Spectral Initialization for Nonconvex Estimation: High-Dimensional Limit and Phase Transitions

Yue M. Lu
Paulson School of Engineering and Applied Sciences
Harvard University, USA
Email: yuelu@seas.harvard.edu

Gen Li
Department of Electronic Engineering
Tsinghua University, China
Email: g-li16@mails.tsinghua.edu.cn

***Abstract*—We study a simple spectral method that serves as a key ingredient in a growing line of work using efficient iterative algorithms for estimating signals in nonconvex settings. Unlike previous work, which focuses on the phase retrieval setting and provides only bounds on the performance, we consider arbitrary generalized linear sensing models and provide an exact characterization of the performance of the spectral method in the high-dimensional regime. Our analysis reveals a phase transition phenomenon that depends on the sampling ratio. When the ratio is below a critical threshold, the estimates given by the spectral method are no better than random guesses drawn uniformly from the hypersphere; above the threshold, however, the estimates become increasingly aligned with the underlying signal. Worked examples and numerical simulations are provided to illustrate and verify the analytical predictions.**

I. Introduction

Let $\boldsymbol{\xi}$ be an unknown vector in $\mathbb{R}^n$ and $\{\boldsymbol{a}_i\}_{1\le i\le m}$ be a set of sensing vectors. We are interested in estimating $\boldsymbol{\xi}$ from a number of generalized linear measurements $\{y_i\}_{1\le i\le m}$: given $\boldsymbol{\xi}$ and $\{\boldsymbol{a}_i\}$, the ith measurement y_i is drawn independently from some distribution

$$\mathbb{P}_i(y) = f(y \,|\, \boldsymbol{a}_i^T\boldsymbol{\xi}), \tag{1}$$

where $f(\cdot\,|\,\cdot)$ is a conditional density function modeling the acquisition process. This model arises in many problems in signal processing and learning. Examples include photon-limited imaging [1], [2], phase retrieval [3], signal recovery from quantized measurements [4], and generalized linear regression [5].

The standard method for recovering $\boldsymbol{\xi}$ is to use the estimator

$$\widehat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{x}} \sum_{i=1}^m \ell(y_i, \boldsymbol{a}_i^T\boldsymbol{x}), \tag{2}$$

where $\ell : \mathbb{R}^2 \to \mathbb{R}$ is some loss function (*e.g.*, the negative log-likelihood of the observation model.) In many applications, however, the natural loss function is not convex with respect to $\boldsymbol{x}$. It is often the case that there is no effective way to convexify (2), or, if there is, the resulting convex relaxation can be computationally expensive. The problem of phase retrieval, where $y_i = (\boldsymbol{a}_i^T\boldsymbol{\xi})^2 + \varepsilon_i$ for some noise terms $\{\varepsilon_i\}$, is an example in the latter scenario. One could use convex relaxation schemes such as those based on semidefinite programming (*e.g.*, [6]–[9]), but the challenges facing these schemes lie in their actual implementation. In practice, the computational complexity and memory requirement associated with these methods are prohibitive for signal dimensions that are encountered in real-word applications such as imaging.

In light of these issues, there is recent interest in developing and analyzing efficient iterative methods that directly solve nonconvex forms of (2). Examples include the alternating minimization scheme for phase retrieval [10], the Wirtinger Flow algorithm and its variants [11]–[14], as well as iterative projection methods [15], [16]. A common ingredient that contributes to the success of these iterative algorithms is that they all use some sort of spectral methods as an initialization step, which is then followed by iterative refinement.

In this paper, we provide an exact high-dimensional analysis of a spectral method [6], [10], [12] for estimating $\boldsymbol{\xi}$. The method consists of only two steps: First, construct a data matrix from the sensing vectors and measurements as

$$\boldsymbol{D} \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m \mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^T, \tag{3}$$

where $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ is a preprocessing function (*e.g.* a truncation step.) Second, compute a normalized eigenvector, denoted by $\boldsymbol{x}_1$, that corresponds to the largest eigenvalue of $\boldsymbol{D}$. The vector $\boldsymbol{x}_1$ is then our estimate of $\boldsymbol{\xi}$ (up to an unknown scalar.)

