TWF algorithm in Cai et al. [6], we propose a gradient-based algorithm for the loss function $\ell_n(\boldsymbol{\beta})$. In addition, due to the cardinality constraint in (2.6), in each step, we apply a thresholding operator to the estimate after a gradient descent step. Note that the constraint in (2.6) involves $s = \|\boldsymbol{\beta}^*\|_0$, which is usually unknown. By selecting the threshold value adaptively, our proposed TWF algorithm actually does not require any prior knowledge of s, and can iteratively determine the sparsity of the estimate. Specifically, at the k-th iteration, we compute a threshold value

$$\tau(\boldsymbol{\beta}^{(k)}) = \kappa\left\{\frac{\log(np)}{n^2}\sum_{i=1}^n \Big[y^{(i)} - \big(\boldsymbol{x}^{(i)\top}\boldsymbol{\beta}^{(k)}\big)^2 - \mu_n + \|\boldsymbol{\beta}^{(k)}\|^2\Big]^2 \big(\boldsymbol{x}^{(i)\top}\boldsymbol{\beta}^{(k)}\big)^2\right\}^{1/2}, \tag{2.7}$$

where κ is an appropriately chosen constant. We then take a thresholded gradient descent step

$$\boldsymbol{\beta}^{(k+1)} = \mathcal{T}_{\eta\cdot\tau(\boldsymbol{\beta}^{(k)})}\big[\boldsymbol{\beta}^{(k)} - \eta\nabla\ell_n(\boldsymbol{\beta}^{(k)})\big], \tag{2.8}$$

where $\eta > 0$ is the step size and $[\mathcal{T}_\tau(\boldsymbol{w})]_j = \boldsymbol{w}_j \cdot \mathbb{1}(|\boldsymbol{w}_j| \ge \tau)$ is the hard-thresholding operator with threshold value $\tau > 0$. The gradient $\nabla\ell_n(\cdot)$ in (2.8) is given by

$$\nabla\ell_n(\boldsymbol{\beta}) = \frac{4}{n}\sum_{i=1}^n \big[y^{(i)} - (\boldsymbol{x}^{(i)\top}\boldsymbol{\beta})^2 - \xi_n(\boldsymbol{\beta})\big]\big(\mathbf{I}_p - \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\top}\big)\boldsymbol{\beta}. \tag{2.9}$$

After running the algorithm for T iterations for some sufficiently large T, the last estimate $\boldsymbol{\beta}^{(T)}$ is standardized by $\boldsymbol{\beta}^{(T)}/\|\boldsymbol{\beta}^{(T)}\|$, and we take it as the final estimate of $\boldsymbol{\beta}^*$ (recall that we assume $\|\boldsymbol{\beta}^*\| = 1$). We note that the gradient step in (2.8) requires starting from an proper initialization $\boldsymbol{\beta}^{(0)}$. In the following section, we present a thresholded spectral algorithm to obtain a good initializer.

2.3 Initialization via thresholded spectral method

Motivated by Proposition 2.2, we introduce the thresholded spectral method (TSM) to obtain an appropriate initialization for the TWF algorithm. Note that $\boldsymbol{\beta}^*$ is the leading eigenvector of $\mathbb{E}[(Y-\mu)\boldsymbol{X}\boldsymbol{X}^\top]$ when ρ is positive. A natural idea is to apply spectral methods to the empirical counterpart of matrix $\mathbb{E}[(Y-\mu)\boldsymbol{X}\boldsymbol{X}^\top]$ based on the samples. However, since $\mathbb{E}[(Y-\mu)\boldsymbol{X}\boldsymbol{X}^\top]$ is a high-dimensional matrix, the leading eigenvector this sample matrix can have large estimation error. To resolve this issue, we apply spectral methods on a submatrix selected by the following screening step.

Specifically, we first conduct a thresholding step to select a subset of coordinates $\hat{S}_0 \subseteq [p]$ by

$$\hat{S}_0 = \left\{ j \in [p] : \left| \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \big[(\boldsymbol{x}_j^{(i)})^2 - 1 \big] \right| > \gamma \sqrt{\log(np)/n} \right\}, \tag{2.10}$$

where γ is an appropriately chosen constant. The thresholding step in (2.10) is motivated by Proposition 2.2, which shows that the diagonal entries of the matrix $\mathbb{E}[Y(\boldsymbol{X}\boldsymbol{X}^\top - \mathbf{I})]$ are non-zero on the support of $\boldsymbol{\beta}^*$. When n is sufficiently large, $n^{-1}\sum_{i=1}^n y^{(i)}[(\boldsymbol{x}_j^{(i)})^2 - 1]$ is close to its expectation for all $j \in [p]$. Thus one would expect that, with an appropriately chosen $\gamma > 0$, $|\mathbb{E}[Y \cdot (\boldsymbol{X}_j^2 - 1)]| > 0$ for all $j \in \hat{S}_0$. This implies that the thresholding step constructs an $\hat{S}_0$, which is a subset of the true support $\mathrm{supp}(\boldsymbol{\beta}^*)$. We remark that this step is closely related to the diagonal thresholding algorithm for sparse principle component analysis [27].

After obtaining $\hat{S}_0$, to simplify the notation, we denote $\boldsymbol{x}_{\hat{S}_0}^{(i)}$ by $\boldsymbol{w}^{(i)}$ for each $i \in [n]$. Recall that μ_n is the sample average of $y^{(1)}, \ldots, y^{(n)}$. We define $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ by

$$\boldsymbol{W} = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \mu_n) \cdot \boldsymbol{w}^{(i)} (\boldsymbol{w}^{(i)})^\top. \tag{2.11}$$

Notice that the the $|\hat{S}_0| \times |\hat{S}_0|$ non-zero entries of $\boldsymbol{W}$ form the submatrix of $n^{-1}\sum_{i=1}^n (y^{(i)} - \mu_n) \cdot \boldsymbol{x}^{(i)} (\boldsymbol{x}^{(i)})^\top$ with both rows and columns in $\hat{S}_0$. Let $\hat{\boldsymbol{v}}$ be the the eigenvector of $\boldsymbol{W}$ corresponding to the largest eigenvalue in magnitude. Finally, we define the initial estimator $\boldsymbol{\beta}^{(0)}$ by

$$\boldsymbol{\beta}^{(0)} = \hat{\boldsymbol{v}} \sqrt{|\rho_n|/2} \quad \text{where} \quad \rho_n = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \big(\boldsymbol{x}^{(i)\top} \hat{\boldsymbol{v}} \big)^2 - \mu_n. \tag{2.12}$$

With this initial estimator, the algorithm proceeds with thresholded gradient descent defined in (2.8) and outputs the normalized estimator $\boldsymbol{\beta}^{(T)}/\|\boldsymbol{\beta}^{(T)}\|$. In the next section, we establish the statistical and computational rates of convergence for the TWF algorithm.

Remark 1 In the phase retrieval problem, the relative signs/phases essentially capture the difficulties of solving this problem. For instance, the solution of the real case is identifiable up to a global sign. For the complex case, the final solution is identifiable up to a global phase, which is very similar to the real case. The key ideas of this paper extend naturally to the complex case. In particular, for the complex case, our objective (2.1), thresholding function (2.7) and gradient (2.9) all remain similar (slight modifications are needed, e.g., change transpose operation to conjugate and change the absolute value to the modulus of the complex variable). Other differences are the concentration inequalities and properties of the input distribution. However, these are not essential for the establishment of the algorithm. For conciseness, we focus on the real cases in this paper.

3 Statistical and computational guarantees

In this section, we show that the TWF algorithm is robust for misspecified phase retrieval models. In particular, we prove that the proposed algorithm linearly converges to an estimator, and the estimator achieves optimal statistical rate of convergence even under the unknown link function. Before going further, we first impose the following assumption on the distribution of Y to facilitate our discussion.

Assumption 3.1 The response variable Y of the misspecified phase retrieval model in (1.3) follows a sub-exponential distribution as specified in Definition 1.1. Specifically, we assume that $\|Y\|_{\psi_1} \leq \Psi$ for some constant $\Psi > 0$. In addition, we assume that $\mathrm{Cov}[Y, (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2] = \rho \neq 0$ for some constant ρ.

The assumption that Y has sub-exponential tail is mild, and is satisfied if both the link function h and the random noise ϵ are well-behaved. For example, if ϵ is a sub-exponential random variable, and $h(z, w)$ is Lipschitz continuous in both z and w, this assumption is satisfied. Meanwhile, Assumption 3.1 also postulates that Y has a non-null correlation with $(\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2$. Note that in Proposition 2.1, we assume that $\rho > 0$. When ρ is negative, we can still estimate $\boldsymbol{\beta}^*$ by applying (2.7) and (2.8) with $y^{(i)}$ replaced by $-y^{(i)}$ for all $i \in [n]$. As we show in the main result, the sign of ρ is well estimated by the sign of ρ_n, which is a byproduct of our thresholded spectral method for initialization described in Sect. 2.3. Thus, a non-zero ρ is sufficient for our TWF algorithm to recover the direction of $\boldsymbol{\beta}^*$ accurately.

To provide more insight on when $\rho \neq 0$ holds, we define the function $\varphi \colon \mathbb{R} \to \mathbb{R}$ by $\varphi(z) = \mathbb{E}_\epsilon[h(z, \epsilon)]$. Since $\boldsymbol{X}^\top \boldsymbol{\beta}^* \sim N(0, 1)$, by the Stein's identity [50, Lemma 4], we have

$$\rho = \mathrm{Cov}\big[Y, (\boldsymbol{X}^\top \boldsymbol{\beta}^*)^2\big] = \mathbb{E}\big[\varphi(Z) \cdot (Z^2 - 1)\big] = \mathbb{E}\big[D^2 \varphi(Z)\big],$$

where $Z \sim N(0,1)$, and $D^2\varphi\colon \mathbb{R} \to \mathbb{R}$ denotes the second-order distributional derivative[2] of φ. Thus, the non-zero assumption of ρ essentially imposes certain smoothness condition on φ, and hence indirectly on h.

