Convolutional Phase Retrieval

Qing Qu
Columbia University
qq2105@columbia.edu

Yuqian Zhang
Columbia University
yz2409@columbia.edu

Yonina C. Eldar
Technion
yonina@ee.technion.ac.il

John Wright
Columbia University
jw2966@columbia.edu

Abstract

We study the *convolutional phase retrieval* problem, which considers recovery of an unknown signal $\boldsymbol{x} \in \mathbb{C}^n$ from m measurements consisting of the magnitude of its cyclic convolution with a known kernel $\boldsymbol{a}$ of length m. This model is motivated by applications to channel estimation, optics, and underwater acoustic communication, where the signal of interest is acted on by a given channel/filter, and phase information is difficult or impossible to acquire. We show that when $\boldsymbol{a}$ is random and m is sufficiently large, $\boldsymbol{x}$ can be efficiently recovered up to a global phase using a combination of spectral initialization and generalized gradient descent. The main challenge is coping with dependencies in the measurement operator; we overcome this challenge by using ideas from decoupling theory, suprema of chaos processes and the restricted isometry property of random circulant matrices, and recent analysis for alternating minimizing methods.

1 Introduction

We study the problem of recovering a unknown signal $\boldsymbol{x} \in \mathbb{C}^n$ from measurements $\boldsymbol{y} = |\boldsymbol{a} \circledast \boldsymbol{x}|$, which consist of the magnitude of the convolution of $\boldsymbol{x}$ and a given filter $\boldsymbol{a} \in \mathbb{C}^m$,

$$\text{find} \quad \boldsymbol{z}, \quad \text{s.t.} \quad \boldsymbol{y} = |\boldsymbol{a} \circledast \boldsymbol{z}|\,, \tag{1}$$

where $\circledast$ denotes cyclic convolution. Let $\boldsymbol{C}_{\boldsymbol{a}} \in \mathbb{C}^{m\times m}$ be a circulant matrix generated by $\boldsymbol{a}$, and let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ be a matrix formed by the first n columns of $\boldsymbol{C}_{\boldsymbol{a}}$. Then the *convolutional phase retrieval* problem can be rewritten in the common matrix-vector form

$$\text{find} \quad \boldsymbol{z}, \quad \text{s.t.} \quad \boldsymbol{y} = |\boldsymbol{A}\boldsymbol{z}|\,. \tag{2}$$

This problem is motivated by applications like *channel estimation* [37, 1], *(non)coherent optical communication* [14, 24], and *underwater acoustic communication* [31]. For example, in millimeter-wave (mm-wave) wireless communications for 5G networks [27], one important problem is to reconstruct signal angle of arrival (AoA) from measurements, which are taken by the convolution of signal AoA and the antenna pattern.

Because of technical difficulties that the phase measurements are either very noisy and unreliable, or expensive to acquire, it is preferred to only take measurements of signal magnitude and the phase information is lost.

Most known results on the exact solution of phase retrieval problems [8, 29, 10, 38, 36, 35] pertain to *generic random matrices*, where the entries of $\boldsymbol{A}$ are independent subgaussian random variables. However, in practice it is almost impossible to design purely random measurement matrices: in many cases as we mentioned above, the measurement is much more

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

structured – generated by passing a signal through a manually designed filter. Moreover, the structured measurements often admit more efficient numerical methods: by using the *fast Fourier transform* for matrix-vector products, the benign structure of the convolutional model (1) allows us to design methods with $\mathcal{O}(m)$ memory and $\mathcal{O}(m \log m)$ computation cost per iteration. While for generic measurements, the cost is around $\mathcal{O}(mn)$.

In this work, we study the convolutional phase retrieval problem (1) under the assumption that the kernel $\boldsymbol{a} = [a_1, \cdots, a_m]^\top$ is random, with each entry i.i.d. complex Gaussian,

$$\boldsymbol{a} = \boldsymbol{u} + \mathrm{i}\boldsymbol{v}, \quad \boldsymbol{u}, \boldsymbol{v} \sim_{\mathrm{iid}} \mathcal{N}\left(\boldsymbol{0}, \tfrac{1}{2}\boldsymbol{I}\right). \tag{3}$$

Compared to the generic random measurement, as we can see, the random convolution model we study here is far more structured: it is parameterized by only $\mathcal{O}(m)$ independent complex normal random variables, whereas the generic model involves $\mathcal{O}(mn)$ ones. Since the rows and columns of $\boldsymbol{A}$ are probabilistically dependent, standard techniques (based on concentration of functions of independent random vectors) do not apply.

We propose and analyze a local[1] gradient descent type method, minimizing a weighted, *nonconvex* and *nonsmooth* objective

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} f(\boldsymbol{z}) = \frac{1}{2m} \left\| \boldsymbol{b}^{1/2} \odot (\boldsymbol{y} - |\boldsymbol{A}\boldsymbol{z}|) \right\|^2, \tag{4}$$

where $\odot$ denotes the Hadamard product and $\boldsymbol{b} \in \mathbb{R}^m_{++}$ is a weighting vector. Our result can be informally summarized as follows.

Theorem 1.1 (Informal) *When $m \geq \Omega(n \operatorname{poly}\log n)$, with high probability, spectral initialization [25, 5] produces an initialization $\boldsymbol{z}^{(0)}$ that is $\mathcal{O}(1/\operatorname{poly}\log n)$ close to the optimum. Moreover, when $m \geq \Omega\left(\frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2} n \operatorname{poly}\log n\right)$, with high probability, a certain gradient descent method based on* (4) *converges linearly from this initialization to the set $\mathcal{X} = \left\{\boldsymbol{x}e^{\mathrm{i}\phi} \mid \phi \in [0, 2\pi)\right\}$ of points that differ from the true signal $\boldsymbol{x}$ only by a global phase.*

Here, $\boldsymbol{C}_{\boldsymbol{x}} \in \mathbb{C}^{m \times m}$ denotes the circulant matrix corresponding to cyclic convolution with a length m zero padding of $\boldsymbol{x}$, and $\operatorname{poly}\log n$ denotes a polynomial in $\log n$. A dependence of the sample complexity m on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ seems inevitable[2] and is corroborated by experiments.