This method was introduced by Netrapalli, Jain, and Sanghavi in [10] to address the problem of phase retrieval. Under the assumption that the sensing vectors consist of i.i.d. Gaussian random variables, the authors show that the leading eigenvector $\boldsymbol{x}_1$ is aligned with the true vector $\boldsymbol{\xi}$ in direction when there are sufficiently many measurements. More specifically, they show that the squared cosine similarity

$$\rho(\boldsymbol{\xi}, \boldsymbol{x}_1) \overset{\text{def}}{=} \left(\frac{\boldsymbol{\xi}^T\boldsymbol{x}_1}{\|\boldsymbol{\xi}\|\|\boldsymbol{x}_1\|}\right)^2, \tag{4}$$

which measures the degree of the alignment between the two vectors, approaches 1 with high probability, whenever $m \ge c\,n\log^3 n$. This requirement on sample complexity was later improved to $m \ge c\,n\log n$ in [6], and further improved to $m \ge c\,n$ in [12] with an additional truncation step on the measurements. In all these expressions, c denotes some unspecified large constant.

In this paper, we present an exact asymptotic analysis of the performance of the spectral method under Gaussian

measurements. Although it was originally proposed for the special case of phase retrieval, the spectral method is applicable to a much broader class of models. Therefore, in our analysis, we consider general acquisition models under arbitrary conditional distributions $f(y \,|\, \boldsymbol{a}_i^T \boldsymbol{\xi})$. Unlike previous work, which only provides bounds for $\rho(\boldsymbol{\xi}, \boldsymbol{x}_1)$, we derive the exact high-dimensional limit of this value. In particular, we show that, as n and m both tend to infinity with the sampling ratio $\alpha \overset{\text{def}}{=} m/n$ kept fixed, the squared cosine similarity ρ converges in probability to a limit value $\rho(\alpha)$. Our analysis also reveals a phase transition phenomenon that occurs at a critical sampling ratio α_c:

(a) Below the threshold $\alpha < \alpha_c$: The limiting value $\rho(\alpha) = 0$, meaning that the estimate from the spectral method is orthogonal to the true vector $\boldsymbol{\xi}$. In this case, the spectral method is not effective. Its estimate $\boldsymbol{x}_1$ is not any better than a random guess drawn uniformly from the hypersphere $\mathcal{S}^{n-1}$ embedded in $\mathbb{R}^n$.

(b) Above the threshold $\alpha > \alpha_c$: The limiting value $\rho(\alpha) > 0$ and it tends to 1 as $\alpha \to \infty$.

This phase transition phenomenon also has implications in terms of the computational complexity of the spectral method. When $\alpha > \alpha_c$, there is a nonzero gap between the largest and the second largest eigenvalues of $\boldsymbol{D}$. As a result, the leading eigenvector $\boldsymbol{x}_1$ can be efficiently computed by power iterations on $\boldsymbol{D}$. In contrast, when $\alpha < \alpha_c$, the gap of the eigenvalues converges to zero, rendering power iterations ineffective.

The rest of the paper is organized as follows. After precisely laying out the various technical assumptions, we present the main results of this work in Section II. Worked-examples and numerical simulations are provided in Section III to demonstrate and verify these results. We highlight in Section IV several key propositions and proof sketches, but leave the full technical details to [17]. Section V concludes the paper.

Notations: To study the high-dimensional limit of the spectral initialization method, we consider a sequence of problem instances, indexed by the ambient dimension n. For each n, we seek to estimate an underlying signal denoted by $\boldsymbol{\xi}^n \in \mathbb{R}^n$. Formally, we should use D_n and $\boldsymbol{x}_{n,1}$ to denote the data matrix and its leading eigenvector. However, to lighten the notation, we shall drop the subscript n whenever doing so causes no ambiguity. We use $\xrightarrow{\mathcal{P}}$ and $\xrightarrow{\text{a.s.}}$ to denote convergence in probability and almost sure convergence, respectively.

II. Main Results

A. Technical Assumptions

In what follows, we first state the assumptions under which our results are proved.

(A1) The sensing vectors are independent random vectors whose entries are i.i.d. standard normal random variables.

(A2) $m = m(n)$ with $\alpha_n = m(n)/n \to \alpha > 0$ as $n \to \infty$.

(A3) $\|\boldsymbol{\xi}\| = \theta > 0$.