To further understand the TWF algorithm, we first provide more intuition on the initialization step. Recall that, as shown in Proposition 2.2, for all $j \in [p]$, we have

$$\mathbb{E}\big[Y(\boldsymbol{X}_j^2 - 1)\big] = \rho \cdot \boldsymbol{\beta}_j^{*2}. \tag{3.1}$$

Using standard concentration inequalities, we have that

$$\frac{1}{n} \cdot \sum_{i=1}^{n} y^{(i)} \cdot \big[(\boldsymbol{x}_j^{(i)})^2 - 1\big] = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} (\boldsymbol{x}_j^{(i)})^2 - \mu_n$$

is close to its expected value $\rho \cdot \boldsymbol{\beta}_j^{*2}$ up to an additive error of order $\mathcal{O}(\sqrt{\log(np)/n})$, where μ_n is the sample average of $\{y^{(i)}\}_{i\in[n]}$. Thus, if for some $j \in [p]$, the magnitude of $\boldsymbol{\beta}_j^*$ is larger than $C \cdot [\log(np)/n]^{1/4}$ for some sufficiently large constant C, the signal term $\rho \cdot \boldsymbol{\beta}_j^{*2}$ dominates the additive error, which implies that j is in $\hat{S}_0$ defined in (2.10).

Thus, the thresholding step in (2.10) ensures that, with high probability, $\hat{S}_0$ contains the effective support of $\boldsymbol{\beta}^*$, which is defined as the set of coordinates of $\boldsymbol{\beta}^*$ whose magnitudes are at least $\Omega([\log(np)/n]^{1/4})$. Furthermore, we show that the coordinates outside the effective support of $\boldsymbol{\beta}^*$ (i.e., the entries of $\boldsymbol{\beta}^*$ with smaller magnitudes) can be safely ignored in the initialization step, since they are indistinguishable from the error. In other words, the thresholding step can be viewed as applying the diagonal thresholding method in Johnstone and Lu [27] to the empirical counterpart of $\mathbb{E}[Y \cdot (\boldsymbol{X}\boldsymbol{X}^\top - \mathbf{I}_p)]$.

More importantly, conducting the thresholding step in (2.10) before constructing $\boldsymbol{W}$ is indispensable. Since $|\hat{S}_0| \le s$, after the thresholding step, the noise introduced to the spectral estimator is roughly proportional to $|\hat{S}_0|$ linearly, which is significantly lower than the dimension p. This allows us to obtain a proper initial estimator with $\mathcal{O}[s^2 \log(np)]$ samples. On the other hand, since the dimensionality p is much larger than the sample size n, the spectral initial estimator without the thresholding step can incur a large error which makes the TWF algorithm diverge.

In what follows, we characterize the estimation error of the initial estimator $\boldsymbol{\beta}^{(0)}$ obtained by the thresholded spectral method. For ease of presentation, in the initialization step, we assume that $\hat{S}_0$, $\boldsymbol{W}$, and ρ_n are constructed using independent samples. Specifically, assuming we have $3n$ i.i.d. samples $\{(y^{(i)}, \boldsymbol{x}^{(i)})\}_{i\in[3n]}$, we use the first n samples to construct $\hat{S}_0$ in (2.10), and use the next n samples $\{(y^{(i)}, \boldsymbol{x}^{(i)})\}_{n+1\le i\le 2n}$ to construct $\boldsymbol{W}$ by (2.11). In addition, we let μ_n be the sample average of $y^{(n+1)}, \ldots, y^{(2n)}$; we denote $\boldsymbol{x}_{\hat{S}_0}^{(i)}$ by $\boldsymbol{w}^{(i)}$ for each $i \in \{n+1, \ldots, 2n\}$, and we let $\hat{\boldsymbol{v}}$ be the the leading eigenvector of $\boldsymbol{W}$ corresponding to the largest eigenvalue in magnitude. Finally, given $\hat{\boldsymbol{v}}$, we use the last n observations to construct ρ_n and the initial estimator $\boldsymbol{\beta}^{(0)}$ as in (2.12). Such construction frees us from possibly complicated dependent structures and allows us to present a simpler proof. In practice, the

[2] See, for example, Foucart and Rauhut [13] for the definition of the distributional derivative.

construction can be achieved by data splitting. Since we only raise the sample size by a constant factor of 3, the statistical and computational rates of our proposed method do not compromise. Note that the additional $2n$ independent samples can be avoided with more involved analysis using the Cauchy interlacing theorem [18]. Since our scope focuses on nonconvex optimization for misspecified models, we adopt such construction with independence to ease the presentation and analysis. Moreover, since we only raise the sample size by a constant factor of 3, the statistical and computational rates of our proposed method do not sacrifice. In the second step of the TWF algorithm, we still assume the sample size to be n so as to be consistent with the description of the algorithm.

In the following lemma, we provide an error bound for the initial estimator $\boldsymbol{\beta}^{(0)}$ obtained by the thresholded spectral method in (2.12). To measure the estimation error, we define

$$\operatorname{dist}(\mathbf{u}, \mathbf{v}) = \min\big(\|\mathbf{u} - \mathbf{v}\|, \|\mathbf{u} + \mathbf{v}\|\big)$$

for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$. As can be seen, this measure is invariant to the sign of $\boldsymbol{\beta}^*$, which is not identifiable. That is, for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, we have $\operatorname{dist}(\mathbf{u}, \mathbf{v}) = \operatorname{dist}(\mathbf{u}, -\mathbf{v})$. Note that we provide an outline of the proof here and defer the detailed arguments to §A.1.

Lemma 3.1 *Suppose Assumption 3.1 holds. Let $\boldsymbol{\beta}^{(0)}$ be the initial solution obtained by the thresholded spectral method defined in (2.12) with appropriately chosen constant γ specified in (2.10). Recall that the constant Ψ is specified in Assumption 3.1. For any constant $\alpha \in (0, 1]$, given large enough n satisfying $n \geq Cs^2 \log\big(np\big)$, where C is a generic constant depending only on Ψ and α, we have*

$$|\rho_n - \rho| \leq \alpha \cdot |\rho|, \quad \operatorname{supp}(\boldsymbol{\beta}^{(0)}) \subseteq \operatorname{supp}(\boldsymbol{\beta}^*), \quad \textit{and} \;\; \operatorname{dist}\big(\boldsymbol{\beta}^{(0)}, \overline{\boldsymbol{\beta}}\big) \leq \alpha \cdot \|\overline{\boldsymbol{\beta}}\|,$$

where $\overline{\boldsymbol{\beta}}$ is the minimizer of $\operatorname{Var}[Y - (\boldsymbol{X}^\top \boldsymbol{\beta})^2]$ *defined in* (2.1)*, with probability at least* $1 - \mathcal{O}(1/n)$.

Proof Sketch of Lemma 3.1 The proof contains two steps. In the first step, we show that $\hat{S}_0 \subseteq \operatorname{supp}(\boldsymbol{\beta}^*)$ with high probability. For each $j \in [p]$, we define

$$I_j = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} \cdot (\boldsymbol{x}_j^{(i)})^2.$$

Then $\hat{S}_0$ in (2.10) can be written as $\hat{S}_0 = \big\{j \in [p] \colon |I_j| > \gamma \sqrt{\log(np)/n}\big\}$. Note that, for any $j \in [p]$, we have $\mathbb{E}[Y \cdot (X_j^2 - 1)] = \rho \boldsymbol{\beta}_j^{*2}$. By the law of large numbers, we expect I_j to be close to zero for all $j \notin \operatorname{supp}(\boldsymbol{\beta}^*)$. Using concentration inequalities, we show that with probability at least $1 - \mathcal{O}\big(1/n\big)$, for any $j \in [p]$, we obtain

$$|I_j - \rho \cdot \boldsymbol{\beta}_j^{*2}| < \gamma \sqrt{\log(np)/n}.$$

Thus, for any $j \notin \operatorname{supp}(\boldsymbol{\beta}^*)$, we have $|I_j| < \gamma\sqrt{\log(np)/n}$ with probability at least $1 - \mathcal{O}(1/n)$, which implies that $j \notin \hat{S}_0$. Thus we conclude that $\hat{S}_0 \subseteq \operatorname{supp}(\boldsymbol{\beta}^*)$ with high probability.

In addition, by similar arguments, it is not difficult to see that, if $\boldsymbol{\beta}_j^{*2} \geq C_1\sqrt{\log(np)/n}$ with $C_1 \geq 2\gamma/|\rho|$, we have, by the triangle inequality,

$$|I_j| \geq |\rho \cdot \boldsymbol{\beta}_j^{*2}| - |I_j - \rho \cdot \boldsymbol{\beta}_j^{*2}| \geq \gamma\sqrt{\log(np)/n}$$

with high probability. This implies that $j \in \hat{S}_0$ high probability. Thus, $\hat{S}_0$ captures all entries $j \in \operatorname{supp}(\boldsymbol{\beta}^*)$ for which $\boldsymbol{\beta}_j^{*2}$ is sufficiently large. Then we obtain that

$$\left\|\boldsymbol{\beta}^*_{\hat{S}_0} - \boldsymbol{\beta}^*\right\|^2 = \mathcal{O}\Big[s\sqrt{\log(np)/n}\Big], \tag{3.2}$$

where $\boldsymbol{\beta}^*_{\hat{S}_0}$ is the restriction of $\boldsymbol{\beta}^*$ on $\hat{S}_0$.