Our proof is based on ideas from *decoupling theory* [11], the *suprema of chaos processes* and *restricted isometry property* of random circulant matrices [26, 20], and a new iterative analysis of alternating minimizing methods [35]. Our analysis draws connections between the convergence properties of gradient descent and the classical alternating direction method. This allows us to avoid the need to argue that high-degree polynomials in the structured random matrix $\boldsymbol{A}$ concentrate uniformly, as would be required by a straightforward translation of existing analysis to this new setting. Instead, we control the bulk effect of phase errors uniformly in a neighborhood around the ground truth. This requires us to develop new decoupling and concentration tools for controlling nonlinear phase functions of circulant random matrices, which could be potentially useful for analyzing other random circulant convolution problems such as blind deconvolution [40], and convolutional dictionary learning [18].

Prior art for phase retrieval. The challenge of developing efficient, guaranteed methods for phase retrieval has attracted substantial interest over the past decade [28, 19]. For the generalized phase retrieval problem in which the sensing matrix $\boldsymbol{A}$ is i.i.d. random, the first result on global recovery is based on *semidefinite programming* (SDP) [8, 3, 36]. However, the computational cost of SDP limits its practicality. Nonconvex methods can be more efficient. [25] showed that the alternating minimization method provably converges to the

[1] It would be nicer to characterize the global geometry of the problem as in [15, 33, 34, 32]. However, the nonhomogeneity of $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ over the space causes tremendous difficulties for concentration with $m \geq \Omega(n \operatorname{poly}\log n)$ samples.

[2] The operator norm of $\boldsymbol{C}_{\boldsymbol{x}}$ is nonhomogeneous over $\boldsymbol{x} \in \mathbb{CS}^{n-1}$, ranging from constant to $\mathcal{O}(\sqrt{n})$. For instance, $\|\boldsymbol{C}_{\boldsymbol{x}}\| = 1$ when $\boldsymbol{x}$ is a standard basis vector; and $\|\boldsymbol{C}_{\boldsymbol{x}}\| = \sqrt{n}$ when $\boldsymbol{x} = \frac{1}{\sqrt{n}}\boldsymbol{1}$.

truth, when initialized using a spectral method and provided with fresh samples at each iteration. Candes et al. [5] showed with the same initalization, gradient descent for the nonconvex least squares objective,

$$\min_{\boldsymbol{z}\in\mathbb{C}^n} f_1(\boldsymbol{z}) = \frac{1}{2m}\left\|\boldsymbol{y}^2 - |\boldsymbol{A}\boldsymbol{z}|^2\right\|^2, \tag{5}$$

provably recovers the ground truth, with near-optimal sample complexity $m \geq \Omega(n\log n)$. The work [10, 39, 38] further reduce the sample complexity to $m \geq \Omega(n)$ by using different nonconvex objectives and truncation techniques. Moreover, [34] reveals that the nonconvex objective (5) has a benign *global geometry*: with high probability, it has no bad critical points with $m \geq \Omega(n\log^3 n)$ samples[3].

Structured random measurements. The study of structured random measurement in signal processing [21] includes the study of random Fourier measurements [7, 9, 12] and partial random convolutions [26, 20] in compressed sensing [6]. However, the study of structured random measurement for phase retrieval is still quite limited. In particular, [17] and [4] studied the performance of SDP methods with t-designs and random masked Fourier transform measurements. The authors in [5, 2] show that the phase retrieval problem with random coded diffraction and STFT measurements can be solved by minimizing nonconvex objectives, while [5] requires resampling for the initialization, and in [2] the contraction radius is not large enough for initialization. In addition, the motivation of these measurement schemes are quite different from ours. For more detailed review of this subject, we refer the readers to Section 4 of [21].

Notations. We use $(\cdot)^\top$ and $(\cdot)^*$ to denote the real and Hermitian transpose, respectively. We use $\mathbb{CS}^{n-1}$ to denote a n dimensional complex sphere. Let $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and imaginary parts of a complex variable, respectively. Throughout the paper, we assume the optimal solution is $\boldsymbol{x} \in \mathbb{C}^n$. Because the solution is only optimal to a global phase shift, we define the optimal solution set as $\mathcal{X} = \left\{\boldsymbol{x}e^{\mathrm{i}\theta} \mid \theta \in [0, 2\pi)\right\}$, and define the distance from a point $\boldsymbol{z} \in \mathbb{C}^n$ to the set $\mathcal{X}$ as

$$\mathrm{dist}(\boldsymbol{z}, \mathcal{X}) \doteq \inf_{\theta\in[0,2\pi)} \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|.$$

For any $z \in \mathbb{C}$ with $|z| \neq 0$, we use $\phi(z)$ to denote the phase of z, that is, $e^{\mathrm{i}\phi(z)} = z/|z|$.

2 Algorithm

We develop an approach to convolutional phase retrieval based on local nonconvex optimization. Our proposed algorithm has two components: (1) a careful initialization using the spectral method; (2) local refinement by (generalized) gradient descent. We introduce the two steps in reverse order.

2.1 Minimization of a nonconvex and nonsmooth objective

We consider minimizing a weighted *nonconvex* and *nonsmooth* objective

$$f(\boldsymbol{z}) = \frac{1}{2m}\left\|\boldsymbol{b}^{1/2} \odot (\boldsymbol{y} - |\boldsymbol{A}\boldsymbol{z}|)\right\|^2. \tag{6}$$

The introduction of the positive weights $\boldsymbol{b}$ facilitates our analysis, by enabling us to compare certain functions of the dependent random matrix $\boldsymbol{A}$ to functions involving more independent random variables. We will substantiate this claim in the next section.

Although the function (4) is not complex-differentiable, if one identifies $\mathbb{C}^n$ with $\mathbb{R}^{2n}$ and treats $f(\boldsymbol{z})$ as a function in the real domain, f is still differentiable in the real sense. Thus, we adopt the *Wirtinger calculus* [22], which can be thought of as a clean way of organizing the real partial derivatives [29, 34].

[3] [30] tightened the sample complexity to $m \geq \Omega(n\log n)$ by using advanced probability tools.