(A4) Let s, y and z be three random variables such that

$$s \sim \mathcal{N}(0,1),\ \mathbb{P}(y \,|\, s) = f(y \,|\, \theta s),\ \text{and}\ z = \mathcal{T}(y), \tag{5}$$

where $f(\cdot \,|\, \cdot)$ is the conditional density function associated with the observation model, and $\mathcal{T}(\cdot)$ is the preprocessing step used in the construction of $\boldsymbol{D}$ in (3). We assume that the probability distribution of the random variable z is supported within a finite interval $[0, \tau]$. In what follows, we take τ to be the tightest such upper bound.

(A5) The random variables z and s^2 are positively correlated, *i.e.*, $\operatorname{cov}(z, s^2) = \mathbb{E}\, zs^2 - \mathbb{E}z\, \mathbb{E}s^2 = \mathbb{E}\, zs^2 - \mathbb{E}z > 0$.

The last two assumptions require some explanations. First, although the observations $\{y_i\}$ are not necessarily bounded, the boundedness of z, as required in (A4), can be enforced by choosing a suitable function $\mathcal{T}(\cdot)$. For example, in the problem of phase retrieval, the measurement model is $y = s^2$, and we can choose $z = \mathcal{T}(y) = y\, \mathbb{1}(y < \tau)$, where $\mathbb{1}(\cdot)$ is the indicator function. This is the truncation strategy proposed in [12].

The inequality in (A5) is also a natural requirement. To see this, we note that the data matrix $\boldsymbol{D}$ in (3) is the sample average of m i.i.d. random rank-one matrices $\left\{y_i \boldsymbol{a}_i \boldsymbol{a}_i^T\right\}_{i \le m}$. When the number of samples m is large, this sample average should be "close" to the statistical expectation:

$$\boldsymbol{D} \approx \mathbb{E}(z_i\, \boldsymbol{a}_i \boldsymbol{a}_i^T), \tag{6}$$

where $z_i \overset{\text{def}}{=} \mathcal{T}(y_i)$. To compute the above expectation, it will be convenient to assume that the underlying signal $\boldsymbol{\xi} = \theta \boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the first vector of the canonical basis of $\mathbb{R}^n$. (This assumption can be made without loss of generality, due to the rotational invariance of the multivariate normal distribution.) Correspondingly, we can partition each sensing vector into two parts, as

$$\boldsymbol{a}_i^T = \begin{bmatrix} s_i & \boldsymbol{u}_i^T \end{bmatrix}, \tag{7}$$

so that $\boldsymbol{a}_i^T \boldsymbol{\xi} = \theta s_i$ and $\mathbb{P}(y_i \,|\, s_i) = f(y \,|\, \theta s_i)$. Since s_i, y_i and z_i are all independent to $\boldsymbol{u}_i$, we get

$$\begin{aligned} \mathbb{E}(z_i\, \boldsymbol{a}_i \boldsymbol{a}_i^T) &= \mathbb{E} \begin{bmatrix} z_i s_i^2 & z_i s_i \boldsymbol{u}_i^T \\ z_i s_i \boldsymbol{u}_i & z_i \boldsymbol{u}_i \boldsymbol{u}_i^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}\, zs^2 & \boldsymbol{0} \\ \boldsymbol{0} & \mathbb{E}z\, \boldsymbol{I}_{n-1} \end{bmatrix}, \end{aligned} \tag{8}$$

where $\boldsymbol{I}_{n-1}$ is the identity matrix of size $(n-1) \times (n-1)$. We see that, if the inequality $\mathbb{E}\, zs^2 > \mathbb{E}z$, as required in (A5), indeed holds, the leading eigenvector of the expected matrix will be $\boldsymbol{e}_1$, which is perfectly aligned with the true vector $\boldsymbol{\xi}$. Now since the data matrix $\boldsymbol{D}$ is an approximation of the expectation, the sample eigenvector should also be an approximation of $\boldsymbol{\xi}$.

The above argument provides an intuitive but nonrigorous explanation for why the spectral initialization method would work. The approximation in (6) can be made exact if the signal dimension n is kept fixed and the number of measurement m goes to infinity. However, we consider the case when m and n both tend to infinity, at a finite and constant ratio $\alpha = m/n$. In this regime, the approximation in (6) will not become an equality even if $m \to \infty$. As we will show, in this case, the correlation $\rho(\boldsymbol{\xi}, \boldsymbol{x}_1)$ between the sample eigenvector $\boldsymbol{x}_1$ and the true vector $\boldsymbol{\xi}$ will converge to a function $\rho(\alpha)$.