In the second step, we show that the eigenvector $\hat{\boldsymbol{v}}$ of $\boldsymbol{W}$ is a good approximation of $\boldsymbol{\beta}^*_{\hat{S}_0}$. Recall that we denote $\boldsymbol{x}^{(i)}_{\hat{S}_0}$ by $\boldsymbol{w}^{(i)}$ to simplify the notation. In addition, let $\boldsymbol{\beta}_0^* = \boldsymbol{\beta}^*_{\hat{S}_0}/\|\boldsymbol{\beta}^*_{\hat{S}_0}\|$. We can rewrite $\boldsymbol{W}$ as

$$\boldsymbol{W} = \frac{1}{n}\sum_{i=n+1}^{2n}(y^{(i)} - \mu) \cdot |\boldsymbol{\beta}_0^{*\top}\boldsymbol{w}^{(i)}|^2 \cdot \boldsymbol{\beta}_0^*\boldsymbol{\beta}_0^{*\top} + \boldsymbol{N},$$

where $\boldsymbol{N}$ can be viewed as an error matrix. We show that $\|\boldsymbol{N}\|_2 = \mathcal{O}(s\sqrt{\log(np)/n})$ with probability at least $1 - \mathcal{O}(1/n)$. Thus, for $n \geq Cs^2\log(np)$ with a sufficiently large constant C, $\hat{\boldsymbol{v}}$ is close to $\boldsymbol{\beta}_0^*$ with high probability. Since $\|\boldsymbol{\beta}^*\| = 1$, (3.2) implies that $\boldsymbol{\beta}_0^*$ is close to $\boldsymbol{\beta}^*$. Thus, we show that $\hat{\boldsymbol{v}}$ and $\boldsymbol{\beta}^*$ are close in the sense that

$$|\hat{\boldsymbol{v}}^\top\boldsymbol{\beta}^*|^2 \geq 1 - C/|\rho| \cdot s\sqrt{\log(np)/n} \tag{3.3}$$

for some absolute constant C. Furthermore, since ρ_n in (2.12) is calculated using samples independent of $\hat{\boldsymbol{v}}$, by standard concentration inequalities, we have

$$\Big|\rho_n - \mathbb{E}\big\{Y \cdot \big[(\boldsymbol{X}^\top\hat{\boldsymbol{v}})^2 - 1\big]\big\}\Big| = \Big|\rho_n - \rho \cdot |\hat{\boldsymbol{v}}^\top\boldsymbol{\beta}^*|^2\Big| \leq \mathcal{O}\Big(\sqrt{\log n/n}\Big) \tag{3.4}$$

with probability at least $1 - \mathcal{O}(1/n)$. Hence, combining (3.3) and (3.4), we obtain that

$$|\rho_n - \rho| \leq \mathcal{O}\Big[s\sqrt{\log(np)/n}\Big] \leq \alpha|\rho| \tag{3.5}$$

for any fixed constant $\alpha \in (0, 1)$. Here the last inequality holds when $n \geq Cs^2\log(np)$ with constant C sufficiently large. Finally, to bound $\operatorname{dist}(\boldsymbol{\beta}^{(0)}, \overline{\boldsymbol{\beta}})$, by direct calculation, we have

$$\operatorname{dist}(\boldsymbol{\beta}^{(0)}, \overline{\boldsymbol{\beta}}) \leq \operatorname{dist}(\hat{\boldsymbol{v}}, \boldsymbol{\beta}^*) \cdot \sqrt{|\rho_n|/2} + \sqrt{|\rho_n - \rho|/2}. \tag{3.6}$$

Combining (3.3), (3.5), and (3.6), we obtain that

$$\operatorname{dist}(\boldsymbol{\beta}^{(0)}, \overline{\boldsymbol{\beta}}) \leq \mathcal{O}\big\{\big[s^2 \log(np)/n\big]^{1/4}\big\} \leq \alpha \|\overline{\boldsymbol{\beta}}\|,$$

which concludes the proof. □

Lemma 3.1 proves that the initial estimator is close to $\overline{\boldsymbol{\beta}}$ or $-\overline{\boldsymbol{\beta}}$ defined in (2.1) up to a constant order of error. We then proceed to characterize the computational and statistical properties of the thresholded Wirtinger flow algorithm in the next theorem.

Theorem 3.1 *Suppose Assumption 3.1 holds. Let the initial solution $\boldsymbol{\beta}^{(0)}$ be given by (2.12). Given $\alpha \leq 1/100$, $\kappa = \sqrt{80}$, $\eta \leq \eta_0/\rho$ for a constant η_0, and large enough n satisfying*

$$n \geq C_1\big[s^2 \log(np) + s \log^5 n\big],$$

for some constant C_1, there exists some constant C_2 such that for all $k = 0, 1, 2, \ldots, \operatorname{polylog}(pn)$, we have

$$\begin{aligned}\operatorname{supp}(\boldsymbol{\beta}^{(k+1)}) &\subseteq \operatorname{supp}(\boldsymbol{\beta}^*) \ \ \textit{and} \ \ \operatorname{dist}\big(\boldsymbol{\beta}^{(k+1)}, \overline{\boldsymbol{\beta}}\big) \\ &\leq \big(1-\eta\rho\big)^k \cdot \operatorname{dist}\big(\boldsymbol{\beta}^{(0)}, \overline{\boldsymbol{\beta}}\big) + C_2\sqrt{s \log(np)/n},\end{aligned}$$

with probability at least $1 - \mathcal{O}\big(1/n\big)$.

This theorem implies that given an appropriate initialization, the TWF algorithm maintains the solution sparsity throughout iterations, and attains a linear rate of convergence to $\overline{\boldsymbol{\beta}}$ up to some unavoidable finite-sample statistical error. Therefore, when the number of iterations T satisfies

$$T \geq \left\lceil \frac{1}{\eta\rho} \cdot \log\left[\frac{\operatorname{dist}(\boldsymbol{\beta}^{(0)}, \overline{\boldsymbol{\beta}})}{\sqrt{s \log p/n}}\right]\right\rceil = \mathcal{O}[\log(n/s)] \tag{3.7}$$

and $T \leq \operatorname{poly}\log(np)$, with probability at least $1 - \mathcal{O}(1/n)$, there exists a constant $C_2' > 0$ such that

$$\operatorname{dist}\big(\boldsymbol{\beta}^{(T)}, \overline{\boldsymbol{\beta}}\big) \leq C_2'\sqrt{s \log p/n}.$$

Here we denote by $\lceil x \rceil$ the smallest integer no less than x.

Thus, Theorem 3.1 establishes a unified computational and statistical guarantees for our TWF algorithm. Specifically, this theorem implies that the estimation error of each $\boldsymbol{\beta}^{(t)}$ is the sum of the statistical error and the computation error. In specific, the statistical error is of order $\sqrt{s \log p/n}$, which cannot be further improved; the optimization error converges to zero in a linear rate as the algorithm proceeds. We point out that our analysis matches the optimal statistical rate of convergence for sparse phase retrieval as discussed in Cai et al. [6]. However, the theoretical analysis in Cai et al. [6] only works for the correctly specified phase retrieval model, which is a special case of our model with link function $h(u, v) = u^2 + v$.

Moreover, note that the estimation error $\mathrm{dist}(\boldsymbol{\beta}^{(T)}, \overline{\boldsymbol{\beta}})$ is of order $\sqrt{s \log p/n}$ when T is sufficiently large. While we require the sample complexity to be $n = \Omega(s^2 \log p)$. Similar phenomena have been observed by Cai et al. [6], which is also conjectured to be the computational barrier of sparse phase retrieval models.

Proof Sketch of Theorem 3.1 Due to space limit, we provide an outline of the proof for the main theorem, and the detailed proof is provided in §A.2.

The proof of this theorem consists of two major steps. In the first step, let $S = \mathrm{supp}(\boldsymbol{\beta}^*)$ and let $z \in \mathbb{R}^d$ be a point in the neighborhood of $\overline{\boldsymbol{\beta}}$ (or $-\overline{\boldsymbol{\beta}}$) such that $\mathrm{supp}(z) \subseteq S$. We show that, starting from z, the TWF update step restricted on S, moves toward the true solution $\overline{\boldsymbol{\beta}}$ (or $-\overline{\boldsymbol{\beta}}$) with high probability. In the second step, we show that the thresholding map forces the TWF updates to be supported on S with high probability. The formal statement of the first step is presented in the following proposition. □

Proposition 3.1 *Let κ and η be appropriately chosen positive constants as in (2.7) and (2.8). We denote $S = \mathrm{supp}(\boldsymbol{\beta}^*)$, and let $\mathcal{S} \subseteq \mathbb{R}^p$ be the set of all vectors in $\mathbb{R}^p$ that are supported on S. For any $z \in \mathbb{R}^p$, we define a mapping $t\colon \mathbb{R}^p \to \mathbb{R}^p$ as*

$$t(z) = \mathcal{T}_{\eta\tau(z)}\big\{z - \eta[\nabla \ell_n(z)]_S\big\},$$

where $\mathcal{T}$ is the hard-thresholding operator defined in (2.8), and the threshold value $\tau(\cdot)$ is defined in (2.7).

For all $z \in \mathcal{S}$ satisfying $\mathrm{dist}(z, \overline{\boldsymbol{\beta}}) \le \alpha \|\overline{\boldsymbol{\beta}}\|$ with $\alpha \in (0, 1/100)$, given large enough n satisfying $n \ge C[s^2 \log(np) + s \log^5 n]$, we have

$$\mathrm{dist}\big[t(z), \overline{\boldsymbol{\beta}}\big] \le \big(1 - \eta\rho\big) \cdot \mathrm{dist}(z, \overline{\boldsymbol{\beta}}) + C'\sqrt{s \log(np)/n}$$

for some constant C', with probability at least $1 - 1/(5n)$.

Proof Sketch of Proposition 3.1 Here we provide the outline of the proof, and the full proof of this proposition is presented in §A.3.