On the other hand, it should also be noted that the absolute value $|\cdot|$ is nonsmooth at 0 and hence $f(\boldsymbol{z})$ is not differentiable everywhere even in the real sense. Similar to [38], for any complex number $u \in \mathbb{C}$, if we uniquely define its phase $\phi(u)$ at 0 by

$$\exp\left(\mathrm{i}\phi(u)\right) \doteq \begin{cases} u/|u| & \text{if } |u| \neq 0, \\ 1 & \text{otherwise,} \end{cases}$$

then the Wirtinger gradient of (4) can be uniquely determined as

$$\frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z}) = \frac{1}{m} \boldsymbol{A}^* \operatorname{diag}(\boldsymbol{b}) \left[\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y} \odot \exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z})\right)\right]. \tag{7}$$

Starting from some initialization $\boldsymbol{z}^{(0)}$, we minimize the objective (6) by gradient descent

$$\boldsymbol{z}^{(r+1)} = \boldsymbol{z}^{(r)} - \tau \frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z}^{(r)}), \tag{8}$$

where $\tau > 0$ is the stepsize. Indeed, $\frac{\partial}{\partial \boldsymbol{z}} f(\boldsymbol{z})$ can be interpreted as the gradient of $f(\boldsymbol{z})$ as in the real case; this method is also referred to as *amplitude flow* [38].

2.2 Initialization via spectral method

Similar to [25, 29], we compute the initialization $\boldsymbol{z}^{(0)}$ via a spectral method, detailed in [29, Algorithm 1]. More specifically, $\boldsymbol{z}^{(0)}$ is a scaled leading eigenvector of

$$\boldsymbol{Y} = \frac{1}{m} \sum_{k=1}^{m} y_k^2 \boldsymbol{a}_k \boldsymbol{a}_k^* = \frac{1}{m} \boldsymbol{A}^* \operatorname{diag}\left(\boldsymbol{y}^2\right) \boldsymbol{A}, \tag{9}$$

which is constructed from the knowledge of the sensing vectors and observations. The leading eigenvector of $\boldsymbol{Y}$ can be efficiently computed via the power method. Note that $\mathbb{E}[\boldsymbol{Y}] = \|\boldsymbol{x}\|^2 \boldsymbol{I} + \boldsymbol{x}\boldsymbol{x}^*$, so the leading eigenvector of $\mathbb{E}[\boldsymbol{Y}]$ is proportional to the optimal solution $\boldsymbol{x}$. Under the random convolutional model of $\boldsymbol{A}$, by using probability tools from [21], we show that $\boldsymbol{v}^* \boldsymbol{Y} \boldsymbol{v}$ concentrates to its expectation $\boldsymbol{v}^* \mathbb{E}[\boldsymbol{Y}] \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathbb{CS}^{n-1}$ whenever $m \geq \Omega(n \operatorname{poly} \log n)$, ensuring the initialization $\boldsymbol{z}^{(0)}$ close to the optimal set $\mathcal{X}$.

3 Main Result and Analysis

In this section, we describe our main theoretical result, which shows that with high probability, the algorithm described in the previous section succeeds.

Theorem 3.1 (Main Result) *Whenever $m \geq C_0 n \log^{31} n$, the spectral method [29, Algorithm 1] produces an initialization $\boldsymbol{z}^{(0)}$ that satisfies*

$$\operatorname{dist}\left(\boldsymbol{z}^{(0)}, \mathcal{X}\right) \leq c_0 \log^{-6} n \, \|\boldsymbol{x}\|$$

with probability at least $1 - c_1 m^{-c_2}$. Suppose $\boldsymbol{b} = \zeta_{\sigma^2}(\boldsymbol{y})$, where

$$\zeta_{\sigma^2}(t) = 1 - 2\pi\sigma^2 \xi_{\sigma^2}(t), \qquad \xi_{\sigma^2}(t) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad t > 0, \tag{10}$$

with $\sigma^2 > 1/2$. Starting from $\boldsymbol{z}^{(0)}$, with $\sigma^2 = 0.51$ and stepsize $\tau = 2.02$, whenever $m \geq C_1 \frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2} \max\left\{\log^{17} n, n \log^4 n\right\}$, with probability at least $1 - c_3 m^{-c_4}$ for all iterate $\boldsymbol{z}^{(r)} (r \geq 1)$ defined in (8)*, we have*

$$\operatorname{dist}\left(\boldsymbol{z}^{(r)}, \mathcal{X}\right) \leq (1 - \varrho)^r \operatorname{dist}\left(\boldsymbol{z}^{(0)}, \mathcal{X}\right), \tag{11}$$

holds for a small scalar $\varrho \in (0, 1)$. Here, $c_0, c_1, c_2, c_3, c_4, C_0, C_1 > 0$ are numerical constants.

Remark: Our result shows that by initializing the problem $\mathcal{O}(1/\mathrm{polylog}(n))$-close to the optimum via spectral method, the gradient descent (8) converges linearly to the optimal solution. As we can see, the sample complexity here also depends on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$, which is quite different from the i.i.d. case. For a typical $\boldsymbol{x} \in \mathbb{CS}^{n-1}$ (e.g., $\boldsymbol{x}$ is drawn uniformly random from $\mathbb{CS}^{n-1}$), $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ remains as $\mathcal{O}(\log n)$, the sample complexity $m \geq \Omega\left(n \operatorname{poly}\log n\right)$ matches the i.i.d. case up to log factors. However, $\|\boldsymbol{C}_{\boldsymbol{x}}\|$ is nonhomogeneous over $\boldsymbol{x} \in \mathbb{CS}^{n-1}$: if $\boldsymbol{x}$ is sparse in the Fourier domain (e.g., $\boldsymbol{x} = \frac{1}{\sqrt{n}}\mathbf{1}$), the sample complexity can be as large as $m \geq \Omega\left(n^2 \operatorname{poly}\log n\right)$. Such a behavior is also demonstrated in the experiments of Section 4. We believe the (very large!) number of logarithms in our result is an artifact of our analysis, rather than a limitation of the method. We expect to reduce the sample complexity to $m \geq \Omega\left(\frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|^2}{\|\boldsymbol{x}\|^2} n \log^6 n\right)$ by a tighter analysis, which is left for future work. The choices of the weighting $\boldsymbol{b} \in \mathbb{R}^m$ in (10), $\sigma^2 = 0.51$, and the stepsize $\tau = 2.02$ are purely for the purpose of analysis. In practice, the algorithm converges with $\boldsymbol{b} = \mathbf{1}$ and a choice of small stepsize τ, or by using backtracking linesearch for the stepsize τ.

In the following, we briefly highlight some major challenges and novel proofing ideas behind the analysis. The details can be found in our full paper.

3.1 Proof sketch of iterative contraction

Our analysis is largely inspired by the recent analysis of *alternating direction method* (ADM) [35]. In this following, we draw connections between the gradient descent method (8) and ADM, and sketch basic ideas of convergence analysis.