B. Main Results: Asymptotic Characterizations

In this section, we summarize the main results of our work on an asymptotic characterization of the spectral method with random Gaussian measurements. To state our results, we first need to introduce several helper functions. Let s, z be the random variables defined in (5). We consider two functions

$$\phi(\lambda) \overset{\text{def}}{=} \lambda\, \mathbb{E}\frac{zs^2}{\lambda - z} \tag{9}$$

and

$$\zeta_\alpha(\lambda) \overset{\text{def}}{=} \frac{\lambda}{\alpha} + \lambda\, \mathbb{E}\frac{z}{\lambda - z}, \tag{10}$$

both defined on the open interval $\lambda \in (\tau, \infty)$, where τ is the bound in assumption (A4). Within their domains, both functions are convex. In particular, $\zeta_\alpha(\lambda)$ achieves its minimum at a unique point denoted by

$$\bar{\lambda}_\alpha \overset{\text{def}}{=} \arg\min_{\lambda > \tau} \zeta_\alpha(\lambda). \tag{11}$$

Finally, let $\psi_\alpha(\lambda)$ be a modification of $\zeta_\alpha(\lambda)$ as follows:

$$\psi_\alpha(\lambda) = \zeta_\alpha\big(\max\{\lambda, \bar{\lambda}_\alpha\}\big) \tag{12}$$

Theorem 1: Under assumptions (A1) – (A5), the following holds:

1) There is a unique solution, denoted by λ_α^*, to the equation
$$\psi_\alpha(\lambda) = \phi(\lambda), \qquad \lambda > \tau. \tag{13}$$
2) As $n \to \infty$,
$$\rho(\boldsymbol{\xi}, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} \begin{cases} 0, & \text{if } \zeta_\alpha'(\lambda_\alpha^*) \le 0, \\ \frac{\psi_\alpha'(\lambda_\alpha^*)}{\psi_\alpha'(\lambda_\alpha^*) - \phi'(\lambda_\alpha^*)}, & \text{otherwise}, \end{cases} \tag{14}$$
where $\psi_\alpha'(\cdot), \zeta_\alpha'(\cdot)$ and $\phi'(\cdot)$ denote the derivatives of the three functions.
3) Let $\lambda_1 \ge \lambda_2$ be the top two eigenvalues of $\boldsymbol{D}$. Then
$$\lambda_1 \xrightarrow{\mathcal{P}} \psi_\alpha(\lambda_\alpha^*) \quad \text{and} \quad \lambda_2 \xrightarrow{\mathcal{P}} \psi_\alpha(\bar{\lambda}_\alpha) \tag{15}$$
as $n \to \infty$. Moreover, $\psi_\alpha(\lambda_\alpha^*) \ge \psi_\alpha(\bar{\lambda}_\alpha)$, and this inequality becomes strict if and only if $\zeta_\alpha'(\lambda_\alpha^*) > 0$.

Remark 1: The above theorem provides a complete asymptotic characterization of the performance of the spectral initialization method. In particular, the theorem shows that the squared cosine similarity $\rho(\boldsymbol{\xi}, \boldsymbol{x}_1)$ converges in probability to a deterministic limiting value in the high-dimensional limit. Moreover, there exists a generic phase transition phenomenon: depending on the sign of the derivative $\zeta_\alpha'(\cdot)$ at λ_α^*, the limiting value can be either zero or strictly positive. The computational complexity of the spectral method is also markedly different in the two phases. Within the uncorrelated phase, the gap between the top two leading eigenvalues, λ_1 and λ_2, diminishes to zero, making iterative algorithms such as power iterations increasingly difficult to converge. In contrast, within the correlated phase, the spectral gap converges to a positive value.

It will be more convenient to characterize the above phase transition in terms of a critical value α_c of the sampling ratio. To do so, we first introduce an additional technical assumption:

(A6) Within the interval $(\tau, \frac{1}{1 - \sqrt{\mathbb{E}z / \mathbb{E}zs^2}}\tau)$, the equation

$$\lambda\, \mathbb{E}\frac{z}{(\lambda - z)^2} = \mathbb{E}\frac{zs^2}{\lambda - z} \tag{16}$$

has a *unique* solution, denoted by λ_c.

Note that (16) can be shown to always admit at least one solution within the above interval. (A6) requires the uniqueness of this solution. (See [17] for an example where this uniqueness assumption does not hold, in which case multiple phase transitions take place as we increase the value of α.)