Without loss of generality, we assume $\mathrm{dist}(z, \overline{\boldsymbol{\beta}}) = \|z - \overline{\boldsymbol{\beta}}\|$, where the proof for the other case, $\mathrm{dist}(z, \overline{\boldsymbol{\beta}}) = \|z + \overline{\boldsymbol{\beta}}\|$, follows similarly. By the definition of the hard-thresholding operator $\mathcal{T}$, we first rewrite $t(z)$ as

$$t(z) = \mathcal{T}_{\eta\tau(z)}\big\{z - \eta[\nabla \ell_n(z)]_S\big\} = z - \eta[\nabla \ell_n(z)]_S + \eta\tau(z)\boldsymbol{v},$$

where the last equality holds for some $\boldsymbol{v}$ satisfying $\boldsymbol{v} \in \mathcal{S}$ and $\|\boldsymbol{v}\|_\infty \le 1$.

Let $\boldsymbol{h} = z - \overline{\boldsymbol{\beta}}$. We have $\boldsymbol{h} \in \mathcal{S}$ by definition. We then bound $\|t(z) - \overline{\boldsymbol{\beta}}\|$ using $\|\boldsymbol{h}\|$. By the triangle inequality, we immediately have

$$\|t(z) - \overline{\boldsymbol{\beta}}\| \le \underbrace{\big\|\boldsymbol{h} - \eta\nabla_S \ell_n(z)\big\|}_{R_1} + \underbrace{\eta\tau(z) \cdot \|\boldsymbol{v}\|}_{R_2}.$$

In the following, we establish upper bounds for R_1 and R_2, respectively. We start with bounding R_2, which is essentially the error introduced by the hard-thresholding operator. Recall that by the definition of $\tau(\cdot)$ in (2.7), we have

$$\tau(z)^2 = \kappa^2 \cdot \frac{\log(np)}{n^2} \cdot \sum_{i=1}^{n} \big[y^{(i)} - (\boldsymbol{x}^{(i)\top} z)^2 - \mu_n + \|z\|^2\big]^2 \big(\boldsymbol{x}^{(i)\top} z\big)^2.$$

Since $z = \boldsymbol{h} + \overline{\boldsymbol{\beta}}$, we can rewrite $\tau(z)^2$ as a sum of terms of the form

$$\sum_{i=1}^{n} (y^{(i)})^a \big(\boldsymbol{x}^{(i)\top} \overline{\boldsymbol{\beta}}\big)^b \big(\boldsymbol{x}^{(i)\top} \boldsymbol{h}\big)^c,$$

where a, b, and c are nonnegative integers. By carefully expanding these terms, we observe that the dominating term in $\tau(z)^2$ is

$$\kappa^2 \frac{\log(np)}{n^2} \sum_{i=1}^{n} \big(\boldsymbol{x}^{(i)\top} \overline{\boldsymbol{\beta}}\big)^2 \big(\boldsymbol{x}^{(i)\top} \boldsymbol{h}\big)^4.$$

Here, we derive an upper bound for this term as an example; see §A.3 for the details of bounding $\tau(z)^2$. By Hölder's inequality, we have

$$\begin{aligned} \sum_{i=1}^{n} \big(\boldsymbol{x}^{(i)\top} \overline{\boldsymbol{\beta}}\big)^2 \big(\boldsymbol{x}^{(i)\top} \boldsymbol{h}\big)^4 &\le \left(\sum_{i=1}^{n} \big|\boldsymbol{x}^{(i)\top} \overline{\boldsymbol{\beta}}\big|^6\right)^{2/6} \left(\sum_{i=1}^{n} \big|\boldsymbol{x}^{(i)\top} \boldsymbol{h}\big|^6\right)^{4/6} \\ &\le \|\boldsymbol{A}\|_{2\to 6}^6 \|\overline{\boldsymbol{\beta}}\|^2 \cdot \|\boldsymbol{h}\|^4, \end{aligned}$$

where $\|\boldsymbol{A}\|_{2\to 6}$ is the induced norm of $\boldsymbol{A}$. We show in §C that, when $n \ge C[s^2 \log(np) + s \log^5 n]$, we have $\|\boldsymbol{A}\|_{2\to 6}^6 = \mathcal{O}(n + s^3)$ with probability at least $1 - \mathcal{O}(1/n)$. Thus we have

$$\kappa^2 \frac{\log(np)}{n^2} \sum_{i=1}^{n} \big(\boldsymbol{x}^{(i)\top} \overline{\boldsymbol{\beta}}\big)^2 \big(\boldsymbol{x}^{(i)\top} \boldsymbol{h}\big)^4 \le C' \kappa^2 \frac{(n + s^3) \log(np)}{n^2} \|\overline{\boldsymbol{\beta}}\|^2 \cdot \|\boldsymbol{h}\|^4$$

for some constant C'. Using similar methods to bound other terms, we eventually obtain that

$$\tau(z)^2 \le C_\tau^2 \cdot (n + s^3) \cdot \log(np)/(n^2) \cdot \|\overline{\boldsymbol{\beta}}\|^2 \cdot \|\boldsymbol{h}\|^4 + \mathcal{O}\big[\log(np)/n\big],$$

where $C_\tau > 0$ is some constant. Since $\boldsymbol{v}$ is supported on $\mathrm{supp}(\boldsymbol{\beta}^*)$ with $\|\boldsymbol{v}\|_\infty \le 1$, we have $\|\boldsymbol{v}\| \le \sqrt{s}$. Hence, we obtain that

$$\begin{aligned} \tau(z)\|\boldsymbol{v}\| &= \mathcal{O}\Big[\kappa \sqrt{\big(ns + s^4\big) \log(np)/n^2} \cdot \|\overline{\boldsymbol{\beta}}\| \cdot \|\boldsymbol{h}\|^2\Big] + \mathcal{O}\Big[\sqrt{s \log(np)/n}\Big] \\ &= \mathcal{O}\big(\kappa \|\overline{\boldsymbol{\beta}}\| \|\boldsymbol{h}\|^2\big) + \mathcal{O}\Big[\sqrt{s \log(np)/n}\Big] = \mathcal{O}(\alpha\kappa\rho\|\boldsymbol{h}\|) + \mathcal{O}\Big[\sqrt{s \log(np)/n}\Big]. \end{aligned}$$

Here in the second equality, we use $n \ge C[s^2 \log(np) + s \log^5 n]$, and the last equality follows from $\|\boldsymbol{h}\| \le \alpha \|\overline{\boldsymbol{\beta}}\|$ and $\|\overline{\boldsymbol{\beta}}\|^2 = \rho/2$ as shown in Proposition 2.1. Since α can be a arbitrarily small constant, we finally have

$$R_2 \le 3/4 \cdot \eta\rho\|\boldsymbol{h}\| + \mathcal{O}\Big[\sqrt{s\log(np)/n}\Big].$$

In addition, deriving the bound for R_1 follows from similar arguments. In the proof, we heavily rely on the fact that for any $\boldsymbol{g} \in \mathbb{R}$, $[\boldsymbol{g} - (\boldsymbol{g}^\top\boldsymbol{\beta}^*)\boldsymbol{\beta}^*]^\top\boldsymbol{X}$ is independent with Y. This is because $[\boldsymbol{g} - (\boldsymbol{g}^\top\boldsymbol{\beta}^*)\boldsymbol{\beta}^*]^\top\boldsymbol{\beta}^* = 0$, and Y depends on $\boldsymbol{X}$ only through $\boldsymbol{X}^\top\boldsymbol{\beta}^*$. For the details of the proof, we refer the readers to §A.3. In particular, we show that, with high probability,

$$R_1 = (1 - 7/4\eta\rho) \cdot \|\boldsymbol{h}\| + \mathcal{O}\Big[\sqrt{s\log(np)/n}\Big].$$

Combining the bounds for R_1 and R_2, we conclude that for all $\boldsymbol{z} \in \mathcal{S}$,

$$\text{dist}\big[t(z), \overline{\boldsymbol{\beta}}\big] = \big\|t(\boldsymbol{z}) - \overline{\boldsymbol{\beta}}\big\| = (1 - \eta\rho) \cdot \|\boldsymbol{h}\| + \mathcal{O}\Big[\sqrt{s\log(np)/n}\Big],$$

for sufficiently small α and η, which completes the proof of Proposition 3.1.

To complete the proof of Theorem 3.1, in the the second step, we show that with high probability, the hard-thresholding operator keeps

the thresholded gradient updates of the TWF algorithm supported on supp$(\boldsymbol{\beta}^*)$. We establish this result by considering another sequence $\{\tilde{\boldsymbol{\beta}}^{(k)}\}_{k\ge 0}$ which is specified by $\tilde{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}^{(0)}$ and

$$\tilde{\boldsymbol{\beta}}^{(k+1)} = \mathcal{T}_{\eta\tau(\tilde{\boldsymbol{\beta}}^{(k)})}\Big\{\tilde{\boldsymbol{\beta}}^{(k)} - \eta[\nabla\ell_n(\tilde{\boldsymbol{\beta}}^{(k)})]_S\Big\} \tag{3.8}$$

for each $k \ge 0$, where we let $S = \text{supp}(\boldsymbol{\beta}^*)$. That is, these two sequences $\{\tilde{\boldsymbol{\beta}}^{(k)}\}_{k\ge 0}$ and $\{\boldsymbol{\beta}^{(k)}\}_{k\ge 0}$ have the same starting point $\boldsymbol{\beta}^{(0)}$, and the former is constructed by restricting the TWF updates to supp$(\boldsymbol{\beta}^*)$.