ADM iteration. ADM is a classical method for solving phase retrieval problems [16, 25, 35], which can be considered as a heuristic method that solves the problem

$$\min_{\boldsymbol{z}\in\mathbb{C}^n, |\boldsymbol{u}|=\mathbf{1}} \tfrac{1}{2}\left\|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{y}\odot\boldsymbol{u}\right\|^2.$$

At every iterate $\widehat{\boldsymbol{z}}^{(r)}$, ADM proceeds in two steps:

$$\boldsymbol{c}^{(r+1)} = \boldsymbol{y}\odot\exp\left(\boldsymbol{A}\widehat{\boldsymbol{z}}^{(r)}\right), \qquad \widehat{\boldsymbol{z}}^{(r+1)} = \arg\min_{\boldsymbol{z}} \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{z} - \boldsymbol{c}^{(r+1)}\right\|^2,$$

which leads to the following update

$$\widehat{\boldsymbol{z}}^{(r+1)} \;=\; \boldsymbol{A}^{\dagger}\left(\boldsymbol{y}\odot\exp\left(\boldsymbol{A}\widehat{\boldsymbol{z}}^{(r)}\right)\right),$$

where $\boldsymbol{A}^{\dagger} = \left(\boldsymbol{A}^*\boldsymbol{A}\right)^{-1}\boldsymbol{A}^*$ is the pseudo-inverse of $\boldsymbol{A}$. Let $\widehat{\theta}_r = \arg\min_{\overline{\theta}}\left\|\widehat{\boldsymbol{z}}^{(r)} - \boldsymbol{x}e^{\mathrm{i}\overline{\theta}}\right\|$. The distance between $\widehat{\boldsymbol{z}}^{(r+1)}$ and $\mathcal{X}$ is bounded by

$$\operatorname{dist}\left(\widehat{\boldsymbol{z}}^{(r+1)}, \mathcal{X}\right) = \left\|\widehat{\boldsymbol{z}}^{(r+1)} - \boldsymbol{x}e^{\mathrm{i}\widehat{\theta}_{r+1}}\right\| \leq \left\|\boldsymbol{A}^{\dagger}\right\| \left\|\boldsymbol{A}\boldsymbol{x}e^{\mathrm{i}\widehat{\theta}_r} - \left(\boldsymbol{y}\odot\exp\left(\boldsymbol{A}\widehat{\boldsymbol{z}}^{(r)}\right)\right)\right\|. \tag{12}$$

Gradient descent with $\boldsymbol{b} = \mathbf{1}$. For simplicity, let us consider the gradient descent update (8) with $\boldsymbol{b} = \mathbf{1}$. Let $\theta_r = \arg\min_{\overline{\theta}}\left\|\boldsymbol{z}^{(r)} - \boldsymbol{x}e^{\mathrm{i}\overline{\theta}}\right\|$, with stepsize $\tau = 1$. The distance between the iterate $\boldsymbol{z}^{(r)}$ and the optimal set $\mathcal{X}$ is bounded by

$$\begin{aligned}\operatorname{dist}\left(\boldsymbol{z}^{(r+1)}, \mathcal{X}\right) = \left\|\boldsymbol{z}^{(r+1)} - \boldsymbol{x}e^{\mathrm{i}\theta_{r+1}}\right\| \leq{}& \left\|\boldsymbol{I} - \frac{1}{m}\boldsymbol{A}^*\boldsymbol{A}\right\| \left\|\boldsymbol{z}^{(r)} - \boldsymbol{x}e^{\mathrm{i}\theta_r}\right\| \\ &+ \frac{1}{m}\left\|\boldsymbol{A}\right\| \left\|\boldsymbol{A}\boldsymbol{x}e^{\mathrm{i}\theta_r} - \boldsymbol{y}\odot\exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z}^{(r)})\right)\right\|. \end{aligned}\tag{13}$$

Towards iterative contraction. By measure concentration, it can be shown that

$$\left\|\boldsymbol{I} - \frac{1}{m}\boldsymbol{A}^*\boldsymbol{A}\right\| = o(1), \quad \left\|\boldsymbol{A}\right\| \approx \sqrt{m}, \quad \left\|\boldsymbol{A}^{\dagger}\right\| \approx 1/\sqrt{m}, \tag{14}$$

holds with high probability whenever $m \geq \Omega(n \operatorname{poly}\log n)$. Therefore, to show iterative contraction of both methods, based on (12) and (13), it is sufficient to show that

$$\left\|\boldsymbol{A}\boldsymbol{x}e^{\mathrm{i}\theta} - \boldsymbol{y} \odot \exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z})\right)\right\| \leq (1-\eta)\sqrt{m}\left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta}\right\|, \tag{15}$$

for some constant $\eta \in (0,1)$, where $\theta = \arg\min_{\overline{\theta}\in[0,2\pi)} \left\|\boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\overline{\theta}}\right\|$ such that $e^{\mathrm{i}\theta} = \boldsymbol{x}^*\boldsymbol{z}/\left|\boldsymbol{x}^*\boldsymbol{z}\right|$. By similar ideas of controlling (15) for the ADM method [35], this observation provides a new way of analyzing the gradient descent method. As an attempt to show (15) for the random circulant matrix $\boldsymbol{A}$, we invoke the following lemma, which controls the error in a first order approximation to $\exp(\mathrm{i}\phi(\cdot))$.

Lemma 3.2 (Lemma 3.2, [35]) *For any $\rho > 0$, and for any $z,\ z' \in \mathbb{C}$, we have*

$$\left|\exp\left(\mathrm{i}\phi(z'+z)\right) - \exp\left(\mathrm{i}\phi(z')\right)\right| \ \leq\ 2\mathbb{1}_{|z|\geq\rho|z'|} \ +\ (1-\rho)^{-1}\left|\Im\left(z/z'\right)\right|.$$

Let us decompose $\boldsymbol{z} = \alpha\boldsymbol{x} + \beta\boldsymbol{w}$, where $\boldsymbol{w} \in \mathbb{CS}^{n-1}$ with $\boldsymbol{w} \perp \boldsymbol{x}$, and $\alpha, \beta \in \mathbb{C}$. Note that $\phi(\alpha) = \theta$. Then by Lemma 3.2, for any $\rho \in (0,1)$, we have

$$\begin{aligned}\left\|\boldsymbol{A}\boldsymbol{x}e^{\mathrm{i}\theta} - \boldsymbol{y} \odot \exp\left(\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{z})\right)\right\| \quad &= \quad \left\|\left|\boldsymbol{A}\boldsymbol{x}\right| \odot \left[\exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x}\right)\right) - \exp\left(\mathrm{i}\phi\left(\boldsymbol{A}\boldsymbol{x} + \frac{\beta}{\alpha}\boldsymbol{A}\boldsymbol{w}\right)\right)\right]\right\| \\ \leq \quad \underbrace{\left\|\left|\boldsymbol{A}\boldsymbol{x}\right| \odot \mathbb{1}_{\left|\frac{\beta}{\alpha}\right|\left|\boldsymbol{A}\boldsymbol{w}\right| \geq \rho\left|\boldsymbol{A}\boldsymbol{x}\right|}\right\|}_{\mathcal{T}_1} \quad &+ \quad \frac{1}{1-\rho}\left|\frac{\beta}{\alpha}\right| \underbrace{\left\|\Im\left(\left(\boldsymbol{A}\boldsymbol{w}\right) \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right)\right\|}_{\mathcal{T}_2}.\end{aligned}$$

The first term $\mathcal{T}_1$ can be bounded using the *restricted isometry property* of random circulant matrices [20], together with some auxiliary analysis.

The second term $\mathcal{T}_2$ involves a nonlinear function $\exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)$ of the random circulant matrix $\boldsymbol{A}$. Controlling this nonlinear, highly dependent random process $\mathcal{T}_2(\boldsymbol{w})$ for all $\boldsymbol{w}$ is a nontrivial task. Next, we explain why controlling $\mathcal{T}_2$ is technically challenging, and sketch the key ideas about how to control a smoothed variant of $\mathcal{T}_2$, by using the weighting $\boldsymbol{b} = \zeta_{\sigma^2}(\boldsymbol{y})$ introduced in (10). We also provide intuition for why the weighting $\boldsymbol{b}$ is helpful.

3.2 Controlling the phase term $\mathcal{T}_2$

As elaborated above, the major challenge of showing iterative contraction is bounding the suprema of the nonlinear, dependent random process $\mathcal{T}_2(\boldsymbol{w})$ over the set $\mathcal{S} = \left\{\boldsymbol{w} \in \mathbb{CS}^{n-1} \mid \boldsymbol{w} \perp \boldsymbol{x}\right\}$. By using the fact that $\Im(u) = \frac{1}{2\mathrm{i}}\left(u - \overline{u}\right)$ for any $u \in \mathbb{C}$, we have

$$\sup_{\boldsymbol{w}\in\mathcal{S}} \mathcal{T}_2^2 \ \leq\ \tfrac{1}{2}\left\|\boldsymbol{A}\right\|^2 + \tfrac{1}{2}\sup_{\boldsymbol{w}\in\mathcal{S}}\left|\underbrace{\boldsymbol{w}^\top\boldsymbol{A}^\top \operatorname{diag}\left(\psi(\boldsymbol{A}\boldsymbol{x})\right)\boldsymbol{A}\boldsymbol{w}}_{\mathcal{L}(\boldsymbol{a},\boldsymbol{w})}\right|,$$

where $\psi(t) \doteq \exp\left(-2\mathrm{i}\phi(t)\right)$. As from (14), $\left\|\boldsymbol{A}\right\| \approx \sqrt{m}$, the major task left is to show that

$$\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\mathcal{L}(\boldsymbol{a},\boldsymbol{w})\right| < (1-\eta')m \tag{16}$$

for some constant $\eta' \in (0,1)$.

Why decoupling? Let $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{a}_1^* \\ \cdots \\ \boldsymbol{a}_m^*\end{bmatrix}$. The term

$$\mathcal{L}(\boldsymbol{a},\boldsymbol{w}) \ =\ \boldsymbol{w}^\top\boldsymbol{A}^\top \operatorname{diag}\left(\psi(\boldsymbol{A}\boldsymbol{x})\right)\boldsymbol{A}\boldsymbol{w} \ =\ \sum_{k=1}^{m} \underbrace{\psi(\boldsymbol{a}_k^*\boldsymbol{x})\boldsymbol{w}^\top\overline{\boldsymbol{a}}_k\overline{\boldsymbol{a}}_k^\top\boldsymbol{w}}_{\text{dependence across } k}$$

is a summation of dependent random variables, for which our probability tools are very limited. To overcome this problem, we deploy ideas from *decoupling* [11]. Informally, decoupling allows us to compare moments of the original random function to functions of more independent random variables, which are usually easier to analyze. The book [11] provides

a beautiful introduction to this area. In our problem, notice that the random vector $\boldsymbol{a}$ occurs twice in the definition of $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ – one in the phase term $\psi(\boldsymbol{A}\boldsymbol{x}) = \exp(-2\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x}))$, and another in the quadratic term. The general spirit of decoupling is to seek to replace one $\boldsymbol{a}$ with an *independent* copy $\boldsymbol{a}'$ of the same random vector, yielding a random process with fewer dependencies. Here, we seek to replace $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ with

$$\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{A}^\top \operatorname{diag}\left(\psi(\boldsymbol{A}'\boldsymbol{x})\right) \boldsymbol{A}\boldsymbol{w}. \tag{17}$$

The usefulness of this new, decoupled form $\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$, is that it introduces extra randomness — $\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$ is now a *chaos* process of $\boldsymbol{a}$ conditioned on $\boldsymbol{a}'$. This makes analyzing $\sup_{\boldsymbol{w}\in\mathcal{S}} \mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$ amenable to existing analysis of *suprema of chaos processes* for random circulant matrices [21]. However, achieving the decoupling requires additional work; the most general existing results on decoupling pertain to *tetrahedral polynomials*, which are polynomials with no monomials involving any power larger than one of any random variable. By appropriately tracking cross terms, these results can also be applied to more general (non-tetrahedral) polynomials in Gaussian random variables [23]. However, our random process $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ involves a nonlinear phase term $\psi(\boldsymbol{A}\boldsymbol{w})$ which is not a polynomial, and hence is not amenable to a direct appeal to existing results.

Decoupling is "recoupling". Existing results [23] for decoupling polynomials of Gaussian random variables are derived from two simple facts: (i) orthogonal projections of Gaussian variables are independent, and (ii) Jensen's inequality. Indeed, for $\boldsymbol{a} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$, let us introduce an independent vector $\boldsymbol{\delta} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I})$. Write

$$\boldsymbol{g}^1 = \boldsymbol{a} + \boldsymbol{\delta}, \qquad \boldsymbol{g}^2 = \boldsymbol{a} - \boldsymbol{\delta}.$$

Because of Fact (i), these are *independent* $\mathcal{CN}(\boldsymbol{0}, 2\boldsymbol{I})$ vectors. By conditional expectation,

$$\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2, \boldsymbol{w})\right] \;=\; \mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}+\boldsymbol{\delta}, \boldsymbol{a}-\boldsymbol{\delta}, \boldsymbol{w})\right] \;\doteq\; \widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w}). \tag{18}$$

Thus, we can see that the key idea of decoupling $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ into $\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$, is essentially "recoupling" $\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2, \boldsymbol{w})$ via conditional expectation – the "recoupled" term $\widehat{\mathcal{L}}$ can be viewed as an approximation of $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$. Notice that by Fact (ii), for any convex function φ,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{a}}\left[\sup_{\boldsymbol{w}\in\mathcal{S}} \varphi\left(\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right)\right] \;&=\; \mathbb{E}_{\boldsymbol{a}}\left[\sup_{\boldsymbol{w}\in\mathcal{S}} \varphi\left(\mathbb{E}_{\boldsymbol{\delta}}\left[\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}+\boldsymbol{\delta}, \boldsymbol{a}-\boldsymbol{\delta}, \boldsymbol{w})\right]\right)\right] \\ &\leq\; \mathbb{E}_{\boldsymbol{a},\boldsymbol{\delta}}\left[\sup_{\boldsymbol{w}\in\mathcal{S}} \varphi\left(\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}+\boldsymbol{\delta}, \boldsymbol{a}-\boldsymbol{\delta}, \boldsymbol{w})\right)\right] \;=\; \mathbb{E}_{\boldsymbol{g}^1,\boldsymbol{g}^2}\left[\sup_{\boldsymbol{w}\in\mathcal{S}} \varphi\left(\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2, \boldsymbol{w})\right)\right].\end{aligned}$$

Thus, by choosing $\varphi(t) = |t|^p$, we can control moments of $\sup_{\boldsymbol{w}\in\mathcal{S}} \widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})$ via

$$\left\|\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right|\right\|_{L^p} \;\leq\; \left\|\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{g}^1, \boldsymbol{g}^2, \boldsymbol{w})\right|\right\|_{L^p}. \tag{19}$$

For tetrahedral polynomials, $\widehat{\mathcal{L}} = \mathcal{L}$, so the approximation is exact. As the tail bound of $\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right|$ can be controlled via its moments bounds [13, Chapter 7.2], this allows us to directly control the object $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ of interest. The reason that this control obtains is because the conditional expectation operator $\mathbb{E}_{\boldsymbol{\delta}}[\cdot \mid \boldsymbol{a}]$ "recouples" $\mathcal{Q}^{\mathcal{L}}_{dec}(\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{w})$ back to the target $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$. In slogan form, (Gaussian) *decoupling is recoupling*.

"Recoupling" is Gaussian smoothing. A distinctive feature in convolutional phase retrieval is that $\mathcal{L}$ is *not* a polynomial. Hence, it may be challenging to posit a $\mathcal{Q}^{\mathcal{L}}_{dec}$ which "recouples" back to $\mathcal{L}$. In other words, in the existing form, we need to tolerate an *approximation error* as $\widehat{\mathcal{L}} \neq \mathcal{L}$. By the triangle inequality,

$$\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})| \;\leq\; \sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right| + \sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w}) - \mathcal{L}(\boldsymbol{a}, \boldsymbol{w})\right|. \tag{20}$$

As discussed above, the $\sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right|$ can be sharply controlled via its moments bound in (19). Now the bound (20) is useful to derive tight control for $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$, if $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ is very

close to $\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})$ uniformly. The question is: *for what $\mathcal{L}$ is it possible to find a "well-behaved" $\mathcal{Q}_{dec}^{\mathcal{L}}$ for which the approximation error is small?* To understand this question, recall that the mechanism that links $\mathcal{Q}_{dec}$ back to $\widehat{\mathcal{L}}$ is the conditional expectation operator $\mathbb{E}_{\boldsymbol{\delta}}[\cdot \mid \boldsymbol{a}]$. For our case, from (18) orthogonality leads to

$$\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w}) = \boldsymbol{w}\boldsymbol{A}^{\top} \operatorname{diag}\left(h(\boldsymbol{A}\boldsymbol{x})\right) \boldsymbol{A}\boldsymbol{w}, \qquad h(t) \doteq \mathbb{E}_{s\sim\mathcal{CN}(0,\|\boldsymbol{x}\|^2)}\left[\psi(t+s)\right]. \tag{21}$$

Thus, by combining the results in (20) and (21), we have

$$\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})| \;\leq\; \sup_{\boldsymbol{w}\in\mathcal{S}} \left|\widehat{\mathcal{L}}(\boldsymbol{a}, \boldsymbol{w})\right| \;+\; \underbrace{\|h-\psi\|_{L^\infty}}_{\text{approximation error}} \|\boldsymbol{A}\|^2. \tag{22}$$

Note that the function h is not exactly ψ, but generated by convolving ψ with a multivariate Gaussian *pdf*: indeed, *recoupling is Gaussian smoothing*. The Fourier transform of a multivariate Gaussian is again a Gaussian; it decays quickly with frequency. So, in order to admit a small approximation error, the target $\mathcal{L}$ must be *smooth*. However, in our case, the function $\psi(t) = \exp(-2\mathrm{i}\phi(t))$ is discontinuous at $t = 0$; it changes extremely rapidly in the vicinity of $t = 0$, and hence its Fourier transform (appropriately defined) does not decay quickly at all. Therefore, $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$ is a poor target for approximation with a smooth function $\widehat{\mathcal{L}} = \mathbb{E}_{\boldsymbol{\delta}}[\mathcal{Q}_{dec}^{\mathcal{L}}]$. From Fig. 1, the difference between h and ψ increases as $|t| \searrow 0$. The poor approximation error $\|h-\psi\|_{L^\infty} = 1$ results in a trivial bound for $\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})|$ instead of (16).

Decoupling and convolutional phase retrieval. The key idea to reduce the approximation error $\|\psi - h\|_{L^\infty} = 1$ is to smooth ψ. More specifically, we introduce a new objective (6) with Gaussian weighting $\boldsymbol{b} = \zeta_{\sigma^2}(\boldsymbol{y})$ in (10), replacing the analyzing target $\mathcal{T}_2$ with

$$\widehat{\mathcal{T}}_2 = \left\| \operatorname{diag}\left(\boldsymbol{b}^{1/2}\right) \Im\left((\boldsymbol{A}\boldsymbol{w}) \odot \exp\left(-\mathrm{i}\phi(\boldsymbol{A}\boldsymbol{x})\right)\right) \right\|.$$

Consequently, we obtain a smoothed variant $\mathcal{L}_s(\boldsymbol{a}, \boldsymbol{w})$ of $\mathcal{L}(\boldsymbol{a}, \boldsymbol{w})$,

$$\mathcal{L}_s(\boldsymbol{a}, \boldsymbol{w}) = \boldsymbol{w}^{\top}\boldsymbol{A}^{\top} \operatorname{diag}\left(\zeta_{\sigma^2}(\boldsymbol{y}) \odot \psi(\boldsymbol{A}\boldsymbol{x})\right) \boldsymbol{A}\boldsymbol{w}.$$

Now the approximation error $\|h-\psi\|_{L^\infty}$ in (22) is replaced by $\|h(t) - \zeta_{\sigma^2}(t)\psi(t)\|_{L^\infty}$. As observed from Fig. 1, the function $\zeta_{\sigma^2}(t)$ smoothes $\psi(t)$ especially near the vicinity of $t = 0$, such that the new approximation error $\|f(t) - \zeta_{\sigma^2}(t)\psi(t)\|_{L^\infty}$ is significantly reduced. Thus, by using similar ideas as above, we can prove a desired bound $\sup_{\boldsymbol{w}\in\mathcal{S}} |\mathcal{L}_s(\boldsymbol{a}, \boldsymbol{w})| < (1-\eta_s)m$. Finally, because the new weighting $\boldsymbol{b} = \zeta_{\sigma^2}(\boldsymbol{y})$, the overall analysis needs to be slightly modified correspondingly. We refer the readers to our full paper for more details.

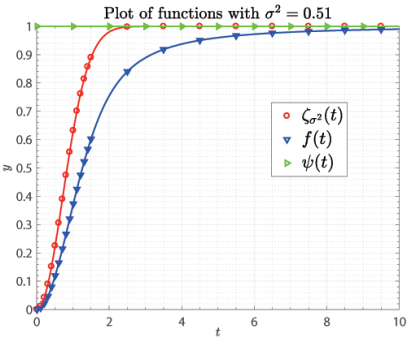

Figure 1: Plots of functions $\zeta_{\sigma^2}(t)$, $f(t)$ and $\psi(t)$ for $t \in \mathbb{R}_+$.

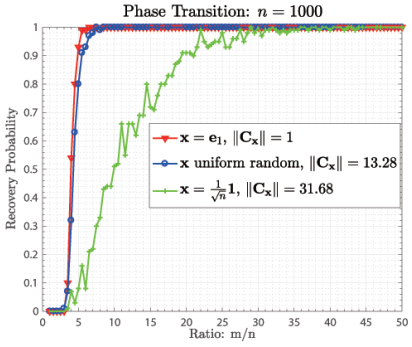

Figure 2: Phase transition for recovering the signal $x \in \mathbb{CS}^{n-1}$ with different $\|\boldsymbol{C}_{\boldsymbol{x}}\|$.

4 Experiments

Dependence of sample complexity on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$. First, we investigate the dependence of the sample complexity m on $\|\boldsymbol{C}_{\boldsymbol{x}}\|$. We assume the ground truth $\boldsymbol{x} \in \mathbb{CS}^{n-1}$, and consider three

cases: (1) $\boldsymbol{x} = \boldsymbol{e}_1$ with $\|\boldsymbol{C}_{\boldsymbol{x}}\| = 1$, where $\boldsymbol{e}_1$ the standard basis vector; (2) $\boldsymbol{x}$ is uniformly random generated from $\mathbb{CS}^{n-1}$; (3) $\boldsymbol{x} = \frac{1}{\sqrt{n}}\mathbf{1}$, with $\|\boldsymbol{C}_{\boldsymbol{x}}\| = \sqrt{n}$. For each case, we fix the signal length $n = 1000$ and vary the ratio m/n. For each ratio m/n, we randomly generate the kernel $\boldsymbol{a} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{I})$ and repeat the experiment for 100 times. We initialize the algorithm by the spectral method [29, Algorithm 1] and run the gradient descent (8). Given the algorithm output $\widehat{\boldsymbol{x}}$, we judge the success of recovery by $\inf_{\phi\in[0,2\pi)} \left\|\widehat{\boldsymbol{x}} - \boldsymbol{x}e^{\mathrm{i}\phi}\right\| \leq \epsilon$, where $\epsilon = 10^{-5}$. From Fig. 2, we can see that the larger the $\|\boldsymbol{C}_{\boldsymbol{x}}\|$, the more samples are needed for exact recovery.

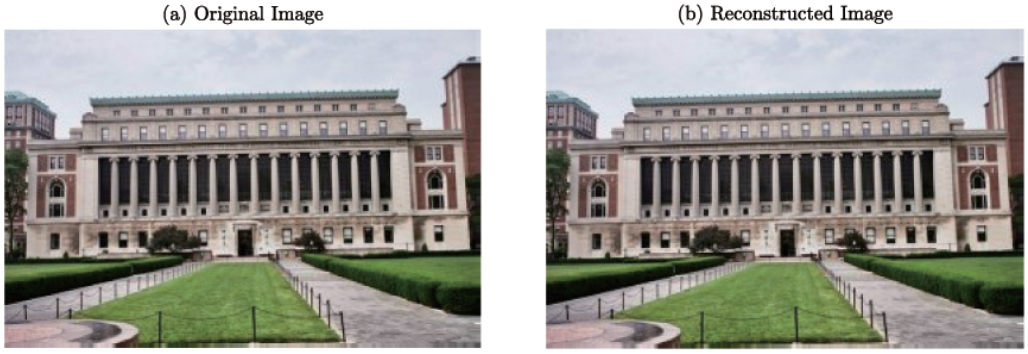

Figure 3: Experiment on real images.

Experiments on real image. Finally, we run the experiment on some real dataset to demonstrate the effectiveness and the efficiency of the proposed method. We choose an image of size 200×300 as in Fig. 4, we use $m = 5n \log n$ samples for reconstruction. The kernel $\boldsymbol{a} \in \mathbb{C}^m$ is randomly generated as complex Gaussian $\mathcal{CN}(\mathbf{0}, \boldsymbol{I})$. We run power method for 100 iterations for initialization, and stop the algorithm once the error is smaller than 1×10^{-4}. It takes $197.08s$ to reconstruct all the RGB channels. Experiment using general Gaussian measurements $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ could easily run out of memory on a personal computer for problems of this size.

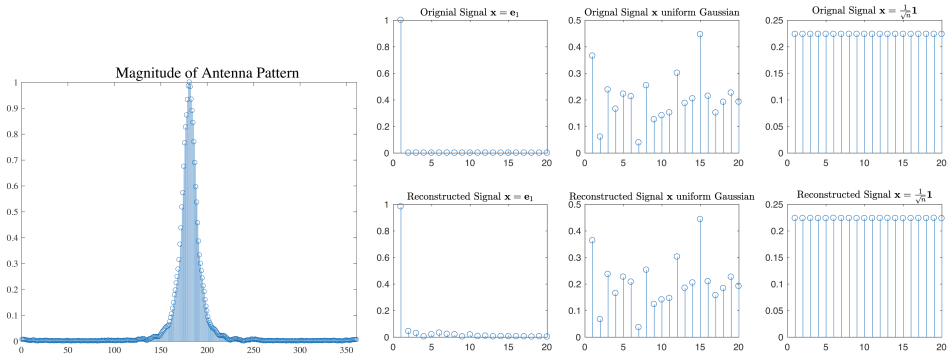

Figure 4: Experiment with real antenna pattern.

Experiments on signal AoA phase recovery for 5G communications. Finally, we demonstrate the effectiveness of the proposed method on a problem arising in 5G communication, as we mentioned in the introduction. Fig. 4 (left) shows an antenna pattern $\boldsymbol{a} \in \mathbb{C}^{361}$ obtained from Bell labs. We observe the modulus of the convolution of this pattern with the signal of interest. For three different types of signals with length $n = 20$, (1) $\boldsymbol{x} = \boldsymbol{e}_1$, (2) $\boldsymbol{x}$ is uniformly random generated from $\mathbb{CS}^{n-1}$, (3) $\boldsymbol{x} = \frac{1}{\sqrt{n}}\mathbf{1}$, our result in Fig. 4 shows that we can achieve almost perfect recovery.

5 Acknowledgement

This work was partially supported by the grants NSF CCF 1527809 and NSF IIS 1546411, the grants from the European Unions Horizon 2020 research and innovation program under grant agreement No. 646804-ERCCOGBNYQ, and the grant from the Israel Science Foundation under grant no. 335/14. QQ thanks the generous support of the Microsoft graduate research fellowship. We would like to thank Shan Zhong for the helpful discussion for real applications and providing the antenna data for experiments, and we thank Ju Sun and Han-wen Kuo for helpful discussion and input regarding the analysis of this work.

References

[1] Sercan Ö. Arik and Joseph M. Kahn. Direct-detection mode-division multiplexing in modal basis using phase retrieval. *Opt. Lett.*, 41(18):4265–4268, Sep 2016.

[2] T. Bendory, Y. C. Eldar, and N. Boumal. Non-convex phase retrieval from stft measurements. *IEEE Transactions on Information Theory*, PP(99):1–1, 2017.

[3] Emmanuel J. Candès, Yonina C. Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1), 2013.

[4] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2015.

[5] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, April 2015.

[6] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[7] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[8] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[9] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[10] Yuxin Chen and Emmanuel J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *arXiv preprint arXiv:1505.05114*, 2015.

[11] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer, 1999.

[12] Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Springer, 2013.

[14] Robert M Gagliardi and Sherman Karp. Optical communications. *New York, Wiley-Interscience, 1976. 445 p.*, 1, 1976.

[15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[16] R. W. Gerchberg and W. Owen Saxton. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[17] David Gross, Felix Krahmer, and Richard Kueng. A partial derandomization of phaselift using spherical designs. *arXiv preprint arXiv:1310.2267*, 2013.

[18] Felix Heide, Wolfgang Heidrich, and Gordon Wetzstein. Fast and flexible convolutional sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5135–5143, 2015.

[19] Kishore Jaganathan, Yonina C. Eldar, and Babak Hassibi. Phase retrieval: An overview of recent developments. *Chapter V, Optical Compressive Imaging*, 2016.

[20] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

[21] Felix Krahmer and Holger Rauhut. Structured random measurements in signal processing. *GAMM-Mitteilungen*, 37(2):217–238, 2014.

[22] Ken Kreutz-Delgado. The complex gradient operator and the $\mathbb{CR}$-calculus. *arXiv preprint arXiv:0906.4835*, 2009.

[23] Stanislaw Kwapien. Decoupling inequalities for polynomial chaos. *The Annals of Probability*, pages 1062–1071, 1987.

[24] Antonio Mecozzi, Cristian Antonelli, and Mark Shtaif. Kramers–kronig coherent receiver. *Optica*, 3(11):1220–1227, Nov 2016.

[25] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

[26] Holger Rauhut. Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9:1–92, 2010.

[27] Arash Shahmansoori, Gabriel E Garcia, Giuseppe Destino, Gonzalo Seco-Granados, and Henk Wymeersch. 5g position and orientation estimation through millimeter wave mimo. In *Globecom Workshops (GC Wkshps), 2015 IEEE*, pages 1–6. IEEE, 2015.

[28] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *Signal Processing Magazine, IEEE*, 32(3):87–109, May 2015.

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

[30] Mahdi Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *CoRR*, abs/1702.06175, 2017.

[31] Milica Stojanovic, Josko A Catipovic, and John G Proakis. Phase-coherent digital communications for underwater acoustic channels. *IEEE Journal of Oceanic Engineering*, 19(1):100–111, 1994.

[32] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*, 2015.

[33] Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

[34] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retreival. *arXiv preprint arXiv:1602.06664*, 2016.

[35] Irène Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *arXiv preprint arXiv:1609.03088*, 2016.

[36] Irène Waldspurger, Alexandre d`Aspremont, and Stéphane Mallat. Phase recovery, maxcut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[37] P. Walk, H. Becker, and P. Jung. OFDM channel estimation via phase retrieval. In *Asilomar 2015*, 2015.

[38] G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, PP(99):1–1, 2017.

[39] Huishuai Zhang and Yingbin Liang. Reshaped wirtinger flow for solving quadratic systems of equations. *arXiv preprint arXiv:1605.07719*, 2016.

[40] Yuqian Zhang, Yenson Lau, Han-wen Kuo, Sky Cheung, Abhay Pasupathy, and John Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.