Proposition 1: Under (A1) – (A6), and as $n \to \infty$,

$$\rho(\boldsymbol{\xi}, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} \begin{cases} 0, & \text{if } \alpha < \alpha_c, \\ \rho(\alpha), & \text{if } \alpha > \alpha_c, \end{cases}$$

where $\alpha_c = \big(\mathbb{E}\frac{z^2}{(\lambda_c - z)^2}\big)^{-1}$ and $\rho(\alpha)$ is a function with the following parametric representation:

$$\alpha = \Big(\mathbb{E}\frac{zs^2 - z}{\lambda - z}\Big)^{-1}$$
$$\rho = \Big[1 + \mathbb{E}\frac{z^2 s^2}{(\lambda - z)^2} \Big/ \Big(\mathbb{E}\frac{zs^2 - z}{\lambda - z} - \mathbb{E}\frac{z^2}{(\lambda - z)^2}\Big)\Big]^{-1},$$

for $\lambda > \lambda_c$. Moreover, $\rho(\alpha) \to 1$ as $\alpha \to \infty$.

III. Examples and Numerical Simulations

A. One-Bit Measurement Models

To illustrate the results presented in Section II, we consider here a special case where z_i takes only binary values $\{0, 1\}$. This type of one-bit measurement models appear in problems such as binary logistic regression and one-bit quantized sensing. The simplicity of this setting allows us to get closed-form expressions for the various quantities in Theorem 1 and Proposition 1.

To proceed, we first explicitly compute the functions $\phi(\lambda)$ and $\zeta_\alpha(\lambda)$ defined in Section II-B and get

$$\phi(\lambda) = \frac{c\lambda}{\lambda - 1} \quad \text{and} \quad \zeta_\alpha(\lambda) = \frac{\lambda}{\alpha} + \frac{d\lambda}{\lambda - 1},$$

where

$$c \overset{\text{def}}{=} \mathbb{E}\, zs^2 \quad \text{and} \quad d \overset{\text{def}}{=} \mathbb{E}\, z \tag{17}$$

and both functions are defined on the interval $\lambda > 1$. The minimum of $\zeta_\alpha(\lambda)$ is achieved as $\bar{\lambda}_\alpha = 1 + \sqrt{\alpha d}$, and thus

$$\psi_\alpha(\lambda) = \begin{cases} \lambda/\alpha + \lambda d/(\lambda - 1), & \text{for } \lambda \ge 1 + \sqrt{\alpha d} \\ (\sqrt{d} + 1/\sqrt{\alpha})^2, & \text{for } 1 < \lambda < 1 + \sqrt{\alpha d}. \end{cases}$$

Solving equation (13) and using (14), we get

$$\rho(\boldsymbol{\xi}, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} \begin{cases} 0, & \text{for } \alpha < \alpha_c, \\ \frac{\alpha - d/(c-d)^2}{\alpha + 1/(c-d)}, & \text{for } \alpha > \alpha_c, \end{cases} \tag{18}$$

where $\alpha_c = \frac{d}{(c-d)^2}$. Finally, the asymptotic predictions (15) for the top two eigenvalues can be computed as

$$\lambda_1 \xrightarrow{\mathcal{P}} \begin{cases} (\sqrt{d} + 1/\sqrt{\alpha})^2, & \text{for } \alpha < \alpha_c, \\ c + \frac{c}{\alpha(c-d)}, & \text{for } \alpha > \alpha_c, \end{cases} \tag{19}$$

and $\lambda_2 \xrightarrow{\mathcal{P}} (\sqrt{d} + 1/\sqrt{\alpha})^2$ for all α.

B. Numerical Simulations

Example 1 (Logistic regression): Consider the case where $\{y_i\}$ are binary random variables generated according to the following conditional distribution:

$$f(y \,|\, \boldsymbol{a}^T\boldsymbol{\xi}) = \text{Bernoulli}\left(\frac{1}{1+\exp\{-\boldsymbol{a}^T\boldsymbol{\xi}-\beta\}}\right), \qquad (20)$$

where β is some constant. Let $z_i = \mathcal{T}(y_i) = y_i$. Since $z_i \in \{0,1\}$, we just need to compute the quantities c and d in (17), after which we can use the closed-form formulas (18) and (19) to obtain the asymptotic predictions. In Figure 1(a) we compare the analytical prediction (18) of the squared cosine similarity with results obtained by numerical simulations. In our experiment, we set the signal dimension to be $n = 5000$. The norm of $\boldsymbol{\xi}$ is $\theta = 3$, and $\beta = 6$. The sample averages and error bars (corresponding to one standard deviation) shown in the figure are calculated over 16 independent Monte Carlo trials. We can see that the analytical predictions match with numerical results very well. Figure 1(b) shows the results for the top two eigenvalues. When $\alpha < \alpha_c$, the two eigenvalues are asymptotically equal, but they start to diverge as α becomes larger than α_c. To clearly illustrate this effect, we plot in the insert the eigengap $\lambda_1 - \lambda_2$ as a function of α.

Example 2 (Phase retrieval): In the second example, we consider the problem of phase retrieval. For simplicity, we assume the noiseless setting, where $y_i = (\boldsymbol{a}_i^T\boldsymbol{\xi})^2$. In [12], the authors show that is important to truncate large values of $\{y_i\}$, proposing to use

$$z_i = \mathcal{T}(y_i) = y_i\,\mathbb{1}(y_i < t), \qquad (21)$$

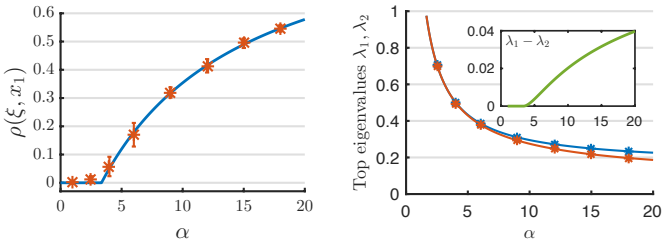
for some threshold value $t > 0$, when constructing the data matrix $\boldsymbol{D}$. A different strategy can be found in [14], where the authors propose to use

$$z_i = \mathbb{1}(y_i > t). \qquad (22)$$

In what follows, we shall refer to (21) and (22) as the truncation algorithm and the subset algorithm, respectively. Figure 2(a) shows the asymptotic performance of these two algorithms and compare them with numerical results ($n = 5000$ and 16 independent trials.) The performance of the subset algorithm (for which we choose the parameter $t = 1.5$) can be characterized by the closed-form formula (18). The truncation algorithm (for which we use $t = 3$) is more complicated as z_i is no longer binary. We use the parametric characterization in Proposition 1 to obtain its asymptotic performance. Again, our analytical predictions have excellent matches with numerical results. The performance of both algorithms clearly depends on the choice of the thresholding parameter t. To show this, we plot in Figure 2(b) the critical phase transition points α_c of both algorithms as functions of t. This points to the potential of using our analytical prediction to optimally tune the algorithmic parameters.

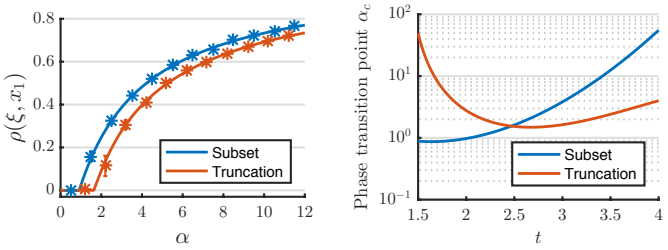
IV. Proof Sketch of Theorem 1

There are three main ingredients in the proof, stated below as Proposition 2, Proposition 3, and Proposition 4. We give some intuition and comment on their statements. Detailed proofs can be found in [17].



(a) Squared cosine similarity (b) Top two eigenvalues

Fig. 1. Analytical predictions v.s. numerical simulations for the binary logistic model in (20). Numerical results are averaged over 16 independent Monte Carlo trials.



(a) Squared cosine similarity (b) Phase transition points

Fig. 2. (a) Analytical predictions v.s. numerical simulations for two different algorithms for phase retrieval. (b) The critical sampling ratio α_c of the two algorithms as functions of the threshold value t.

In what follows, we assume without loss of generality that $\boldsymbol{\xi} = \theta\boldsymbol{e}_1$. Partitioning each sensing vector into two parts as in (7), we can rewrite the data matrix $\boldsymbol{D}$ in (3) in a block-partitioned form

$$\boldsymbol{D} = \begin{bmatrix} a & \boldsymbol{q}^T \\ \boldsymbol{q} & \boldsymbol{P} \end{bmatrix}.$$

Here, $a \overset{\text{def}}{=} \sum_{i=1}^m z_i s_i^2/m$ is a scalar; $\boldsymbol{q} = \frac{1}{m}\boldsymbol{U}\boldsymbol{v} \in \mathbb{R}^{n-1}$; $\boldsymbol{U}$ is an $(n-1)\times m$ matrix consisting of i.i.d. standard normal random variables; $\boldsymbol{v} \overset{\text{def}}{=} [z_1 s_1, z_2 s_2, \ldots, z_m s_m]^T \in \mathbb{R}^m$; and $\boldsymbol{P} = \frac{1}{m}\boldsymbol{U}\text{diag}\{z_1, z_2, \ldots, z_m\}\boldsymbol{U}^T$.