In the sequel, we prove by induction that these two sequences coincide with high probability. Specifically, we show that

$$\boldsymbol{\beta}^{(k)} = \tilde{\boldsymbol{\beta}}^{(k)}, \quad \text{supp}(\tilde{\boldsymbol{\beta}}^{(k)}) \subseteq \text{supp}(\boldsymbol{\beta}^*), \quad \text{and } \text{dist}\big(\tilde{\boldsymbol{\beta}}^{(k)}, \overline{\boldsymbol{\beta}}\big) \le \alpha\|\overline{\boldsymbol{\beta}}\| \tag{3.9}$$

for all $k \ge 0$ with high probability. As shown in Lemma 3.1, (3.9) holds for $\boldsymbol{\beta}^{(0)}$ with probability at least $1 - \mathcal{O}(1/n)$.

Assuming (3.9) holds for some k, we consider $\tilde{\boldsymbol{\beta}}^{(k+1)}$. We first note that, by the construction of $\tilde{\boldsymbol{\beta}}^{(k+1)}$ in (3.8), we immediately have supp$(\tilde{\boldsymbol{\beta}}^{(k+1)}) \subseteq S$ provided supp$(\tilde{\boldsymbol{\beta}}^{(k)}) \subseteq S$. Moreover, by Proposition 3.1, when n is sufficiently large, we have

$$\begin{aligned}\text{dist}\big(\tilde{\boldsymbol{\beta}}^{(k+1)}, \overline{\boldsymbol{\beta}}\big) &\le (1 - \eta\rho) \cdot \text{dist}\big(\tilde{\boldsymbol{\beta}}^{(k)}, \overline{\boldsymbol{\beta}}\big) + C'\sqrt{s\log(np)/n} \\ &\le (1 - \eta\rho) \cdot \text{dist}\big(\tilde{\boldsymbol{\beta}}^{(k)}, \overline{\boldsymbol{\beta}}\big) + \eta\rho\alpha\big\|\overline{\boldsymbol{\beta}}\big\| \le \alpha\|\overline{\boldsymbol{\beta}}\|.\end{aligned}$$

Thus, it only remains to show that $\tilde{\boldsymbol{\beta}}^{(k+1)} = \boldsymbol{\beta}^{(k+1)}$ under the inductive assumption $\tilde{\boldsymbol{\beta}}^{(k)} = \boldsymbol{\beta}^{(k)}$. It suffices to show that the hard-thresholding operator sets any $j \notin S$ to zero. That is, we show that

$$\max_{j \notin S} \big| [\nabla \ell_n(\tilde{\boldsymbol{\beta}}^{(k)})]_j \big| \le \tau\big(\tilde{\boldsymbol{\beta}}^{(k)}\big)$$

holds with high probability. Since $\tilde{\boldsymbol{\beta}}^{(k)}$ is supported on S, for any $j \notin S$, $\boldsymbol{x}_j^{(i)}$ is independent of

$$\varphi_i = \big[y^{(i)} - \big(\boldsymbol{x}^{(i)\top} \tilde{\boldsymbol{\beta}}^{(k)}\big)^2 - \xi_n(\tilde{\boldsymbol{\beta}}^{(k)})\big] \cdot \big(\boldsymbol{x}^{(i)\top} \tilde{\boldsymbol{\beta}}^{(k)}\big).$$

Moreover, by the definition of $\nabla \ell_n(\cdot)$ in (2.9), we have $[\nabla \ell_n(\tilde{\boldsymbol{\beta}}^{(k)})]_j = -4n^{-1} \sum_{i=1}^n \varphi_i \cdot \boldsymbol{x}_j^{(i)}$. Therefore, when we condition on $\{\varphi_i\}_{i \in [n]}$, $[\nabla \ell_n(\tilde{\boldsymbol{\beta}}^{(k)})]_j$ is a centered Gaussian random variable with variance $16/(n^2) \cdot \sum_{i=1}^n \varphi_i^2$. Thus, with probability at least $1 - 1/(n^2 p)$, for all $j \notin S$, we have

$$\big| [\nabla \ell_n(\tilde{\boldsymbol{\beta}}^{(k)})]_j \big| \le \left[\frac{80 \log(np)}{n^2} \sum_{i=1}^n \varphi_j^2 \right]^{1/2} \le \tau(\tilde{\boldsymbol{\beta}}^k).$$

Therefore, we conclude that, with probability at least $1 - 1/(n^2 p)$, we obtain

$$\tilde{\boldsymbol{\beta}}^{(k+1)} = \mathcal{T}_{\eta\tau(\tilde{\boldsymbol{\beta}}^k)} \Big\{ \tilde{\boldsymbol{\beta}}^{(k)} - \eta [\nabla \ell_n(\tilde{\boldsymbol{\beta}}^{(k)})]_S \Big\} = \mathcal{T}_{\eta\tau(\tilde{\boldsymbol{\beta}}^k)} \big[\tilde{\boldsymbol{\beta}}^{(k)} - \eta \nabla \ell_n(\tilde{\boldsymbol{\beta}}^{(k)}) \big] = \boldsymbol{\beta}^{(k+1)},$$

which completes the induction step. By summing up the probabilities that the aforementioned events do not happen, we conclude that the theorem statement holds for all $k = 1, 2, \ldots \text{poly} \log(np)$. □

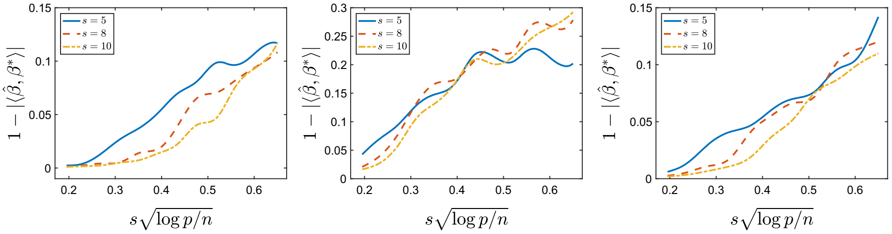
4 Experimental results

In this section, we evaluate the empirical finite-sample performance of the proposed variant of TWF algorithm for misspecified phase retrieval models. We present results on both simulated and real data. We present additional simulation results in "Appendix A".

4.1 Simulated data

For simulated data, we consider three link functions for the misspecified phase retrieval model including

$$h_1(u, v) = |u| + v, \quad h_2(u, v) = |u + v|, \quad h_3(u, v) = 4u^2 + 3\sin(|u|) + v. \quad (4.1)$$



(a) $h_1(u,v)=|u|+v$ **(b)** $h_2(u,v)=|u+v|$, **(c)** $h_3(u,v)=4u^2+3\sin(|u|)+v$

Fig. 1 Plots of the cosine distance $1-|\langle\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*\rangle|$ against the inverse signal-to-noise ratio $s\sqrt{\log p/n}$, in which $\hat{\boldsymbol{\beta}}$ is obtained by normalizing the final output $\boldsymbol{\beta}^{(T)}$. The link function is one of h_1, h_2, and h_3 in (4.1). Besides, we set $p=1000$, $s\in\{5,8,10\}$, and let n vary. We generate each of the figures based on 100 independent trials for each (n,s,p)

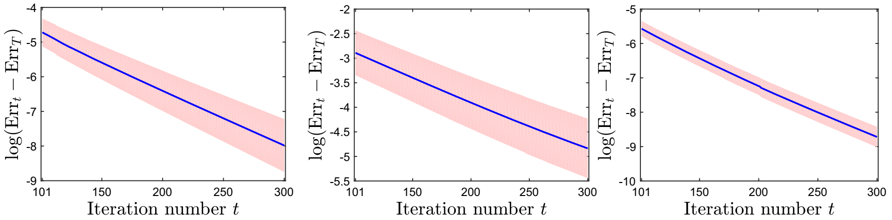
One can verify that the above link functions satisfy Assumption 3.1. Hence, the model in (1.3) with $h\in\{h_1,h_2,h_3\}$ is a misspecified phase retrieval model. Throughout the experiments, we fix $p=1000$, $s\in\{5,8,10\}$, and let n vary. We sample the random noise ϵ from the standard Gaussian distribution, and we let each covariate be sampled independently from $\boldsymbol{X}\sim N(0,\mathbf{I}_p)$. For the signal parameter $\boldsymbol{\beta}^*$, we choose $\text{supp}(\boldsymbol{\beta}^*)$ uniformly at random among all subsets of $[p]$ with cardinality s. Moreover, the nonzero entries of $\boldsymbol{\beta}^*$ is sampled uniformly from the unit sphere in $\mathbb{R}^s$. For tuning parameters of the initialization, we set $\gamma=2$ in (2.10). We observe that our numerical results are not sensitive to the choice of γ. For the other tuning parameters in the TWF algorithm, we set κ in (2.7) and the step size η as 15 and 0.005, respectively. Finally, we terminate the algorithm when $\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t-1)}\|\le 10^{-4}$.

We first investigate the empirical statistical rate of convergence, which is of the order $\sqrt{s\log p/n}$ as quantified in Theorem 3.1. In particular, we compare the estimation error of $\boldsymbol{\beta}^{(T)}$ with $\sqrt{s\log p/n}$, where $\boldsymbol{\beta}^{(T)}$ is the final estimator obtained by the TWF algorithm. Since $\boldsymbol{\beta}^*$ has unit norm and its sign is not identifiable, we use the cosine distance $1-|\langle\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*\rangle|$ as a measure of the estimation error, where $\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}^{(T)}/\|\boldsymbol{\beta}^{(T)}\|$. This is equivalent to using $\text{dist}(\cdot,\cdot)$, since

$$1-|\langle\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*\rangle| = 1/2\cdot\text{dist}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*)^2.$$

Recall that Theorem 3.1 requires the sample size n to be larger than $C\cdot s^2\log p$ for some constant C. The term $s\sqrt{\log p/n}$ essentially quantifies the difficulty of recovering $\boldsymbol{\beta}^*$, which can be viewed as the inverse signal-to-noise ratio (SNR) in our problem. In Fig. 1, for each $s\in\{5,8,10\}$, we plot the averaged estimation error $1-|\langle\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*\rangle|$ using 100 independent trials against the inverse SNR $s\sqrt{\log p/n}$. Since we set s as small constants, as shown in the figures, the estimation error is bounded by a linear function of $\sqrt{s\log p/n}$, corroborating the statistical rate of convergence as discussed in Theorem 3.1.

