A Geometric Analysis of Phase Retrieval

Ju Sun, Qing Qu, John Wright

{js4038, qq2105, jw2966}@columbia.edu

Dept. of Electrical Engineering, Columbia University, New York NY 10027, USA

Abstract—**Given measurements of the form $y_k = |\langle a_k, x \rangle|$ for $k = 1, \ldots, m$, is it possible to recover $x \in \mathbb{C}^n$? This is the generalized phase retrieval (GPR) problem which is a fundamental task in various disciplines. Natural nonconvex heuristics often work remarkably well for GPR in practice, but lack clear theoretical explanations. In this paper, we take a step towards bridging this gap. We show that when the sensing vectors a_k's are generic (i.i.d. complex Gaussian) and the number of measurements is large enough ($m \geq Cn \log^3 n$), with high probability, a natural least-squares formulation for GPR has the following benign geometric structure: (1) all global minimizers are the target signal x and its equivalent copies; and (2) the objective function has a negative curvature around each saddle point. Such structure allows a number of algorithmic possibilities for efficient global optimization. We describe a second-order trust-region algorithm that provably finds a global minimizer in polynomial time, from an arbitrary initialization.**

I. Introduction

Generalized phase retrieval (GPR) concerns the recovery of an unknown complex signal $x \in \mathbb{C}^n$ from a set of m measurements: $y_k = |\langle a_k, x \rangle|$ for $k = 1, \ldots, m$. This problem, and particularly its specialized version where a_k's are the Fourier basis vectors, arise in various applied fields, such as electron microscopy [1], astronomy [2], and optical imaging [3]. In these applications, one wants to recover a complex signal from its Fourier magnitudes, as phase information is missing due to physical constraints.

For decades, PR has seen a host of nonconvex formulations and heuristic algorithms that often work surprisingly well in practice. In contrast, investigation into provable recovery algorithms has started only relatively recently. We refer the readers to [4, 5] for an updated account of PR.

In this paper, we show that one can provably solve the GPR problem when the measurement vectors a_k's are *numerous* and *generic* enough. In particular, we focus on a natural least-square formulation [4, 5] studied theoretically in [6],

$$\min_{z \in \mathbb{C}^n} f(z) \doteq \frac{1}{2m} \sum_{k=1}^{m} \left(y_k^2 - |\langle a_k, z \rangle|^2 \right)^2. \quad (1)$$

We assume the a_k's are independent complex Gaussians:

$$a_k = (X_k + iY_k)/\sqrt{2}, \; X_k, Y_k \sim \mathcal{N}(0, I_n). \quad (2)$$

$f(z)$ is a 4-th order polynomial in z, and is nonconvex. Existing theoretical results on this and related formulations require careful initialization in the vicinity of the ground truth [6]. A-priori, there is little reason to believe that simple iterative methods can solve this problem without special initialization.
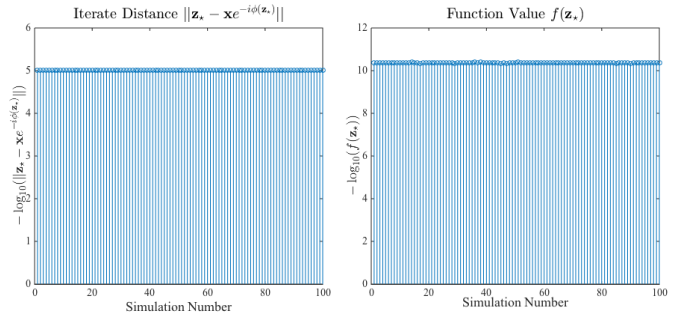

Fig. 1. Gradient descent with random initializations seems to always return a global solution for (1)! Here $n = 100$, $m = 5n \log n$, step size $\mu = 0.05$, and stopping criterion is $\|\nabla_z f(z)\| \leq 10^{-5}$. For the same set of random measurements, the experiments are repeated for 100 times with different random initializations. (Left) Final distance to the target; (Right) Final function value (0 if globally optimized). Both vertical axes are on $-\log_{10}(\cdot)$ scale.