Our goal boils down to studying the largest eigenvalue of $\boldsymbol{D}$ and the magnitude of the first element of the associated eigenvector. Readers familiar with the spiked covariance model (see, *e.g.*, [18], [19]) might recognize that our task bears resemblance to this classical problem. The unique challenge in our setting is that the vector $\boldsymbol{q}$ and the matrix $\boldsymbol{P}$ are correlated through $\{s_i\}$. To "decouple" this correlation, we introduce a parametric family of companion matrices $\boldsymbol{M}(\mu) = \boldsymbol{P} + \mu\boldsymbol{q}\boldsymbol{q}^T$ for $\mu > 0$. We also use $L(\mu)$ to denote the largest eigenvalue of $\boldsymbol{M}(\mu)$.

The following proposition shows that the desired quantities of the data matrix $\boldsymbol{D}$, such as its leading eigenvalue λ_1 and the squared cosine similarity between $\boldsymbol{\xi}$ and the leading eigenvector $\boldsymbol{x}_1$, can be obtained from the function $L(\mu)$.

Proposition 2: Let μ^* be the unique solution to $L(\mu) =$

$a + 1/\mu$. Then, $\lambda_1 = L(\mu^*)$ and

$$\rho(\boldsymbol{\xi}, \boldsymbol{x}_1) = \frac{L'(\mu^*)}{L'(\mu^*) + (1/\mu^*)^2}. \tag{23}$$

The above characterization, which is the first ingredient of the proof of Theorem 1, is valid for any finite dimension n and for any deterministic sensing vectors $\{\boldsymbol{a}_i\}$. Next, we specialize to the case of i.i.d. Gaussian measurements and show that $L(\mu)$ converges almost surely to some deterministic function as $n \to \infty$. To that end, we note that $\boldsymbol{M}(\mu)$ can be written as

$$\boldsymbol{M}(\mu) = \tfrac{1}{m}\boldsymbol{U}\boldsymbol{W}_m\boldsymbol{U}^T,$$

where $\boldsymbol{W}_m \overset{\text{def}}{=} \text{diag}\,\{z_1, z_2, \ldots, z_m\} + \frac{\mu}{m}\boldsymbol{v}\boldsymbol{v}^T$ is a rank-one perturbation of a diagonal matrix. Since $\boldsymbol{U}$ and $\boldsymbol{W}_m$ are independent, we first study the spectrum of $\boldsymbol{W}_m$.

Let $\lambda_1^{W_m} \geq \lambda_2^{W_m} \geq \ldots \geq \lambda_m^{W_m}$ be the set of eigenvalues of $\boldsymbol{W}_m$ in descending order. Let $F^{W_m}(\lambda) \overset{\text{def}}{=} \frac{1}{m-1}\sum_{i=2}^{m}\delta(\lambda - \lambda_i^{W_m})$ denote the empirical distribution of the last $m-1$ eigenvalues.

Proposition 3: As $n \to \infty$, the empirical distribution $F^{W_m}(\lambda)$ converges almost surely to the probability law of the random variable z, denoted by $F_z(\lambda)$. Meanwhile,

$$\lambda_1^{W_m} \xrightarrow{\text{a.s.}} \lambda_\mu, \tag{24}$$

where λ_μ is the unique solution to $\frac{1}{\mu} = \mathbb{E}\frac{z^2 s^2}{\lambda - z}$ on the interval $\lambda \in (\tau, \infty)$ and τ is the upper bound of the support of $F_z(\lambda)$.

Recall from Assumption (A4) stated in Section II-A that the law of z is supported within the interval $[0, \tau]$. The above proposition, which is the second main ingredient of our proof of Theorem 1, shows that the spectrum of W_m consists of two parts: a base set of $m-1$ eigenvalues supported within $[0, \tau]$ and a single spiked eigenvalue $\lambda_1^{W_m}$ well separated from the base set. This setting is a generalization of the classical spiked population model [18]. Directly applying and specializing the result in [19], we reach the third ingredient of the proof, characterizing the leading eigenvalue of $\boldsymbol{M}(\mu)$ in the asymptotic limit.

Proposition 4: For every fixed $\mu > 0$,

$$L_n(\mu) \xrightarrow{\text{a.s.}} \psi_\alpha(\lambda_\mu), \tag{25}$$

where $\psi_\alpha(\cdot)$ is the function defined in (12) and λ_μ is the limit value in (24).

Substituting the limiting function $\psi_\alpha(\lambda_\mu)$ into (23), and after some simple manipulations, we can (formally) reach the expression for the limiting squared cosine similarity given in (14). To rigorously prove the convergence, we still need to extend the pointwise convergence in (25) to uniform convergence over compact intervals and to establish the convergence of the derivatives. Details can be found in [17].

V. Conclusion

We presented a precise characterization of the asymptotic performance of a spectral method for estimating signals from generalized linear measurements with Gaussian sensing vectors. Our analysis reveals a phase transition phenomenon that takes place at a critical sampling ratio. Below this threshold, estimates given by the methods are nearly orthogonal to the true signal $\boldsymbol{\xi}$, thus carrying no information; above the threshold, the estimates become increasingly aligned with $\boldsymbol{\xi}$. The computational complexity of the spectral method is also markedly different in the two phases. Within the uncorrelated phase, the gap between the top two leading eigenvalues diminishes to zero. In contrast, a nonzero spectral gap emerges within the correlated phase. Although our analysis is asymptotic in nature, numerical simulations show that the predictions are accurate even for moderate signal dimensions.

References

[1] M. Unser and M. Eden, "Maximum likelihood estimation of linear signal parameters for Poisson processes," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 36, no. 6, pp. 942–945, Jun. 1988.

[2] F. Yang, Y. M. Lu, L. Sbaiz, and M. Vetterli, "Bits from photons: Oversampled image acquisition using binary poisson statistics," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1421–1436, 2012.

[3] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied Optics*, vol. 21, no. 15, pp. 2758–2769, 1982.

[4] S. Rangan and V. K. Goyal, "Recursive consistent estimation with bounded noise," *Information Theory, IEEE Transactions on*, vol. 47, no. 1, pp. 457–464, 2001.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd ed. New York, NY: Springer, Apr. 2011.

[6] E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[7] E. J. Candes and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Foundations of Computational Mathematics*, vol. 14, no. 5, pp. 1017–1026, 2014.

[8] K. Jaganathan, S. Oymak, and B. Hassibi, "Sparse phase retrieval: Convex algorithms and limitations," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 1022–1026.

[9] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Mathematical Programming*, vol. 149, no. 1-2, pp. 47–81, 2015.

[10] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[11] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *Information Theory, IEEE Transactions on*, vol. 61, no. 4, pp. 1985–2007, 2015.

[12] Y. Chen and E. J. Candes, "Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems," *arXiv preprint arXiv:1505.05114*, 2015.

[13] H. Zhang and Y. Liang, "Reshaped Wirtinger Flow for Solving Quadratic System of Equations," in *Advances in Neural Information Processing Systems*, 2016, pp. 2622–2630.

[14] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving Systems of Random Quadratic Equations via Truncated Amplitude Flow," *arXiv:1605.08285*, May 2016.

[15] G. Li, Y. Gu, and Y. M. Lu, "Phase retrieval using iterative projections: Dynamics in the large systems limit," in *Proc. Allerton Conference on Communication, Control and Computing.*, Monticello, IL, Oct 2015.

[16] Y. Chi and Y. M. Lu, "Kaczmarz method for solving quadratic equations," *IEEE Signal Process. Lett.*, vol. 23, no. 9, 2016.

[17] Y. M. Lu and G. Li, "Phase transitions of spectral initialization for high-dimensional nonconvex estimation," *arXiv:1702.06435 [cs.IT]*, 2017. [Online]. Available: https://arxiv.org/abs/1702.06435

[18] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *The Annals of Statistics*, vol. 29, no. 2, pp. 295–327, Apr. 2001.

[19] Z. D. Bai and J. Yao, "On sample eigenvalues in a generalized spiked population model," *Journal of Multivariate Analysis*, vol. 106, no. 167–177, 2012.