Next, we study the optimization error of the algorithm empirically. As shown in Theorem 3.1, the optimization error converges to zero at a linear rate. We first define $\text{Err}_t=1-|\langle\boldsymbol{\beta}^{(t)},\boldsymbol{\beta}^*\rangle|/\|\boldsymbol{\beta}^{(t)}\|$ as the cosine error of $\boldsymbol{\beta}^{(t)}$ for all $t\ge 0$. Note that



(a) $h_1(u,v) = |u| + v$ **(b)** $h_2(u,v) = |u+v|$, **(c)** $h_3(u,v) = 4u^2 + 3\sin(|u|) + v$

Fig. 2 Plots of the log optimization error $\log(\mathrm{Err}_t - \mathrm{Err}_T)$ against the number of iterations t for $t \in \{101, 102, \ldots, 300\}$, in which Err_t is the cosine error of the tth iterate of the TWF algorithm and $T = 1000$ is the total number of iterations. Here the link function is one of h_1, h_2, and h_3. In addition, we set $p = 1000$, $s = 5$, and $n = 864 \approx 5s^2 \cdot \log p$. These figures are generated based on 50 independent trials for each link function

Err_t converges to a non-vanishing statistical error term. When T is sufficiently large such that the algorithm converges, $\mathrm{Err}_t - \mathrm{Err}_T$ can be viewed as the optimization error of $\boldsymbol{\beta}^{(t)}$. Hence, viewing $\log(\mathrm{Err}_t - \mathrm{Err}_T)$ as a function of the iteration counter t, $\log(\mathrm{Err}_t - \mathrm{Err}_T)$ is close to a linear function of t with a negative slope by Theorem 3.1. To verify this result, in the following experiments, we set $p = 1000$, $s = 5$, and $n = 863$, which satisfy $n = 5s^2 \cdot \log p$ approximately. Similar to the previous experiments, we study the convergence using three different link functions in (4.1). For each link function, we run the TWF algorithm for $T = 1000$ iterations with the tuning parameters set as $\gamma = 2$, $\kappa = 15$, and $\eta = 0.005$. In Fig. 2, we plot the averaged $\log(\mathrm{Err}_t - \mathrm{Err}_T)$ against t based on 50 independent trials with $t \in \{101, \ldots, 300\}$. For the 50 sequences of $\{\log(\mathrm{Err}_t - \mathrm{Err}_T)\}_{t \in [T]}$, we also compute the mean and standard error. In Fig. 2, the blue curves are the mean values of these sequences, and the red areas are empirical confidence bands with one standard error. As shown in these figures, at the second stage of the algorithm, the optimization error decreases to zero at a linear rate, which corroborates our theory.

4.2 Real Data

We consider an example in image processing using real data. Let $\mathbf{M} \in \mathbb{R}^{H \times W}$ be an image, where H and W denote the height and width of $\mathbf{M}$, correspondingly. Without loss of generality, we assume that $H \leq W$. In addition, we focus on the case that $\mathbf{M}$ has a sparse representation $\sum_{j=1}^{p} \boldsymbol{\alpha}_j \cdot \mathbf{D}_j$, where the coefficient vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)^\top \in \mathbb{R}^p$ is sparse, and the d matrices $\{\mathbf{D}_j\}_{j \in [p]}$ are orthonormal in $\mathbb{R}^{H \times W}$. That is, $\langle \mathbf{D}_j, \mathbf{D}_k \rangle = \mathrm{Tr}(\mathbf{D}_k^\top \mathbf{D}_j) = \mathbb{1}(j = k)$ for all $j, k \in [p]$, where $\mathrm{Tr}(\cdot)$ stands for the trace of a matrix. Given a random matrix $\boldsymbol{Z} \in \mathbb{R}^{H \times W}$ with i.i.d. $N(0,1)$ entries, we consider a model $Y = h\big(\langle \mathbf{Z}, \mathbf{M}/\|\mathbf{M}\|_{\mathrm{fro}} \rangle, \epsilon\big)$. Furthermore, we denote $\boldsymbol{\beta}^* = \boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|$ and define $\boldsymbol{X} \in \mathbb{R}^p$ by letting $\boldsymbol{X}_j = \langle \mathbf{D}_j, \boldsymbol{Z} \rangle$ for each $j \in [p]$. Since $\{\mathbf{D}_j\}_{j \in [p]}$ are orthonormal, we have $\boldsymbol{X} \sim N(0, \mathbf{I}_p)$ and $\|\boldsymbol{\alpha}\| = \|\mathbf{M}\|_{\mathrm{fro}}$. Thus, the model defined above is equivalent to the misspecified phase retrieval model defined in (1.3). In the sequel, reconstruct the signal parameter $\boldsymbol{\beta}^*$ using the TWF algorithm. Let $\widehat{\boldsymbol{\beta}}$ be the estimator, we then use the quality of the reconstructed image $\widehat{\mathbf{M}} = \|\boldsymbol{\alpha}\| \cdot \sum_{j=1}^{p} \widehat{\boldsymbol{\beta}}_j \cdot \mathbf{D}_j$ to evaluate the empirical performance of our algorithm.

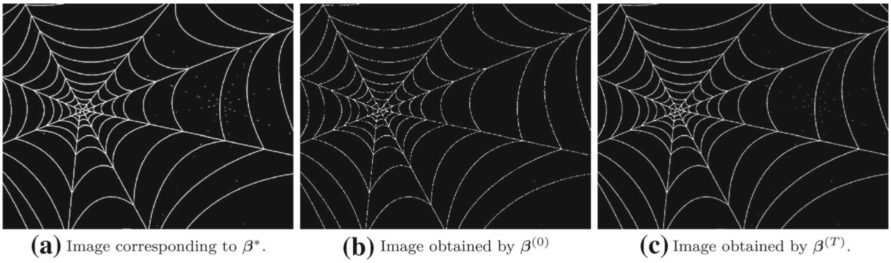

(a) Image corresponding to β^*.
(b) Image obtained by $\beta^{(0)}$
(c) Image obtained by $\beta^{(T)}$.

Fig. 3 Plots of the sparse image used for misspecified phase retrieval and the images reconstructed using the TWF algorithm. The sparse image is plotted in **a**, which is decomposed into $L = 150$ disjoint patches to construct the signal parameters $\{\boldsymbol{\beta}^*_\ell\}_{\ell\in[L]}$. We plot the image obtained using the initialization $\{\boldsymbol{\beta}^{(0)}_\ell\}_{\ell\in[L]}$ in **b**, which, although approximately recovers the pattern of the image in **a**, incurs observable reconstruction error . In **c** we plot the reconstructed image using $\boldsymbol{\beta}^{(T)}$ with $T = 1000$, whose difference from the image in **a** is minute

We first consider the sparse image plotted in Fig. 3a. The size of this image is $H = 385$ by $W = 500$. In this case, we flatten the image matrix and regard $\mathbf{M}$ as a sparse vector in $\mathbb{R}^{H\times M}$. Due to the size of the image, we decompose the image into $L = 150$ disjoint patches, each with size $p = 1290$. That is, we have $\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_L)^\top$, where $\mathbf{M}_\ell \in \mathbb{R}^p$ for all $\ell \in [L]$. Denote the sparsity of $\mathbf{M}_\ell$ by s_ℓ. The maximum and minimum values of $\{s_\ell\}_{\ell\in[L]}$ are 205 and 44, respectively. For each $\ell \in [L]$, we use $\boldsymbol{\beta}^*_\ell = \mathbf{M}_\ell/\|\mathbf{M}_\ell\|_{\text{fro}} \in \mathbb{R}^p$ as the signal parameter of the model in (1.3) with $\epsilon \sim N(0, 1)$. We set h_1 defined in (4.1) as the link function and sample $n_\ell = 5 \cdot s^2_\ell \cdot \log p$ i.i.d. observations. To recover each $\boldsymbol{\beta}^*_\ell$, we run the TWF algorithm and output the final estimator after $T = 1000$ gradient steps. For the tuning parameters of the TWF algorithm, similar to the case in simulated data, we set $\gamma = 2$, $\tau = 15$, and $\eta = 0.005$. To present the result, we plot in Fig. 3c the image constructed using $\{\boldsymbol{\beta}^{(T)}_\ell\}_{\ell\in[L]}$, which is very close to the target image in Fig. 3a. Moreover, we also plot in Fig. 3b the image constructed using $\{\boldsymbol{\beta}^{(0)}_\ell\}_{\ell\in[L]}$ for comparison, which are obtained by spectral initialization. It is perceptible that Fig. 3b already approximately recovers the pattern in the target image, which justifies our theory that spectral initialization has a constant order of error. However, the difference between Fig. 3b and a is also discernible. Comparing Fig. 3b and c, it is clear that the gradient steps in the second stage of the TWF algorithm improves the estimation accuracy.

In addition, we also experiment on a dense image with sparse representations. Let $\mathbf{M} = \sum_{j=1}^{H} \boldsymbol{\alpha}_j \mathbf{u}_j \mathbf{v}_j^\top$ be the singular value decomposition of $\mathbf{M}$, where $\{\boldsymbol{\alpha}_j\}_{j\in[H]}$ are the singular values of $\mathbf{M}$ in the descending order, and $\mathbf{u}_j$ and $\mathbf{v}_j$ are the corresponding singular vectors. The best rank-s approximation of $\mathbf{M}$ is given by $\tilde{\mathbf{M}} = \sum_{j=1}^{s} \boldsymbol{\alpha}_j \mathbf{u}_j \mathbf{v}_j^\top$. The success of image compression implies that real images can be well approximated by some low-rank matrix, that is, $\tilde{\mathbf{M}}$ is close to $\mathbf{M}$ for some $s \ll H$. Let $\mathbf{D}_j = \mathbf{u}_j \mathbf{v}_j^\top$ for each $j \in [H]$. Then $\tilde{\mathbf{M}}$ can be represented by an s-sparse vector of singular values $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_s, 0, \ldots, 0)^\top \in \mathbb{R}^H$ using basis $\{\mathbf{D}_j\}_{j\in[H]}$. In this example, we treat $\boldsymbol{\beta}^* = \boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|$ as the signal parameter in the misspecified phase retrieval model, and and reconstruct $\boldsymbol{\beta}^*$ using the TWF algorithm.