A. A Curious Experiment

We apply gradient descent to $f(z)$, starting from a *random initialization* $z^{(0)}$:

$$z^{(r+1)} = z^{(r)} - \mu \nabla_z f(z^{(r)}),$$

where the step size μ is fixed for simplicity. The result is quite striking (Fig. 1): for a fixed problem instance (fixed random measurements and fixed target x), gradient descent seems to always return a global minimizer, across many random initializations! This contrasts with the typical "mental picture" of nonconvex objectives as possessing many spurious local minimizers.

B. A Geometric Analysis

The numerical surprise described above is not a completely isolated observation – simple heuristic methods for GPR often work surprisingly well (see [4, 5] and references therein). In this paper, we take a step towards explaining this phenomenon. We show that *although the function (1) is nonconvex, it actually has benign global geometry which allows it to be globally optimized by efficient iterative methods, regardless of the initialization.*

Fig. 2 plots the function landscape of $f(z)$ for large m and $z \in \mathbb{R}^2$. It is remarkable that (i) the only local and global minimizers are exactly $\pm x$;[1] (ii) there are saddle points (and a local maximizer), but around them there is negative curvature in the $\pm x$ direction. Intuitively, any algorithm that can successfully escape this kind of saddle points (and local

---

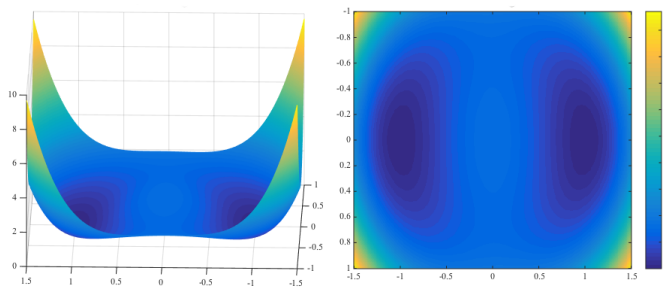[1] Note that the global sign cannot be recovered.

Fig. 2. Function landscape of (1) for $\boldsymbol{x} = [1; 0]$ and $m \to \infty$. The only local and global minimizers are $\pm \boldsymbol{x}$. There are two saddle points near $\pm[0; 1/\sqrt{2}]$, around each there is a negative curvature direction along $\pm \boldsymbol{x}$. (Left) The function graph; (Right) The same function visualized as a color image.

maximizers) can in fact find a global minimizer, i.e., recover the target signal.

We prove an analogous geometric structure exists, with high probability (w.h.p.)[2], for high-dimensional, complex signals, when m is reasonably large (Theorem II.2). Specifically, we show that when $m \geq Cn \log^3 n$, w.h.p., (i) the only local/global minimizers to (1) are the target $\boldsymbol{x}\mathrm{e}^{\mathrm{i}\phi}$ for all $\phi \in [0, 2\pi)$;[3] (ii) at any point in $\mathbb{C}^n$, either the gradient is large, or the curvature is negative in a certain direction (i.e., the Hessian has at least one negative eigenvalue), or it is near a minimizer such that the function is locally strongly convex in nontrivial directions (that $f(\boldsymbol{z}) = f(\boldsymbol{z}\mathrm{e}^{\mathrm{i}\phi})$ for all $\phi \in [0, 2\pi)$ induces a flat direction).

Because of this global geometry, a wide range of efficient iterative methods can obtain the global optimum, regardless of initialization. Examples include the noisy gradient and stochastic gradient methods [7] (see also [8]), curvilinear search [9] and trust region methods [10, 11]. The key property that the methods must possess is the ability to escape saddle points at which the Hessian has a strictly negative eigenvalue.

We corroborate this claim by developing a second-order trust-region method for this problem, and prove that (Theorem III.1) (i) from any initialization, it efficiently obtains a close approximation of the target vector $\boldsymbol{x}$ (up to a global phase) and (ii) it exhibits quadratic convergence in the vicinity of the global minimizer.

C. Prior Arts and Connections

Although heuristic methods for GPR have been used effectively in practice [4, 5, 12, 13], only recently have researchers begun to develop methods with provable performance guarantees. The first results of this nature were obtained using semidefinite programming relaxations [14–19]. While this represented a substantial advance in theory, the computational complexity limits the practicality of this approach.[4]

Recently, several provable *nonconvex* methods have been proposed for phase retrieval. [23] augmented the seminal error-reduction method [12] with spectral initialization and

resampling, and thereby provided the first provable nonconvex method for GPR. [6] studied the nonconvex formulation (1) under the same hypotheses as this paper, and showed that a spectral initialization plus local gradient descent recovers the true signal with near-optimal sample complexity. [24] worked with a different nonconvex formulation, and refined the spectral initialization and the local gradient descent with a step-adaptive truncation. With the modifications, they reduced the sample requirement to the optimal order.[5] All three analyses are local in nature, and depend on the spectral initializer being sufficiently close to the target signal.

In contrast, we explicitly characterize the global function landscape of (1). The geometric structure actually allows several algorithmic choices that need *no special initialization*. In fact, the spectral initialization used in [6] lands the iterate sequence in the restricted strongly convex regions in our results (Theorem II.2). The analysis of [6] is based on a property that ensures the gradient descent method is locally contractive near the target set, which is closely linked to local convexity. [25] and [26] explicitly established local strong convexity near the target set for GPR in $\mathbb{R}^n$.

The approach taken here is similar in spirit to our recent geometric analysis of a nonconvex formulation for complete dictionary learning [27]. For that problem, we also identified a similar geometric structure that allows efficient global optimization. [11] provides a high-level overview of this common structure (see also [7, 28].)

D. Definitions and Notations

We use $\|\cdot\|$ to denote the ℓ^2-norm for vectors, and the operator norm for matrices. We use $\boldsymbol{x} \in \mathbb{C}^n$ to denote the target vector. Since we only hope to recover it up to a global phase shift, let us denote the *target set* as $\mathcal{X} \doteq \left\{ \boldsymbol{x}\mathrm{e}^{\mathrm{i}\theta} : \theta \in [0, 2\pi) \right\}$.

For each complex vector $\boldsymbol{z} \in \mathbb{C}^n$, let $\Re(\boldsymbol{z})$ and $\Im(\boldsymbol{z})$ be the real and imaginary parts of $\boldsymbol{z}$, respectively. Define

$$\phi(\boldsymbol{z}) \doteq \arg\min_{\theta \in [0, 2\pi)} \left\| \boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\theta} \right\|, \quad (3)$$

$$\boldsymbol{h}(\boldsymbol{z}) \doteq \boldsymbol{z} - \boldsymbol{x}e^{\mathrm{i}\phi(\boldsymbol{z})}, \ \mathrm{dist}(\boldsymbol{z}, \mathcal{X}) \doteq \|\boldsymbol{h}(\boldsymbol{z})\|. \quad (4)$$

The function $f(\boldsymbol{z}) : \mathbb{C}^n \mapsto \mathbb{R}$ is not complex-differentiable. We adopt the Wirtinger derivatives [29], which can conveniently be considered as partial derivatives:

$$\nabla f = \left[\frac{\partial f}{\partial \boldsymbol{z}}, \frac{\partial f}{\partial \overline{\boldsymbol{z}}} \right]^*, \nabla^2 f = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{z}} \left(\frac{\partial f}{\partial \boldsymbol{z}} \right)^* & \frac{\partial}{\partial \overline{\boldsymbol{z}}} \left(\frac{\partial f}{\partial \boldsymbol{z}} \right)^* \\ \frac{\partial}{\partial \boldsymbol{z}} \left(\frac{\partial f}{\partial \overline{\boldsymbol{z}}} \right)^* & \frac{\partial}{\partial \overline{\boldsymbol{z}}} \left(\frac{\partial f}{\partial \overline{\boldsymbol{z}}} \right)^* \end{bmatrix},$$

where we have that *formally*

$$\frac{\partial f}{\partial \boldsymbol{z}} \doteq \frac{\partial f(z, \overline{z})}{\partial z} \big|_{\overline{z}=\text{constant}}, \frac{\partial f}{\partial \overline{\boldsymbol{z}}} \doteq \frac{\partial f(z, \overline{z})}{\partial \overline{z}} \big|_{z=\text{constant}}.$$

Let $\nabla_{\boldsymbol{z}} f$ and $\nabla_{\overline{\boldsymbol{z}}} f$ denote the first and second part of ∇f, respectively. The Wirtinger calculus is simple in expression

---

[2]The probability is with respect to drawing of $\boldsymbol{a}_k$'s.

[3]Note that global phase cannot be recovered.

[4]Another line of research [20–22] seeks to co-design the measurements and recovery algorithms based on frame- or graph-theoretic tools.

[5]In addition, [24] shows that the measurements can be non-adaptive, in the sense that a single, randomly chosen collection of vectors $\boldsymbol{a}_i$ can simultaneously recover every $\boldsymbol{x} \in \mathbb{C}^n$. Results in [6, 23] and this paper pertain only to adaptive measurements that recover any fixed signal $\boldsymbol{x}$ with high probability.

and analogous to those we obtain in the real case, we refer the readers to [29] for more details.

All technical proofs are deferred to the full version [30].

II. A Glimpse into High Dimensional Geometry

We have seen the low-dimensional plot of the function graph above. In this section, we present the quantitative results for high-dimensional, complex signals. We work out the "expected" version of the function landscape first, i.e., let $m \to \infty$.

Theorem II.1 (Characterization of $\mathbb{E}[f(z)]$). *When $x \neq 0$, for the asymptotic function $\mathbb{E}_a[f(z)]$, the only critical points are 0, $\mathcal{X}$ (the target set) and $\mathcal{C} \doteq \left\{ z \in \mathbb{C}^n : x^*z = 0, \|z\| = \|x\| / \sqrt{2} \right\}$, which are the local maximizer, the set of global optimizers, and the set of saddle points, respectively. Moreover, the saddle points and local maximizer have a negative curvature in the $xe^{i\phi(z)}$ direction.*

Basically, this says the geometric structure we observed about $f(z)$ for low-dimensional case qualitatively holds for high-dimensional signals also. Interestingly, we can show that the qualitative aspects of the geometric structure remain for high-dimensional, complex signals, even when the number of samples is *large yet finite*.
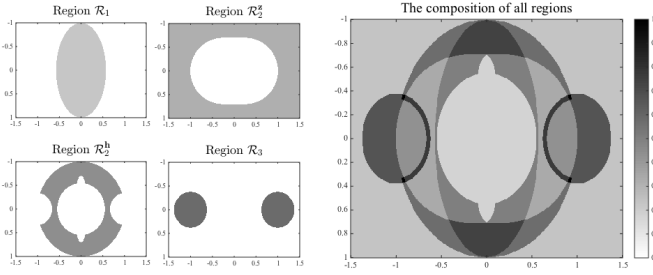

Fig. 3. Schematic illustration of partitioning regions for Theorem II.2. This plot corresponds to Fig. 2, i.e., the target signal is $x = [1; 0]$ and measurements are real Gaussians, such that the function is defined in $\mathbb{R}^2$.

Theorem II.2 (Characterization of $f(z)$). *There exist positive absolute constants c, C, C_1, C_2, C_3 and C_4, such that when $m \geq Cn \log^3 n$, it holds with probability at least $1 - cm^{-1}$ that the function $f(z)$ defined in (1) has no spurious local minimizers and the only local/global minimizers are exactly the target set $\mathcal{X}$. Quantitatively, for regions $\mathcal{R}_1$, $\mathcal{R}_2^z$, $\mathcal{R}_2^h$, and $\mathcal{R}_3$ that together cover $\mathbb{C}^n$,*

$$\begin{bmatrix} xe^{i\phi(z)} \\ \overline{xe^{i\phi(z)}} \end{bmatrix}^* \nabla^2 f(z) \begin{bmatrix} xe^{i\phi(z)} \\ \overline{xe^{i\phi(z)}} \end{bmatrix} \leq -C_1 \|x\|^4, \quad \forall z \in \mathcal{R}_1,$$

$$\frac{z^* \nabla_z f(z)}{\|z\|} \geq C_2 \|x\|^2 \|z\|, \qquad \forall z \in \mathcal{R}_2^z,$$

$$\frac{\Re(h(z)^* \nabla_z f(z))}{\|h(z)\|} \geq C_3 \|x\|^2 \|z\|, \qquad \forall z \in \mathcal{R}_2^h,$$

$$\begin{bmatrix} g(z) \\ \overline{g(z)} \end{bmatrix}^* \nabla^2 f(z) \begin{bmatrix} g(z) \\ \overline{g(z)} \end{bmatrix} \geq C_4 \|x\|^2, \qquad \forall z \in \mathcal{R}_3,$$

where $h(z)$ is defined in (4)*, and*

$$g(z) \doteq \begin{cases} h(z) / \|h(z)\| & \text{if } \operatorname{dist}(z, \mathcal{X}) \neq 0, \\ w \in \mathbb{S}^{n-1}, \Im(w^*z) = 0 & \text{if } z \in \mathcal{X}. \end{cases}$$

Here

$$\mathcal{R}_1 \doteq \left\{ z : 8 |x^*z|^2 + \frac{401}{100} \|x\|^2 \|z\|^2 \leq \frac{398}{100} \|x\|^4 \right\},$$

$$\mathcal{R}_2^z \doteq \left\{ z : \Re(\langle z, \nabla_z \mathbb{E}[f] \rangle) \geq \frac{1}{100} \|z\|^4 + \frac{1}{500} \|x\|^2 \|z\|^2 \right\},$$

$$\mathcal{R}_2^h \doteq \left\{ z : \frac{11}{20} \|x\| \leq \|z\| \leq \|x\|, \operatorname{dist}(z, \mathcal{X}) \geq \frac{\|x\|}{3}, \right.$$
$$\left. \Re(\langle h(z), \nabla_z \mathbb{E}[f] \rangle) \geq \frac{1}{250} \|x\|^2 \|z\| \|h(z)\| \right\},$$

$$\mathcal{R}_3 \doteq \left\{ z : \operatorname{dist}(z, \mathcal{X}) \leq \|x\| / \sqrt{7} \right\}.$$

Fig. 3 gives an idea on how the different regions cover the whole space.

Since (1) and associated derivatives take the form of summation of m independent random variables, the proof involves concentration and covering arguments [31]. The main challenge is that the function (1) is 4-th-order polynomial, and so the quantities of interest are heavy-tailed random variables. With only $O(n \log^3 n)$ samples, the gradients and Hessians of f do not concentrate uniformly about their expectations. Fortunately, this heavy-tailed behavior does not prevent the objective function from being globally well-structured for optimization. By carefully partitioning the "large gradient region" into two regions, we can show that the gradient is uniformly large using quantities that depend only on the *lower tails* of sums of fourth powers of independent (real) Gaussian random variables. These quantities *do* concentrate uniformly, because the fourth power of a Gaussian is nonnegative.

Theorem II.2 implies that the Wirtinger Hessian is indefinite in region $\mathcal{R}_1$, and the Wirtinger gradient is lower bounded in regions $\mathcal{R}_2^z$ and $\mathcal{R}_2^h$, so that the function value can always be decreased by following either $xe^{i\phi(z)}$ or the negative gradient directions $-\nabla f$. In region $\mathcal{R}_3$ which is around the global minimizers, although the function is flat on the complex circle $\left\{ ze^{i\theta} : \theta \in [0, 2\pi) \right\}$, it is strongly convex in any orthogonal direction $h(z)$ to the circle.

In sum, our objective $f(z)$ has the benign geometry that each $z \in \mathbb{C}^n$ has either large gradient or directional negative curvature, or lies in the vicinity of a local minimizer around which the function is locally restrictedly strongly convex. Functions with this property lies in the ridable-saddle function class [7, 11]. The particular geometric structure allows several optimization methods that can escape the saddle points to efficiently find a local minimizer[6].

---

[6]The definition of ridable-saddle class does not require all local minimizers to be global; here our objective has this additional good property that effectively allows us to do global optimization.

III. MODIFIED TRUST-REGION ALGORITHM

To demonstrate that the geometric structure we proved above is favorable for global optimization, we describe a second-order trust-region method (TRM) [10] that escapes saddle points, and provably finds one global minimizer of (1) in polynomial time, from an arbitrary initialization.

In each step, the method forms a quadratic approximation to the function $f(z)$ at the current iterate $z^{(r)}$ as

$$\widehat{f}(\delta; z^{(r)}) = f(z^{(r)}) + \begin{bmatrix} \delta \\ \overline{\delta} \end{bmatrix}^* \nabla f(z^{(r)})$$
$$+ \frac{1}{2} \begin{bmatrix} \delta \\ \overline{\delta} \end{bmatrix}^* \nabla^2 f(z^{(r)}) \begin{bmatrix} \delta \\ \overline{\delta} \end{bmatrix}.$$

When $z^{(r)}$ is a saddle point, the approximation reduces to

$$\widehat{f}(\delta; z^{(r)}) = f(z^{(r)}) + \frac{1}{2} \begin{bmatrix} \delta \\ \overline{\delta} \end{bmatrix}^* \nabla^2 f(z^{(r)}) \begin{bmatrix} \delta \\ \overline{\delta} \end{bmatrix}.$$

It is obvious that if $\nabla^2 f(z^{(r)})$ has a negative eigenvalue, which we proved above, the associated eigenvector is a local descent direction. This is the rationale for choosing second-order method.

Typical trust-region method goes by minimizing the approximation within a small ball, i.e., the trust region. Here we consider a modified version:

$$\min_{\delta \in \mathbb{C}^n} \widehat{f}(\delta, z^{(r)}), \text{ s.t. } \Im(\delta^* z^{(r)}) = 0, \|\delta\| \leq \Delta,$$

where Δ controls the trust region radius. Since the function is flat on the complex circle $\mathcal{C} = \{z^{(r)} e^{i\theta} : \theta \in [0, 2\pi)\}$, along $iz^{(r)}$ which is tangent to the $\mathcal{C}$, the reduction in function value is limited. The additional linear constraint forces the movement δ to be geometrically orthogonal to the trivial direction iz. This simple modification helps the algorithm to converge faster in practice, and guarantees quadratic convergence near the optima in theory.

To tackle the modified trust-region subproblem, note that the linear constraint defines a subspace $\mathcal{S}(z^{(r)}) \doteq \{w \in \mathbb{C}^n : \Im(w^* z^{(r)}) = 0\}$ of dimension $2n - 1$ over $\mathbb{R}^{2n}$. Take any matrix $U(z^{(r)}) \in \mathbb{C}^{n \times (2n-1)}$ whose columns form an orthonormal basis[7] for the subspace $\mathcal{S}(z^{(r)})$. Take $U^{(r)}$ for short of $U(z^{(r)})$ and let $\delta = U^{(r)} \xi$ with $\xi \in \mathbb{R}^{2n-1}$, then the subproblem can be reduced to the classical trust-region subproblem

$$\min_{\xi \in \mathbb{R}^{2n-1}} \widehat{f}(\xi; z^{(r)}), \quad \text{s.t. } \|\xi\| \leq \Delta, \quad (5)$$

where

$$\widehat{f}(\xi; z^{(r)}) = f(z^{(r)}) + \xi^\top g(z^{(r)}) + \frac{1}{2} \xi^\top H(z^{(r)}) \xi,$$

with

$$g = \begin{bmatrix} U^{(r)} \\ \overline{U^{(r)}} \end{bmatrix}^* \nabla f, \ H = \begin{bmatrix} U^{(r)} \\ \overline{U^{(r)}} \end{bmatrix}^* \nabla^2 f \begin{bmatrix} U^{(r)} \\ \overline{U^{(r)}} \end{bmatrix}.$$

---

[7]Here a matrix U is orthonormal means $\Re(U_i^* U_j) = \delta_{ij}$ for any columns U_i and U_j of U, where δ_{ij} denotes the Kronecker delta function.

The trust-region subproblem (5) can be efficiently solved by rooting finding [32] or SDP relaxation [33]. Once the solution $\xi_\star$ to the subproblem (5) is obtained, the iterate is updated by $z^{(r+1)} \leftarrow z^{(r)} + \delta_\star$ with $\delta_\star = U\xi_\star$. The choice of trust region size Δ is important both for the convergence theory and practical effectiveness. Following standard recommendations, we use a backtracking approach which modifies Δ from iteration to iteration based on the accuracy of the quadratic approximation of $f(z)$.

For analysis, we fix the trust region size Δ and prove the following convergence result.

Theorem III.1 (TRM Convergence). *Suppose $m \geq Cn \log^3 n$ for a sufficiently large constant C. Then with probability at least $1 - c_a m^{-1}$, the TRM algorithm with an arbitrary initialization $z^{(0)} \in \mathbb{CB}(R_0)$, where $R_0 \doteq 3(\frac{1}{m} \sum_{k=1}^m y_k^2)^{1/2}$, will return a solution that is ε-close to the optimal set $\mathcal{X}$ in*

$$\frac{c_b}{\Delta^2 \|x\|^2} f(z^{(0)}) + \log\log\left(\frac{c_c \|x\|}{\varepsilon}\right)$$

steps, provided that

$$\Delta \leq c_d (n^{7/2} \log^{7/2} m)^{-1} \|x\|.$$

Here c_a through c_d are positive absolute constants.

The proof is straightforward.

- When either the curvature is negative or the gradient is strong (i.e., region $\mathcal{R}_1, \mathcal{R}_2^z$ and $\mathcal{R}_2^h$), one step reduces the function value by a concrete amount.
- Under mild conditions, the iterates ultimately move into the restricted strong convex region $\mathcal{R}_3$ around the optimal.
- In $\mathcal{R}_3$, the TRM algorithm behaves like a typical second-order method on strongly convex functions with quadratic sequence convergence.

Remark. Our analysis here is very conservative, Theorem III.1 suggests that, for fixed trust region size Δ, at least $\Omega(n^7 \log^7 m)$ iterates are required for convergence. While in practice, TRM algorithm converges in $30 - 50$ iterations for our simulations with the backtracking strategy.

IV. EXPERIMENTAL RESULTS

To corroborate our theory, we run our trust-region methods alongside gradient descent, on simulated data.

We set $m : 1 \sim 500$ and $n : 1 \sim 100$. For each individual pair of (m, n), the sensing vectors $\{a_k\}_{k=1}^m \in \mathbb{C}^n$ are generated as i.i.d. complex Gaussian $\mathcal{CN}(n)$, and ground truth $x \in \mathbb{C}^n$ is uniformly drawn from the complex unit sphere $\mathbb{CB}(1)$. All simulations are repeated independently for 10 times. Both the gradient descent and the trust-region solvers are randomly initialized for each simulation. Since we only hope to recover x up to a global phase, we judge success by the following criteria

$$\left\| z_\star - x e^{i\phi(z_\star)} \right\| / \|x\| \leq \varepsilon,$$

where $z_\star$ is the algorithm output, and the error tolerance $\varepsilon = 10^{-4}$. Fig. 4 shows the phase transition in the (n, m)-plane. Our

observation suggests that, using random initializations, both gradient descent and trust-region methods can successfully recover the signal with around $m = 5n$ measurement with at least constant probability. It also suggests that our current analysis is still log factors away from the optimal. Finally, it is quite striking that randomly initialized gradient descent also succeeds here despite the presence of saddle points.
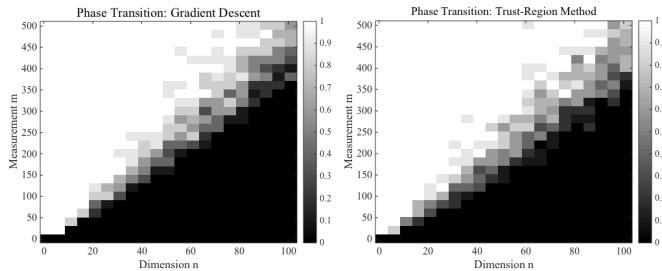

Fig. 4. Phase Transition for gradient descent (left) and trust region methods (right).

V. CONCLUSION AND DISCUSSION

In this paper, we explicitly characterized the global geometry of the phase retrieval problem. We conclude the paper by some open challenges for future work: (i) While our analysis is for complex Gaussian measurement, it is interesting to investigate the possible extension to other more practical sensing schemes, such as the masked Fourier measurements [17]. (ii) The sample complexity is suboptimal by a poly-logarithmic factor. (iii) Simulations in Section II and Section IV suggest that simple gradient descent with random initialization succeeds despite the presence of saddle points, how to understand such behaviors in theory remains an open problem. (iv) This work joins recent surge of theoretical understanding of nonconvex optimization (see [11] and the references therein), it is interesting to see whether our current geometric analysis can be generalized to other nonconvex problems.

REFERENCES

[1] J. Miao, T. Ishikawa, B. Johnson, E. H. Anderson, B. Lai, and K. O. Hodgson, "High resolution 3D X-Ray diffraction microscopy," *Phys. Rev. Lett.*, vol. 89, p. 088303, Aug 2002.
[2] C. Dainty and J. R. Fienup, "Phase retrieval and image reconstruction for astronomy," *Image Recovery: Theory and Application*, pp. 231–275, 1987.
[3] A. Walther, "The question of phase retrieval in optics," *Journal of Modern Optics*, vol. 10, no. 1, pp. 41–49, 1963.
[4] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: A contemporary overview," *Signal Processing Magazine, IEEE*, vol. 32, pp. 87–109, May 2015.
[5] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "Phase retrieval: An overview of recent developments," *arXiv preprint arXiv:1510.07713*, 2015.
[6] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *Information Theory, IEEE Transactions on*, vol. 61, pp. 1985–2007, April 2015.
[7] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Proceedings of The 28th Conference on Learning Theory*, pp. 797–842, 2015.
[8] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent converges to minimizers," *arXiv preprint arXiv:1602.04915*, 2016.
[9] D. Goldfarb, "Curvilinear path steplength algorithms for minimization which use directions of negative curvature," *Mathematical programming*, vol. 18, no. 1, pp. 31–40, 1980.
[10] A. R. Conn, N. I. Gould, and P. L. Toint, *Trust region methods*, vol. 1. SIAM, 2000.
[11] J. Sun, Q. Qu, and J. Wright, "When are nonconvex problems not scary?," *arXiv preprint arXiv:1510.06096*, 2015.
[12] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of the phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.
[13] J. R. Fienup, "Phase retrieval algorithms: a comparison," *Applied Optics*, vol. 21, pp. 2758–2769, Aug 1982.
[14] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, 2013.
[15] E. J. Candès, 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.
[16] E. J. Candès 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.
[17] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," *Applied and Computational Harmonic Analysis*, vol. 39, no. 2, pp. 277–299, 2015.
[18] 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.
[19] V. Voroninski and Z. Xu, "A strong restricted isometry property, with an application to phaseless compressed sensing," *arXiv preprint arXiv:1404.3811*, 2014.
[20] R. Balana, P. Casazzab, and D. Edidin, "On signal reconstruction without phase," *Applied and Computational Harmonic Analysis*, vol. 20, no. 3, pp. 345 – 356, 2006.
[21] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients," *Journal of Fourier Analysis and Applications*, vol. 15, no. 4, pp. 488–501, 2009.
[22] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, "Phase retrieval with polarization," *SIAM Journal on Imaging Sciences*, vol. 7, no. 1, pp. 35–66, 2014.
[23] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, pp. 2796–2804, 2013.
[24] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *arXiv preprint arXiv:1505.05114*, 2015.
[25] M. Soltanolkotabi, *Algorithms and theory for clustering and nonconvex quadratic programming*. PhD thesis, Stanford University, 2014.
[26] C. D. White, R. Ward, and S. Sanghavi, "The local convexity of solving quadratic equations," *arXiv preprint arXiv:1506.07868*, 2015.
[27] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere," *arXiv preprint arXiv:1504.06785*, 2015.
[28] A. Anandkumar and R. Ge, "Efficient approaches for escaping higher order saddle points in non-convex optimization," *arXiv preprint arXiv:1602.05908*, 2016.
[29] K. Kreutz-Delgado, "The complex gradient operator and the $\mathbb{CR}$-calculus," *arXiv preprint arXiv:0906.4835*, 2009.
[30] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *arXiv preprint arXiv:1602.06664*, 2016.
[31] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing* (Y. C. Eldar and G. Kutyniok, eds.), pp. 210–268, Cambridge University Press, 2012.
[32] J. J. Moré and D. C. Sorensen, "Computing a trust region step," *SIAM Journal on Scientific and Statistical Computing*, vol. 4, no. 3, pp. 553–572, 1983.
[33] C. Fortin and H. Wolkowicz, "The trust region subproblem and semidefinite programming," *Optimization methods and software*, vol. 19, no. 1, pp. 41–67, 2004.