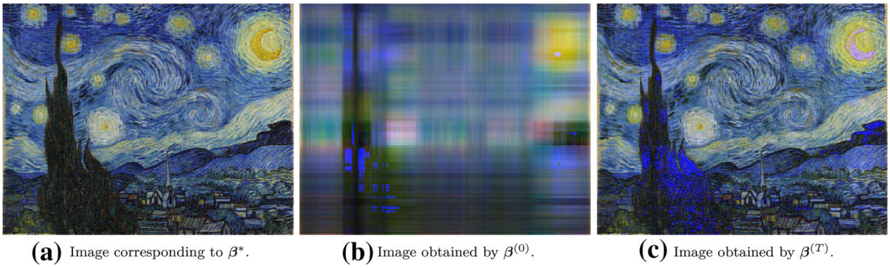

(a) Image corresponding to β^*. (b) Image obtained by $\beta^{(0)}$. (c) Image obtained by $\beta^{(T)}$.

Fig. 4 Plots of the image used for misspecified phase retrieval and the images reconstructed using the TWF algorithm. The rank-80 approximation of the original *Starry Night* is plotted in **a**, which corresponds to the signal parameter $\boldsymbol{\beta}^*$. We plot the image obtained using the initialization $\boldsymbol{\beta}^{(0)}$ in **b**, which is a poor reconstruction of the image in **a**. In **c** we plot the reconstructed image using $\boldsymbol{\beta}^{(T)}$ with $T = 1000$, whose difference from the image in **a** is negligible

In Fig. 4a, we plot the low-rank approximation of *Starry Night* by Vincent van Gogh. The height and width of the image is $H = 1014$ and $W = 1280$. For each channel of the image, we compute the best rank-s approximation with $s = 80$. As for the experiment, we set the link function as h_2 in (4.1). We sample $n = 10s^2 \log(p)$ observations of the model in (1.3) with $\epsilon \sim N(0, 1)$, where $s = 80$, $p = H = 1014$, and the covariate $\boldsymbol{X} \sim N(0, \mathbf{I}_p)$. Similar to the previous examples, we let the tuning parameters of the TWF algorithm be $\gamma = 2$, $\tau = 15$, and $\eta = 0.005$. Here we output the final estimator $\boldsymbol{\beta}^{(T)}$ and the corresponding image after $T = 1000$ gradient steps. For comparison, in Fig. 4b we plot the image constructed using $\boldsymbol{\beta}^{(0)}$, which is obtained by spectral initialization. It is evident that this image is blurred and does not capture the pattern of the target image in Fig. 4a. In sharp contrast, as shown in Fig. 4c, the image constructed using $\boldsymbol{\beta}^{(T)}$ is very close to the target image. Furthermore, the difference is nearly indiscernible, which illustrates the empirical success of the TWF algorithm for real applications.

5 Concluding remarks

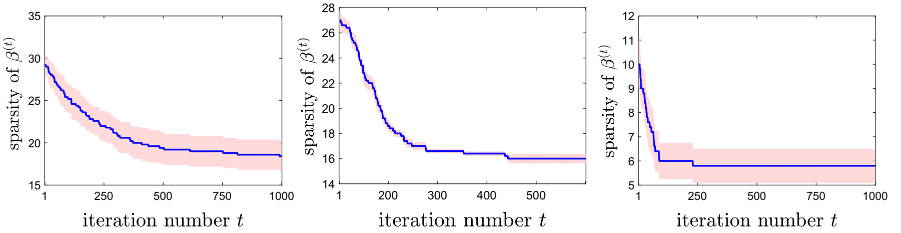
In this paper, we establish unified statistical and computational results for a simple variant of the TWF algorithm, which is robust over a large class of misspecified phase retrieval models. In specific, we prove that, with a proper initialization, the proposed algorithm linearly converges to an estimator with optimal statistical accuracy. Perhaps surprisingly, both our sample complexity $n = \Omega(s^2 \log p)$ and statistical rate of convergence $\sqrt{s \log p/n}$ match the best possible results for the case where the sparse phase retrieval model is correctly specified. To the best of our knowledge, our paper makes the first attempt to understand the robustness of nonconvex statistical optimization under model misspecification. We hope that our techniques can be further extended to more general problems in the future.

Appendix

A Further simulations

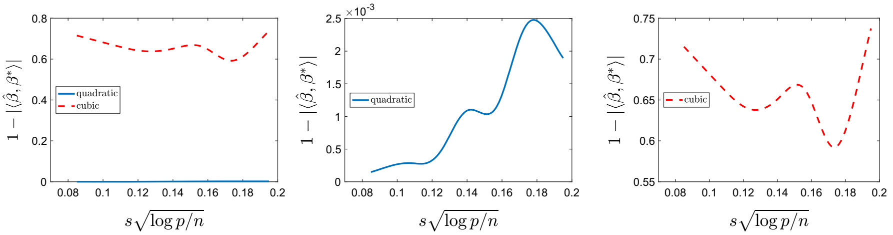
We further investigate how the sparsity level of the estimator changes over iterations empirically. For the three link function h_1, h_2, and h_3 defined in Eq. (4.1), we plot the sparsity of the TWF algorithm in Fig. 5 up to $T = 1000$ iterations. Similar to our experimental study of the optimization error, here we set $p = 1000$, $s = 5$, and $n = 864 \approx 5s^2 \log p$. Moreover, the tuning parameters are set as $\gamma = 2$, $\kappa = 15$, and $\eta = 0.005$. In this case, as we show in Sect. 4.1, our TWF algorithm converges geometrically to an estimator with high statistical accuracy. As shown in Fig. 5, our algorithm tends to over-estimate the sparsity. However, in all of these figures, the estimated sparsity gradually decreases as the algorithm proceeds. Furthermore, combining with the experiments in Sect. 4.1, although the estimator has more nonzero entries than the true parameter, the statistical error is satisfactory.

In addition, we show a failed example, which violates Assumption 3.1, for the readers to better understand to what extent the proposed algorithm is robust to model misspecification. In particular, consider the cubic link function $f(u, v) = u^3 + v$,



(a) $h_1(u, v) = |u| + v$ **(b)** $h_2(u, v) = |u + v|$, **(c)** $h_3(u, v) = 4u^2 + 3\sin(|u|) + v$

Fig. 5 Plots of the sparsity of the TWF updates $\beta^{(t)}$. Here the link function is one of h_1, h_2, and h_3. In addition, we set $p = 1000$, $s = 5$, $n = 864 \approx 5s^2 \cdot \log p$, and run the TWF algorithm for $T = 1000$ iterations. These figures are generated based on 50 independent trials for each link function



(a) Statistical errors for the two models **(b)** The phase retrieval model **(c)** The cubic model

Fig. 6 Plots of the statistical errors for the cubic model $Y = (X^\top \beta^*)^3 + \epsilon$ and the phase retrieval model $Y = (X^\top \beta^*)^2 + \epsilon$. We set $p = 1000$, $s = 5$, and let n vary. The plots are generated based on 100 independent trials for each (n, p, s). In **a**, we plot the two error curves together, which shows that the TWF algorithm incurs much larger error on the cubic model, whereas the error for the phase retrieval model becomes rather negligible in comparison. In **b** and **c**, we plot the two curves in **a** separately for presentation

which violates Assumption 3.1 since in this case we have $\mathrm{Cov}[Y,(X^\top\beta^*)^2]=0$. We compare this cubic model $Y=(X^\top\beta^*)^3+\epsilon$ with the phase retrieval model $Y=(X^\top\beta^*)^2+\epsilon$, where the link function is quadratic. We set $p=1000$, $s=5$, and let n vary. For each setting, we report the estimation error based on 100 independent trials. The statistical error of the TWF algorithm for these two models are plotted in Fig. 6a, which shows that our algorithm has nondecreasing estimation error for the cubic model even when n is very large. This is in sharp contrast with the phase retrieval model, where the estimation error is negligible. Moreover, we plot the two error curves separately in Fig. 6b and c to better see the details.

References

1. Ahmadi, A.A., Parrilo, P.A.: Some recent directions in algebraic methods for optimization and Lyapunov analysis. In: Laumond, J.P., Mansard, N., Lasserre, J.B. (eds.) Geometric and Numerical Foundations of Movements, pp. 89–112. Springer, Cham (2017)
2. Alquier, P., Biau, G.: Sparse single-index model. J. Mach. Learn. Res. **14**, 243–280 (2013)
3. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2014)
4. Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Berlin (2011)
5. Bunk, O., Diaz, A., Pfeiffer, F., David, C., Schmitt, B., Satapathy, D.K., van der Veen, J.F.: Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. Acta Crystallogr. Sect. A Found. Crystallogr. **63**, 306–314 (2007)
6. Cai, T.T., Li, X., Ma, Z.: Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. Ann. Stat. **44**, 2221–2251 (2016)
7. Candès, E.J., Eldar, Y.C., Strohmer, T., Voroninski, V.: Phase retrieval via matrix completion. SIAM Rev. **57**, 225–251 (2015)
8. Candès, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via Wirtinger flow: theory and algorithms. IEEE Trans. Inf. Theory **61**, 1985–2007 (2015)
9. Candès, E.J., Strohmer, T., Voroninski, V.: Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**, 1241–1274 (2013)
10. Chen, Y., Candès, E.: Solving random quadratic systems of equations is nearly as easy as solving linear systems. In: Advances in Neural Information Processing Systems (2015)
11. Coene, W., Janssen, G., de Beeck, M.O., Van Dyck, D.: Phase retrieval through focus variation for ultra-resolution in field-emission transmission electron microscopy. Phys. Rev. Lett. **69**, 3743 (1992)
12. Cook, R.D., Ni, L.: Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J. Am. Stat. Assoc. **100**, 410–428 (2005)
13. Foucart, S., Rauhut, H.: A Mathematical Introduction to Compressive Sensing. Springer, Berlin (2013)
14. Genzel, M.: High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. IEEE Trans. Inf. Theory **63**, 1601–1619 (2017)
15. Ghadimi, S., Lan, G.: Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. **23**, 2341–2368 (2013)
16. Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. **156**, 59–99 (2016)
17. Goldstein, L., Minsker, S., Wei, X.: Structured signal recovery from non-linear and heavy-tailed measurements (2016). arXiv preprint arXiv:1609.01025
18. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Johns Hopkins University Press (2012)
19. Gonçalves, M.L., Melo, J.G., Monteiro, R.D.: Convergence rate bounds for a proximal admm with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems (2017). arXiv preprint arXiv:1702.01850
20. Gu, Q., Wang, Z., Liu, H.: Sparse PCA with oracle property. In: Advances in Neural Information Processing Systems (2014)

21. Han, A.K.: Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. J. Econ. **35**, 303–316 (1987)
22. Harrison, R.: Phase problem in crystallography. J. Opt. Soc. Am. Part A Opt. Image Sci. **10**, 1046–1055 (1993)
23. Hong, M., Luo, Z.-Q., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J. Optim. **26**, 337–364 (2016)
24. Horowitz, J.L.: Semiparametric and Nonparametric Methods in Econometrics. Springer, Berlin (2009)
25. Jaganathan, K., Eldar, Y.C. and Hassibi, B.: Phase retrieval: an overview of recent developments (2015). arXiv preprint arXiv:1510.07713
26. Jiang, B., Liu, J.S.: Variable selection for general index models via sliced inverse regression. Ann. Stat. **42**, 1751–1786 (2014)
27. Johnstone, I.M., Lu, A.Y.: On consistency and sparsity for principal components analysis in high dimensions. J. Am. Stat. Assoc. **104**, 682–693 (2009)
28. Kakade, S.M., Kanade, V., Shamir, O., Kalai, A.: Efficient learning of generalized linear and single index models with isotonic regression. In: Advances in Neural Information Processing Systems (2011)
29. Kalai, A.T., Sastry, R.: The isotron algorithm: high-dimensional isotonic regression. In: Conference on Learning Theory (2009)
30. Kim, S., Kojima, M., Toh, K.-C.: A Lagrangian-DNN relaxation: a fast method for computing tight lower bounds for a class of quadratic optimization problems. Math. Program. **156**, 161–187 (2016)
31. Li, K.-C.: Sliced inverse regression for dimension reduction. J. Am. Stat. Assoc. **86**, 316–327 (1991)
32. Li, K.-C.: On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. J. Am. Stat. Assoc. **87**, 1025–1039 (1992)
33. Li, K.-C., Duan, N.: Regression analysis under link violation. Ann. Stat. **17**, 1009–1052 (1989)
34. Li, X., Yang, L.F., Ge, J., Haupt, J., Zhang, T., Zhao, T.: On quadratic convergence of DC proximal Newton algorithm for nonconvex sparse learning in high dimensions (2017). arXiv preprint arXiv:1706.06066
35. Lin, Q., Zhao, Z., Liu, J.S.: On consistency and sparsity for sliced inverse regression in high dimensions (2015). arXiv preprint arXiv:1507.03895
36. Loh, P.-L., Wainwright, M.J.: Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. J. Mach. Learn. Res. **16**, 559–616 (2015)
37. Lu, Z., Xiao, L.: On the complexity analysis of randomized block-coordinate descent methods. Math. Program. **152**, 615–642 (2015)
38. Marchesini, S., He, H., Chapman, H.N., Hau-Riege, S.P., Noy, A., Howells, M.R., Weierstall, U., Spence, J.C.: X-ray image reconstruction from a diffraction pattern alone. Phys. Rev. B **68**, 140101 (2003)
39. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman and Hall, Boca Raton (1989)
40. Miao, J., Ishikawa, T., Shen, Q., Earnest, T.: Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. Ann. Rev. Phys. Chem. **59**, 387–410 (2008)
41. Millane, R.: Phase retrieval in crystallography and optics. J. Opt. Soc. Am. A Opt. Image Sci. **7**, 394–411 (1990)
42. Netrapalli, P., Jain, P., Sanghavi, S.: Phase retrieval using alternating minimization. In: Advances in Neural Information Processing Systems (2013)
43. Neykov, M., Wang, Z., Liu, H.: Agnostic estimation for misspecified phase retrieval models. In: Advances in Neural Information Processing Systems (2016)
44. Parrilo, P.A.: Semidefinite programming relaxations for semialgebraic problems. Math. Program. **96**, 293–320 (2003)
45. Plan, Y., Vershynin, R.: The generalized lasso with non-linear observations. IEEE Trans. Inf. Theory **62**, 1528–1537 (2016)
46. Plan, Y., Vershynin, R., Yudovina, E.: High-dimensional estimation with geometric constraints (2014). arXiv preprint arXiv:1404.3749
47. Radchenko, P.: High dimensional single index models. J. Multivar. Anal. **139**, 266–282 (2015)
48. Sahinoglou, H., Cabrera, S.D.: On phase retrieval of finite-length sequences using the initial time sample. IEEE Trans. Circuits Syst. **38**, 954–958 (1991)
49. Shechtman, Y., Eldar, Y.C., Cohen, O., Chapman, H.N., Miao, J., Segev, M.: Phase retrieval with application to optical imaging: a contemporary overview. IEEE Signal Process. Mag. **32**, 87–109 (2015)

50. Stein, C.M.: Estimation of the mean of a multivariate normal distribution. Ann. Stat. **9**, 1135–1151 (1981)
51. Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. In: IEEE International Symposium on Information Theory (2016)
52. Sun, W., Wang, Z., Liu, H., Cheng, G.: Non-convex statistical optimization for sparse tensor graphical model. In: Advances in Neural Information Processing Systems (2015)
53. Tan, K.M., Wang, Z., Liu, H., Zhang, T.: Sparse generalized eigenvalue problem: optimal statistical rates via truncated rayleigh flow (2016). arXiv preprint arXiv:1604.08697
54. Thrampoulidis, C., Abbasi, E., Hassibi, B.: Lasso with non-linear measurements is equivalent to one with linear measurements. In: Advances in Neural Information Processing Systems (2015)
55. Waldspurger, I.: Phase retrieval with random gaussian sensing vectors by alternating projections (2016). arXiv preprint arXiv:1609.03088
56. Waldspurger, I., dÁspremont, A., Mallat, S.: Phase recovery, maxcut and complex semidefinite programming. Math. Program. **149**, 47–81 (2015)
57. Wang, Z., Liu, H., Zhang, T.: Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. Ann. Stat. **42**, 2164–2201 (2014)
58. Wang, Z., Lu, H., Liu, H.: Tighten after relax: Minimax-optimal sparse PCA in polynomial time. In: Advances in Neural Information Processing Systems (2014)
59. Weisser, T., Lasserre, J.-B., Toh, K.-C.: Sparse-BSOS: a bounded degree SOS hierarchy for large scale polynomial optimization with sparsity. Math. Program. Comput. **10**, 1–32 (2017)
60. Xu, Y., Yin, W.: A globally convergent algorithm for nonconvex optimization based on block coordinate update. J. Sci. Comput. **72**(2), 700–734 (2017)
61. Yang, Z., Balasubramanian, K., Liu, H.: High-dimensional non-gaussian single index models via thresholded score function estimation (2017)
62. Yang, Z., Wang, Z., Liu, H.: Estimating high-dimensional non-gaussian multiple index models via Stein's lemma. In: Advances in Neural Information Processing Systems (2017)
63. Yang, Z., Wang, Z., Liu, H., Eldar, Y.C., Zhang, T.: Sparse nonlinear regression: Parameter estimation and asymptotic inference. In: International Conference on Machine Learning (2015)
64. Yi, X., Wang, Z., Caramanis, C., Liu, H.: Optimal linear estimation under unknown nonlinear transform. In: Advances in Neural Information Processing Systems (2015)
65. Zhang, H., Chi, Y., Liang, Y.: Provable nonconvex phase retrieval with outliers: median truncated Wirtinger flow. In: International Conference on Machine Learning (2016)
66. Zhao, T., Liu, H., Zhang, T.: Pathwise coordinate optimization for sparse learning: algorithm and theory. Ann. Stat. **46**(1), 180–218 (2018). https://doi.org/10.1214/17-AOS1547
67. Zhu, L., Miao, B., Peng, H.: On sliced inverse regression with high-dimensional covariates. J. Am. Stat. Assoc. **101**, 630–643 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Zhuoran Yang[1] · Lin F. Yang[1] · Ethan X. Fang[2,3] · Tuo Zhao[4,5] · Zhaoran Wang[6] · Matey Neykov[7]

Zhuoran Yang
zy6@princeton.edu

Lin F. Yang
lin.yang@princeton.edu

Tuo Zhao
tuo.zhao@isye.gatech.edu

Zhaoran Wang
zhaoran@princeton.edu

Matey Neykov
mneykov@stat.cmu.edu

[1] Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA

[2] Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA

[3] Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802, USA

[4] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[5] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[6] Department of Industrial Engineering and Management Science, Northwestern University, Evanston, IL 60208, USA

[7] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA