Ann. Appl. Math.
doi: 10.4208/aam.OA-2021-0010

Vol. **38**, No. 1, pp. 62-114
February 2022

The Global Landscape of Phase Retrieval II: Quotient Intensity Models

Jian-Feng Cai[1], Meng Huang[1], Dong Li[2,*] and Yang Wang[1]

[1] *Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*
[2] *SUSTech International Center for Mathematics and Department of Mathematics, Southern University of Science and Technology, Shenzhen, Guangdong 518055, China*

Received 3 April 2021; Accepted (in revised version) 10 November 2021

Abstract. A fundamental problem in phase retrieval is to reconstruct an unknown signal from a set of magnitude-only measurements. In this work we introduce three novel quotient intensity models (QIMs) based on a deep modification of the traditional intensity-based models. A remarkable feature of the new loss functions is that the corresponding geometric landscape is benign under the optimal sampling complexity. When the measurements $a_i \in \mathbb{R}^n$ are Gaussian random vectors and the number of measurements $m \geq Cn$, the QIMs admit no spurious local minimizers with high probability, i.e., the target solution x is the unique local minimizer (up to a global phase) and the loss function has a negative directional curvature around each saddle point. Such benign geometric landscape allows the gradient descent methods to find the global solution x (up to a global phase) without spectral initialization.

AMS subject classifications: 94A12, 65K10, 49K45

Key words: Phase retrieval, landscape analysis, non-convex optimization.

1 Introduction

*Corresponding author.
Emails: jfcai@ust.hk (J. Cai), menghuang@ust.hk (M. Huang), lid@sustech.edu.cn (D. Li), yangwang@ust.hk (Y. Wang)

1.1 Background

The intensity-based model for phase retrieval is

$$y_j=|a_j\cdot u|^2,\quad j=1,\cdots,m,$$

where $a_j\in\mathbb{R}^n$, $j=1,\cdots,m$ are given vectors and m is the number of measurements. The phase retrieval problem aims to recover the unknown signal $x\in\mathbb{R}^n$ based on the measurements $\{(a_j,y_j)\}_{j=1}^m$. A natural approach to solve this problem is to consider the minimization problem

$$\min_{u\in\mathbb{R}^n}\quad f(u)=\frac{1}{m}\sum_{j=1}^m((a_j\cdot u)^2-(a_j\cdot x)^2)^2. \tag{1.1}$$

However, as shown in [28], to guarantee the above loss function to have benign geometric landscape, the requirement of sampling complexity is $\mathcal{O}(n\log^3 n)$. This result is recently improved to $\mathcal{O}(n\log n)$ in [6]. On the other hand, due to the heavy tail of the quartic random variables in (1.1), such results seem to be optimal for this class of loss functions.

To remedy this issue, we propose in this work three novel quotient intensity models (QIMs) to recover x under optimal sampling complexity. We rigorously prove that, for Gaussian random measurements, those empirical loss functions admit the benign geometric landscapes with high probability under the optimal sampling complexity $\mathcal{O}(n)$. Here, the phrase "benign" means: (1) the loss function has no spurious local minimizers; and (2) the loss function has a negative directional curvature around each saddle point. The three quotient intensity models are

QIM1:

$$\min_{u\in\mathbb{R}^n} f(u)=\frac{1}{m}\sum_{k=1}^m\frac{((a_k\cdot u)^2-(a_k\cdot x)^2)^2}{(a_k\cdot x)^2}. \tag{1.2}$$

QIM2:

$$\min_{u\in\mathbb{R}^n} f(u)=\frac{1}{m}\sum_{k=1}^m\frac{((a_k\cdot u)^2-(a_k\cdot x)^2)^2}{\beta\|u\|_2^2+(a_k\cdot x)^2}. \tag{1.3}$$

QIM3:

$$\min_{u\in\mathbb{R}^n} f(u)=\frac{1}{m}\sum_{k=1}^m\frac{((a_k\cdot u)^2-(a_k\cdot x)^2)^2}{\|u\|_2^2+\beta_1(a_k\cdot u)^2+\beta_2(a_k\cdot x)^2}. \tag{1.4}$$

The phase retrieval problem arises in many fields of science and engineering such as X-ray crystallography [17, 23], microscopy [22], astronomy [8], coherent diffractive imaging [16,27] and optics [33] etc. In practical applications due to the physical limitations optical detectors can only record the magnitude of signals while losing the phase information. Many algorithms have been designed to solve the phase retrieval problem, which includes convex algorithms and non-convex ones. The convex algorithms usually rely on a "matrix-lifting" technique, which is computationally inefficient for large scale problems [2, 4, 32]. In contrast, many non-convex algorithms bypass the lifting step and operate directly on the lower-dimensional ambient space, making them much more computationally efficient. Early non-convex algorithms were mostly based on the technique of alternating projections, e.g., Gerchberg-Saxton [15] and Fineup [10]. The main drawback, however, is the lack of theoretical guarantee. Later Netrapalli et al. [24] proposed the AltMinPhase algorithm based on a technique known as *spectral initialization*. They proved that the algorithm linearly converges to the true solution with $\mathcal{O}(n\log^3 n)$ resampling Gaussian random measurements. This work led further to several other non-convex algorithms based on spectral initialization. A common thread is first choosing a good initial guess through spectral initialization, and then solving an optimization model through gradient descent, such as the Wirtinger Flow method [3], Truncated Wirtinger Flow algorithm [7], randomized Kaczmarz method [18,30,35], Gauss-Newton method [12], Truncated Amplitude Flow algorithm [34], Reshaped Wirtinger Flow (RWF) [36] and so on.

1.2 Prior arts and connections

As was already mentioned earlier, producing a good initial guess using spectral initialization seems to be a prerequisite for prototypical non-convex algorithms to succeed with good theoretical guarantees. A natural and fundamental question is:

Is it possible for non-convex algorithms to achieve successful recovery with a random initialization (*i.e., without spectral initialization or any additional truncation*)?

In the recent work [28], Ju Sun et al. carried out a deep study of the global geometric structure of phase retrieval problem. They proved that the loss function does not have any spurious local minima under $\mathcal{O}(n\log^3 n)$ Gaussian random measurements. More specifically, it was shown in [28] that all local minimizers coincide with the target signal $\boldsymbol{x}$ up to a global phase, and the loss function has a negative directional curvature around each saddle point. Thanks to this benign geometric landscape any algorithm which can avoid saddle points converges to the true solution with high probability. A trust-region method was employed in [28] to find the global minimizers with random initialization. To reduce the sampling complexity,

it has been shown in [21] that a combination of the loss function with a judiciously chosen activation function also possesses the benign geometry structure under $\mathcal{O}(n)$ Gaussian random measurements. Recently, a smoothed amplitude flow estimator has been proposed in [5] and the authors show that the loss function has benign geometry structure under the optimal sampling complexity. Numerical tests show that the estimator in [5] yields very stable and fast convergence with random initialization and performs as good as or even better than the existing gradient descent methods with spectral initialization.

The emerging concept of a benign geometric landscape has also recently been explored in many other applications of signal processing and machine learning, e.g., matrix sensing [1,25], tensor decomposition [13], dictionary learning [29] and matrix completion [14]. For general optimization problems there exist a plethora of loss functions with well-behaved geometric landscapes such that all local optima are also global optima and each saddle point has a negative direction curvature in its vincinity. Correspondingly several techniques have been developed to guarantee that the standard gradient based optimization algorithms can escape such saddle points efficiently, see e.g., [9, 19, 20].

1.3 Our contributions

This paper aims to show the intensity-based model (1.1) with some deep modification has a benign geometry structure under the optimal sampling complexity. More specifically, we first introduce three novel quotient intensity models and then we prove rigorously that each loss function of them has no spurious local minimizers. Furthermore, the loss function of quotient intensity model has a negative directional curvature around each saddle point. Such properties allow first order method like gradient descent to locate a global minimum with random initial guess.

Our first result shows that the loss function of (1.2) has the benign geometric landscape, as stated below.

Theorem 1.1 (Informal)**.** *Consider the quotient intensity model* (1.2)*. Assume* $\{a_i\}_{i=1}^m$ *are i.i.d. standard Gaussian random vectors and* $x\neq 0$*. There exist positive absolute constants* c*,* C*, such that if* $m\geq Cn$*, then with probability at least* $1-e^{-cm}$ *the loss function* $f=f(u)$ *has no spurious local minimizers. The only local minimizers are* $\pm x$*. All other critical points are strict saddles.*

The second result is the global analysis for the estimator (1.3).

Theorem 1.2 (Informal)**.** *Consider the quotient intensity model* (1.3)*. Let* $0<\beta<\infty$*. Assume* $\{a_i\}_{i=1}^m$ *are i.i.d. standard Gaussian random vectors and* $x\neq 0$*. There exist positive constants* c*,* C *depending only on* β*, such that if* $m\geq Cn$*, then*

with probability at least $1-e^{-cm}$ *the loss function* $f=f(u)$ *has no spurious local minimizers. The only local minimizer is* $\pm x$ *and all other critical points are strict saddles.*

Remark 1.1. There appears some subtle differences between estimators (1.2) and (1.3). Although the former looks more singular, one can prove full strong convexity in the neighborhood of the global minimizers. In the latter case, however, we only have certain restricted convexity.

The third result is the global landscape for the estimator (1.4).

Theorem 1.3 (Informal)**.** *Consider the quotient intensity model* (1.4)*. Let* $0<\beta_1,\beta_2<\infty$*. Assume* $\{a_i\}_{i=1}^m$ *are i.i.d. standard Gaussian random vectors and* $x\neq 0$*. There exist positive constants* c*,* C *depending only on* β*, such that if* $m\geq Cn$*, then with probability at least* $1-e^{-cm}$ *the loss function* $f=f(u)$ *has no spurious local minimizers. The only local minimizer is* $\pm x$ *and all other critical points are strict saddles.*

Remark 1.2. For this case, thanks to the strong damping, we have full strong convexity in the neighborhood of the global minimizers.

1.4 Notations

Throughout this proof we fix $\beta>0$ as a constant and do not study the precise dependence of other parameters on β. We write $u\in\mathbb{S}^{n-1}$ if $u\in\mathbb{R}^n$ and $\|u\|_2=\sqrt{\sum_j u_j^2}=1$. We use χ to denote the usual characteristic function. For example $\chi_A(x)=1$ if $x\in A$ and $\chi_A(x)=0$ if $x\notin A$. We denote by δ_1, ϵ, η, η_1 various constants whose value will be taken sufficiently small. The needed smallness will be clear from the context. For any quantity X, we shall write $X=\mathcal{O}(Y)$ if $|X|\leq CY$ for some constant $C>0$. We write $X\lesssim Y$ if $X\leq CY$ for some constant $C>0$. We shall write $X\ll Y$ if $X\leq cY$ where the constant $c>0$ will be sufficiently small. In our proof it is important for us to specify the precise dependence of the sampling size m in terms of the dimension n. For this purpose we shall write $m\gtrsim n$ if $m\geq Cn$ where the constant C is allowed to depend on β and the small constants ϵ, ϵ_i etc used in the argument. One can extract more explicit dependence of C on the small constants and β but for simplicity we suppress this dependence here. We shall say an event A happens with **high probability** if $\mathbb{P}(A)\geq 1-Ce^{-cm}$, where $c>0$, $C>0$ are constants. The constants c and C are allowed to depend on β and the small constants ϵ, δ mentioned before.

1.5 Organization

In Sections 2–4 we carry out an in-depth analysis of the corresponding geometric landscape of QIM1, QIM2 and QIM3 under optimal sampling complexity $\mathcal{O}(n)$. In Section 5, we report some numerical experiments to demonstrate the efficiency of our proposed estimators. In Appendix, we collect the technique lemmas which are used in the proof.

2 Quotient intensity model I

In this section, we consider the first quotient intensity model and prove that it has benign geometric landscape, as demonstrated below.

$$f(u)=\frac{1}{m}\sum_{k=1}^{m}\frac{((a_k\cdot u)^2-(a_k\cdot x)^2)^2}{(a_k\cdot x)^2}. \tag{2.1}$$

Theorem 2.1. *Assume $\{a_k\}_{k=1}^m$ are i.i.d. standard Gaussian random vectors and $x\neq 0$. There exist positive absolute constants c, C, such that if $m\geq Cn$, then with probability at least $1-e^{-cm}$ the loss function $f=f(u)$ defined by (2.1) has no spurious local minimizers. The only local minimizer is $\pm x$, and the loss function is strongly convex in a neighborhood of $\pm x$. The point $u=0$ is a local maximum point with strictly negative-definite Hessian. All other critical points are strict saddles, i.e., each saddle point has a neighborhood where the function has negative directional curvature.*

Without loss of generality we shall assume $x=e_1$ throughout the rest of the proof. Note that the set $\bigcup_{k=1}^m\{a_k\cdot e_1=0\}$ has measure zero. Thus for typical realization we have $a_k\cdot e_1\neq 0$ for all k. This means that the loss function $f(u)$ defined by (2.1) is smooth almost surely. We denote the Hessian of the function $f(u)$ along the ξ-direction ($\xi\in\mathbb{S}^{n-1}$) as

$$H_{\xi\xi}(u)=\sum_{i,j=1}^{n}\xi_i\xi_j(\partial_{ij}f)(u)=\frac{4}{m}\sum_{k=1}^{m}\Big(3\frac{(a_k\cdot\xi)^2(a_k\cdot u)^2}{(a_k\cdot e_1)^2}-(a_k\cdot\xi)^2\Big). \tag{2.2}$$

2.1 Strong convexity near the global minimizers $u=\pm e_1$

Theorem 2.2 (Strong convexity near $u=\pm e_1$)**.** *There exists an absolute constant $0<\epsilon_0\ll 1$ such that the following hold. For $m\gtrsim n$, it holds with high probability that*

$$H_{\xi\xi}(u)\geq 1,\quad \forall\xi\in\mathbb{S}^{n-1},\quad \forall u\ \ \textit{with}\ \ \|u\pm e_1\|_2\leq\epsilon_0.$$

Proof. By Lemma A.1, we can take $\epsilon>0$ sufficiently small, N sufficiently large such that

$$\mathbb{E}\frac{(a_k\cdot\xi)^2(a_k\cdot e_1)^2}{\epsilon+(a_k\cdot e_1)^2}\phi\Big(\frac{a_k\cdot\xi}{N}\Big)\geq 0.99,\quad \forall\xi\in\mathbb{S}^{n-1},\quad \forall 1\leq k\leq m.$$

In the above $\phi\in C_c^\infty(\mathbb{R})$ satisfies $0\leq\phi(x)\leq 1$ for all x, $\phi(x)=1$ for $|x|\leq 1$ and $\phi(x)=0$ for $|x|\geq 2$. Clearly if $\|u\pm e_1\|_2\leq\epsilon_0$ and ϵ_0 is sufficiently small (depending on ϵ and N), then

$$\mathbb{E}\frac{(a_k\cdot\xi)^2(a_k\cdot u)^2}{\epsilon+(a_k\cdot e_1)^2}\phi\Big(\frac{a_k\cdot\xi}{N}\Big)\geq 0.98,\quad \forall\xi\in\mathbb{S}^{n-1},\quad \forall 1\leq k\leq m.$$

The above term inside the expectation is clearly OK for union bounds. Thus for $\|u\pm e_1\|\leq\epsilon_0$ and $m\gtrsim n$, it holds with high probability that

$$\frac{1}{4}H_{\xi\xi}(u)\geq\frac{1}{m}\sum_{k=1}^m\Big(\frac{(a_k\cdot\xi)^2(a_k\cdot u)^2}{\epsilon+(a_k\cdot e_1)^2}\phi(\frac{a_k\cdot\xi}{N})-(a_k\cdot\xi)^2\Big)\geq 3\cdot 0.97-1.01,\quad \forall\xi\in\mathbb{S}^{n-1}.$$

Thus the desired inequality follows. □

2.2 The regimes $\|u\|_2\ll 1$ and $\|u\|_2\gg 1$ are fine

We first investigate the point $u=0$. It is trivial to verify that $\nabla f(0)=0$ since $a_k\cdot e_1\neq 0$ for all k almost surely.

Lemma 2.1 ($u=0$ has strictly negative-definite Hessian)**.** *We have $u=0$ is a local maximum point with strictly negative-definite Hessian. More precisely, for $m\gtrsim n$, it holds with high probability that*

$$\sum_{k,l=1}^n\xi_k\xi_l(\partial_{kl}f)(0)\leq -1,\quad \forall\xi\in\mathbb{S}^{n-1}.$$

Proof. By (2.2), it is obvious that

$$H_{\xi\xi}(0)=-4\frac{1}{m}\sum_{k=1}^m(a_k\cdot\xi)^2.$$

The desired conclusion then easily follows from Bernstein's inequality. □

Write $u=\sqrt{R}\hat{u}$ where $\hat{u}\in S^{n-1}$ and $R>0$. Then

$$f(u)=\frac{1}{m}\sum_{k=1}^{m}\frac{\left(R(a_k\cdot\hat{u})^2-(a_k\cdot e_1)^2\right)^2}{(a_k\cdot e_1)^2}.$$

A simple calculation leads to

$$\partial_R f=2R\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^4}{(a_k\cdot e_1)^2}-2\frac{1}{m}\sum_{k=1}^{m}(a_k\cdot\hat{u})^2; \tag{2.3a}$$

$$\partial_{RR} f=2\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^4}{(a_k\cdot e_1)^2}. \tag{2.3b}$$

Lemma 2.2 (The regime $\|u\|_2\geq 1+\epsilon_0$ is OK)**.** *Let $0<\epsilon_0\ll 1$ be any given small constant. Then the following hold: For $m\gtrsim n$, with high probability it holds that*

$$\partial_R f>0,\quad \forall R\geq 1+\epsilon_0,\quad \forall\hat{u}\in\mathbb{S}^{n-1}.$$

Proof. Denote $X_k=a_k\cdot e_1$ and $Z_k=a_k\cdot\hat{u}$. By (2.3a) and Cauchy-Schwartz, we have

$$\begin{aligned}\partial_R f&\geq\frac{2R}{m}\frac{\left(\sum_{k=1}^{m}(a_k\cdot\hat{u})^2\right)^2}{\sum_{k=1}^{m}(a_k\cdot e_1)^2}-\frac{2}{m}\sum_{k=1}^{m}(a_k\cdot\hat{u})^2\\&\geq 2R\cdot(1-\delta_1)-2(1+\delta_1),\quad\forall\hat{u}\in\mathbb{S}^{n-1},\end{aligned}$$

where $0<\delta_1\ll 1$ is an absolute constant which we can take to be sufficiently small, and in the last inequality we have used Bernstein. The desired result then easily follows by taking $R\geq R_1=\frac{1+2\delta_1}{1-\delta_1}$ and choosing δ_1 such that $R_1\leq 1+\epsilon_0$. □

From (2.3a), due to the highly irregular coefficients near R, it is difficult to control the upper bound of $\partial_R f$ in the regime $R\ll 1$. To resolve this difficulty, we shall examine the Hessian in this regime.

Lemma 2.3 (The regime $\|u\|_2\leq\frac{1}{3}$ is OK)**.** *For $m\gtrsim n$, with high probability it holds that*

$$H_{e_1e_1}(u)\leq-\frac{1}{2}<0,\quad\forall 0<\|u\|_2\leq\frac{1}{3},$$

where $H_{e_1e_1}$ is defined in (2.2).

Proof. It follows from (2.2) together with Bernstein's inequality that for $m\gtrsim n$ with high probability, it holds

$$\frac{1}{4}H_{e_1e_1}(u)=\frac{1}{m}\sum_{k=1}^{m}\Big(3(a_k\cdot u)^2-(a_k\cdot e_1)^2\Big)\leq\|u\|_2^2\cdot 3\cdot\frac{10}{9}-\frac{8}{9}\leq-\frac{1}{2}.$$

This completes the proof. □

Theorem 2.3 (The regimes $\|u\|_2\leq\frac{1}{3}$ and $\|u\|_2\geq 1+\epsilon_0$ are OK)**.** *Let $0<\epsilon_0\ll 1$ be a given small constant. For $m\gtrsim n$, with high probability the following hold:*

1. *We have*

$$\partial_R f>0,\quad \forall R\geq 1+\epsilon_0,\quad \forall\hat{u}\in\mathbb{S}^{n-1}.$$

2. *The point $u=0$ is a local maximum point with strictly negative-definite Hessian,*

$$\sum_{k,l=1}^{n}\xi_k\xi_l(\partial_{kl}f)(0)\leq-1,\quad \forall\xi\in\mathbb{S}^{n-1}.$$

3. *We have*

$$H_{e_1e_1}(u)\leq-1,\quad \forall\|u\|_2\leq\frac{1}{3}.$$

Proof. This follows from Lemmas 2.1, 2.2 and 2.3. □

Theorem 2.4 (The regime $\|u\|_2\sim 1$, $||\hat{u}\cdot e_1|-1|\geq\eta_0$)**.** *Let $0<\eta_0\ll 1$ be given. Then for $m\gtrsim n$, the following hold with high probability: Suppose $u=\sqrt{R}\hat{u}$, $1/9\leq R\leq 2$, and $\big||\hat{u}\cdot e_1|-1\big|\geq\eta_0$. If $(\partial_R f)(u)=0$, then we must have*

$$H_{e_1e_1}(u)<0.$$

Proof. By (2.3a), we have if $\partial_R f(u)=0$, then

$$R\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^4}{(a_k\cdot e_1)^2}=\frac{1}{m}\sum_{k=1}^{m}(a_k\cdot\hat{u})^2.$$

By Lemma A.2, we have for $m\gtrsim n$, it holds with high probability that

$$\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^4}{(a_k\cdot e_1)^2}\geq 100,\quad \forall\hat{u}\in\mathbb{S}^{n-1}\ \text{ with }\ ||\hat{u}\cdot e_1|-1|\geq\eta_0.$$

Clearly then $R\leq\frac{1}{50}$ with high probability. Thus it follows easily that $H_{e_1e_1}(u)<0$ also with high probability. □

Theorem 2.5 (Localization of R when $||\hat{u}\cdot e_1|-1|\leq\eta_0$, $R\leq 1+\eta_0$ and u is a critical point)**.** *Let $0<\eta_0\ll 1$ be given. For $m\gtrsim n$, the following hold with high probability: Assume $u=\sqrt{R}\hat{u}$ is a critical point with $\frac{1}{9}\leq R\leq 1+\eta_0$, and $||\hat{u}\cdot e_1|-1|\leq\eta_0$. Then we must have*

$$|R-1|\leq c(\eta_0),$$

where $c(\eta_0)\to 0$ as $\eta_0\to 0$.

Proof. Denote $\partial_\xi f=\xi\cdot\nabla f$ for $\xi\in\mathbb{S}^{n-1}$. It is not difficult to check that

$$\frac{1}{4}\partial_\xi f=\frac{1}{m}\sum_{k=1}^m\frac{(a_k\cdot u)^3(a_k\cdot\xi)}{X_k^2}-\frac{1}{m}\sum_{k=1}^m(a_k\cdot u)(a_k\cdot\xi)=0,$$

where $X_k=a_k\cdot e_1$. Setting $\xi=\hat{u}$ and $\xi=e_1$, respectively give us two equations:

$$R\cdot\Big(\frac{1}{m}\sum_{k=1}^m\frac{(a_k\cdot\hat{u})^4}{X_k^2}\Big)-\frac{1}{m}\sum_{k=1}^m(a_k\cdot\hat{u})^2=0, \tag{2.4a}$$

$$R\cdot\Big(\frac{1}{m}\sum_{k=1}^m\frac{(a_k\cdot\hat{u})^3}{X_k}\Big)-\frac{1}{m}\sum_{k=1}^m(a_k\cdot\hat{u})X_k=0. \tag{2.4b}$$

We then obtain

$$\Big(\frac{1}{m}\sum_{k=1}^m(a_k\cdot\hat{u})^2\Big)\cdot\Big(\frac{1}{m}\sum_{k=1}^m\frac{(a_k\cdot\hat{u})^3}{X_k}\Big)=\Big(\frac{1}{m}\sum_{k=1}^m\frac{(a_k\cdot\hat{u})^4}{X_k^2}\Big)\cdot\Big(\frac{1}{m}\sum_{k=1}^m(a_k\cdot\hat{u})X_k\Big). \tag{2.5}$$

Without loss of generality we assume $\|\hat{u}-e_1\|_2\leq\eta\ll 1$. Then with high probability we have

$$\frac{1}{m}\sum_{k=1}^m(a_k\cdot\hat{u})X_k=1+\mathcal{O}(\eta),$$
$$\frac{1}{m}\sum_{k=1}^m(a_k\cdot\hat{u})^2=1+\mathcal{O}(\eta).$$

Observe that by Cauchy-Schwartz,

$$\sum_{k=1}^m\frac{|a_k\cdot\hat{u}|^3}{|X_k|}\leq\Big(\sum_{k=1}^m\frac{|a_k\cdot\hat{u}|^4}{X_k^2}\Big)^{\frac{1}{2}}\cdot\Big(\sum_{k=1}^m(a_k\cdot\hat{u})^2\Big)^{\frac{1}{2}}.$$

Plugging the above estimates into (2.5), we obtain

$$\sqrt{\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^4}{X_k^2}}\leq 1+\mathcal{O}(\eta).$$

Using (2.4a), we then get

$$R\geq 1+\mathcal{O}(\eta).$$

The desired result then easily follows. □

We now complete the proof of the main theorem.

Proof of Theorem 2.1. We proceed in several steps.

1. By Theorem 2.2, the function $f(u)$ is strongly convex when $\|u\pm e_1\|_2\ll 1$.
2. By Theorem 2.3, f has non-vanishing gradient when $R\geq 1+\epsilon_0$. Also $H_{e_1e_1}(u)\leq -1$ when $\|u\|_2\leq\frac{1}{3}$. The point $u=0$ is a strict local maximum point with strictly negative-definite Hessian.
3. By Theorem 2.4, we have $H_{e_1e_1}(u)<0$ if $\|u\|_2\sim 1$ and $||\hat{u}\cdot e_1|-1|\geq\epsilon_0$.
4. Theorem 2.5 shows that if $R\leq 1+\epsilon_0$, $||\hat{u}\cdot e_1|-1|\leq\epsilon_0$ and u is a critical point, then we must have $|R-1|\leq c(\epsilon_0)\ll 1$. In yet other words we must have $\|u\pm e_1\|_2\ll 1$. This regime is then treated by Step 1.

This completes the proof. □

3 Quotient intensity model II

Consider for $\beta>0$,

$$f(u)=\frac{1}{m}\sum_{k=1}^{m}\frac{((a_k\cdot u)^2-(a_k\cdot x)^2)^2}{\beta\|u\|_2^2+(a_k\cdot x)^2}. \tag{3.1}$$

Theorem 3.1. *Let $0<\beta<\infty$. Assume $\{a_k\}_{k=1}^m$ are i.i.d. standard Gaussian random vectors and $x\neq 0$. There exist positive constants c, C depending only on β, such that if $m\geq Cn$, then with probability at least $1-e^{-cm}$ the loss function $f=f(u)$ defined by* (3.1) *has no spurious local minimizers. The only local minimizer is $\pm x$, and the loss function is restrictively convex in a neighborhood of $\pm x$. The point $u=0$ is a local maximum point with strictly negative-definite Hessian. All other critical points are strict saddles, i.e., each saddle point has a neighborhood where the function has negative directional curvature.*

Remark 3.1. See Theorem 3.4 for the precise statement concerning restrictive convexity.

Without loss of generality we shall assume $x=e_1$ throughout the rest of the proof.

3.1 The regimes $\|u\|_2 \ll 1$ and $\|u\|_2 \gg 1$ are fine

We first investigate the point $u=0$. It is trivial to verify that $\nabla f(0)=0$ since $a_k \cdot e_1 \neq 0$ for all k almost surely.

Lemma 3.1 ($u=0$ has strictly negative-definite Hessian)**.** *We have $u=0$ is local maximum point with strictly negative-definite Hessian. More precisely, for $m \gtrsim n$, it holds with high probability that*

$$\sum_{k,l=1}^{n} \xi_k \xi_l (\partial_{kl} f)(0) \leq -d_1, \quad \forall \xi \in \mathbb{S}^{n-1},$$

where $d_1>0$ is an absolute constant.

Proof. We begin by noting that since almost surely $a_k \cdot e_1 \neq 0$ for all k, the function f is smooth at $u=0$. It suffices for us to consider (write $u=\sqrt{t}\xi$)

$$G(t)=\frac{1}{m}\sum_{k=1}^{m} \frac{(t(a_k\cdot\xi)^2-(a_k\cdot e_1)^2)^2}{\beta t+(a_k\cdot e_1)^2}.$$

Clearly

$$G'(0)=-\beta-2\frac{1}{m}\sum_{k=1}^{m}(a_k\cdot\xi)^2.$$

The desired conclusion then easily follows by using Bernstein's inequality. □

Write $u=\sqrt{R}\hat{u}$ where $\hat{u}\in S^{n-1}$ and $R>0$. Then

$$f(u)=\frac{1}{m}\sum_{k=1}^{m} \frac{\left(R(a_k\cdot\hat{u})^2-(a_k\cdot e_1)^2\right)^2}{\beta R+(a_k\cdot e_1)^2}.$$

Clearly

$$\partial_R f=\frac{1}{m}\sum_{k=1}^{m} \frac{R^2(\beta(a_k\cdot\hat{u})^4)+2R(a_k\cdot\hat{u})^4(a_k\cdot e_1)^2-\beta(a_k\cdot e_1)^4-2(a_k\cdot e_1)^4(a_k\cdot\hat{u})^2}{(\beta R+(a_k\cdot e_1)^2)^2}; \quad (3.2)$$

$$\partial_{RR} f=2\frac{1}{m}\sum_{k=1}^{m} \frac{(a_k\cdot e_1)^4(\beta+(a_k\cdot\hat{u})^2)^2}{(\beta R+(a_k\cdot e_1)^2)^3}. \quad (3.3)$$

Lemma 3.2 (The regime $\|u\|_2 \gg 1$ is OK)**.** *There exist constants $R_1 = R_1(\beta) > 0$, $d_1 = d_1(\beta) > 0$ such that the following hold: For $m \gtrsim n$, with high probability it holds that*

$$\partial_R f \geq d_1, \quad \forall R \geq R_1, \quad \forall \hat{u} \in \mathbb{S}^{n-1}.$$

Proof. We only sketch the proof. Denote $X_k = a_k \cdot e_1$ and $Z_k = a_k \cdot \hat{u}$. Using the inequalities (assume $R \gg 1$ and denote by $C_1 > 0$ a constant depending only on β)

$$\beta R + X_k^2 \leq R(\beta + X_k^2), \quad (\beta R + X_k^2)^2 \geq 4\beta R X_k^2,$$

and

$$\frac{X_k^4}{(\beta R + X_k^2)^2} \leq C_1 \cdot \Big(\frac{R}{R^2} + \chi_{|X_k| \geq R^{\frac{1}{4}}} \Big),$$

we have

$$\partial_R f \geq \frac{1}{m} \sum_{k=1}^m \frac{\beta Z_k^4}{(\beta + X_k^2)^2} \phi(\frac{Z_k}{N}) - \frac{1}{m} \sum_{k=1}^m \frac{1}{4R} X_k^2 - \frac{2}{m} \sum_{k=1}^m C_1 \cdot (R^{-1} + \chi_{|X_k| \geq R^{\frac{1}{4}}}) \cdot Z_k^2,$$

where we have chosen $\phi \in C_c^\infty$ such that $0 \leq \phi(x) \leq 1$ for all x, $\phi(x) = 1$ for $|x| \leq 1$ and $\phi(x) = 0$ for $|x| \geq 2$. Observe that for $a \sim \mathcal{N}(0, \mathrm{I_n})$, $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{E}(a \cdot \hat{u})^4 \chi_{|a \cdot \hat{u}| \geq N} \leq \mathbb{E} Z^4 \chi_{|Z| \geq N} \to 0 \quad \text{as} \quad N \to \infty.$$

It is also easy to show that

$$\inf_{\hat{u} \in \mathbb{S}^{n-1}} \mathbb{E} \frac{(a \cdot \hat{u})^4}{(\beta + (a \cdot e_1)^2)^2} \gtrsim 1.$$

Thus we can take N large such that

$$\inf_{\hat{u} \in \mathbb{S}^{n-1}} \mathbb{E} \frac{(a \cdot \hat{u})^4}{(\beta + (a \cdot e_1)^2)^2} \phi\Big(\frac{a \cdot \hat{u}}{N} \Big) \gtrsim 1.$$

It is easy to show that by taking R large, for $m \gtrsim n$, it holds with high probability that

$$\frac{1}{m} \sum_{k=1}^m \chi_{|X_k| \geq R^{\frac{1}{4}}} Z_k^2 \leq \epsilon.$$

Since all the other terms are OK for union bounds, the desired result then clearly follows by taking R large. □

Lemma 3.3 (The regime $\|u\|_2 \ll 1$ with $\frac{|u_1|}{\|u\|_2} \le \frac{1}{10}$ is OK)**.** *There exist a constant $R_2 = R_2(\beta) > 0$ such that the following hold: For $m \gtrsim n$, with high probability it holds that*

$$\partial_{u_1 u_1} f \le -2 < 0, \quad \forall 0 < R \le R_2, \quad \forall \hat{u} \in \mathbb{S}^{n-1} \quad \textit{with} \quad |\hat{u} \cdot e_1| \le \frac{1}{10}.$$

Proof. We only sketch the proof. Denote $X_k = a_k \cdot e_1$ and $Z_k = a_k \cdot \hat{u}$. A short computation gives

$$\begin{aligned}\partial_{u_1 u_1} f = &\frac{4}{m}\sum_{k=1}^m \frac{3RX_k^2 Z_k^2 - X_k^4}{\beta R + X_k^2} + \frac{1}{m}\sum_{k=1}^m (RZ_k^2 - X_k^2)^2 \cdot \frac{6\beta^2 u_1^2 - 2\beta^2 |u'|^2 - 2\beta X_k^2}{(\beta R + X_k^2)^3} \\ &-16\beta(\hat{u}\cdot e_1) R^2 \frac{1}{m}\sum_{k=1}^m \frac{Z_k^3 X_k}{(\beta R + X_k^2)^2} + 16\beta(\hat{u}\cdot e_1) R \frac{1}{m}\sum_{k=1}^m \frac{X_k^3 Z_k}{(\beta R + X_k^2)^2},\end{aligned}$$

where $u_1 = u \cdot e_1$ and $u' = u - u_1 e_1$. Now observe that

$$\begin{aligned}&\frac{1}{m}\sum_{k=1}^m \frac{Z_k^2}{\beta R + X_k^2} X_k^2 \le \frac{1}{m}\sum_{k=1}^m Z_k^2; \\ &\frac{1}{m}\sum_{k=1}^m \frac{X_k^4}{\beta R + X_k^2} = \frac{1}{m}\sum_{k=1}^m \frac{(\beta R + X_k^2 - \beta R)^2}{\beta R + X_k^2} \\ &\ge \Big(\frac{1}{m}\sum_{k=1}^m (\beta R + X_k^2)\Big) - 2\beta R \ge -\beta R + \frac{1}{m}\sum_{k=1}^m X_k^2; \\ &\frac{1}{m}\sum_{k=1}^m \frac{R^{\frac{3}{2}} |Z_k|^3 R^{\frac{1}{2}} |X_k|}{(\beta R + X_k^2)^2} \le \epsilon_1 \frac{1}{m}\sum_{k=1}^m \frac{R^2 Z_k^4}{(\beta R + X_k^2)^2} + \frac{1}{\epsilon_1^3}\frac{1}{m}\sum_{k=1}^m \frac{R^2 X_k^4}{(\beta R + X_k^2)^2} \\ &\le \frac{R^2}{\epsilon_1^3} + \epsilon_1 \frac{1}{m}\sum_{k=1}^m \frac{R^2 Z_k^4}{(\beta R + X_k^2)^2}; \\ &\frac{R}{m}\sum_{k=1}^m \frac{|X_k|^3 |Z_k|}{(\beta R + X_k^2)^2} \le \frac{R}{m}\sum_{k=1}^m \frac{|X_k|^3 |Z_k|}{(3(\beta R)^{\frac{1}{3}} (\frac{1}{4} X_k^4)^{\frac{1}{3}})^2} \\ &\lesssim R^{\frac{1}{3}} \beta^{-\frac{2}{3}} \frac{1}{m}\sum_{k=1}^m |X_k|^{\frac{1}{3}} |Z_k| \lesssim R^{\frac{1}{3}} \beta^{-\frac{2}{3}} \frac{1}{m}\sum_{k=1}^m (X_k^2 + Z_k^2 + 1),\end{aligned}$$

where in the above the constant $\epsilon_1 > 0$ will be taken sufficiently small. The needed smallness will become clear momentarily. Since $|u_1|/\|u\|_2 \le \frac{1}{10}$, it is clear that for some absolute constant $C_1 > 0$,

$$\frac{6\beta^2 u_1^2 - 2\beta^2 \|u'\|_2^2 - 2\beta X_k^2}{(\beta R + X_k^2)^3} \le -\beta C_1 \cdot \frac{1}{(\beta R + X_k^2)^2}.$$

Now

$$-\frac{1}{m}\sum_{k=1}^{m}\frac{(RZ_k^2-X_k^2)^2}{(\beta R+X_k^2)^2}\leq-\frac{1}{m}\sum_{k=1}^{m}\frac{R^2Z_k^4}{(\beta R+X_k^2)^2}+\frac{2R}{m}\sum_{k=1}^{m}\frac{Z_k^2}{\beta R+X_k^2}.$$

Now take $\epsilon_1=\frac{C_1}{1000}$. By Lemma B.1, we can take R sufficiently small such that with high probability

$$\frac{2R}{m}\sum_{k=1}^{m}\frac{Z_k^2}{\beta R+X_k^2}<\frac{1}{100}.$$

All the other terms can be treated by taking R sufficiently small, and the desired result follows easily. □

Lemma 3.4 (The regime $\|u\|_2\ll 1$ with $\frac{|u_1|}{\|u\|_2}>\frac{1}{10}$ is OK)**.** *There exist a constant $R_3=R_3(\beta)>0$ such that the following hold: For $m\gtrsim n$, with high probability it holds that the loss function $f=f(u)$ has no critical points in the regime*

$$\Big\{u=\sqrt{R}\hat{u}\colon\ 0<R\leq R_3,\ \hat{u}\in\mathbb{S}^{n-1}\ \textit{and}\ \ |\hat{u}\cdot e_1|>\frac{1}{10}\Big\}.$$

Proof. We assume that for $0<R\ll 1$ there exists some critical point. The idea is to examine the necessary conditions for a potential critical point and then derive a lower bound on R. Denote $X_k=a_k\cdot e_1$ and $Z_k=a_k\cdot\hat{u}$. By (3.2), we have $\partial_R f=0$ which gives

$$\underbrace{R\frac{1}{m}\sum_{k=1}^{m}\frac{R\beta Z_k^4+2Z_k^4X_k^2}{(\beta R+X_k^2)^2}}_{=:A_1}=\underbrace{\frac{1}{m}\sum_{k=1}^{m}\frac{\beta X_k^4+2X_k^4Z_k^2}{(\beta R+X_k^2)^2}}_{=:B_1}.$$

On the other hand, by using $\partial_{u_1}f(u)=0$, we obtain

$$\begin{aligned}&\beta u_1R^2\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^4}{(\beta R+X_k^2)^2}-2R^{\frac{3}{2}}\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^3X_k}{\beta R+X_k^2}\\=&\beta u_1\frac{1}{m}\sum_{k=1}^{m}\frac{2RZ_k^2X_k^2-X_k^4}{(\beta R+X_k^2)^2}-2R^{\frac{1}{2}}\frac{1}{m}\sum_{k=1}^{m}\frac{Z_kX_k^3}{\beta R+X_k^2}.\end{aligned}$$

Thus

$$\begin{aligned}&\underbrace{-R\widehat{u}_1\frac{1}{m}\sum_{k=1}^{m}\frac{\beta RZ_k^4}{(\beta R+X_k^2)^2}}_{=:A_2}+R\cdot\underbrace{2\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^3X_k}{\beta R+X_k^2}}_{=:A_3}\\=&\underbrace{\beta\widehat{u}_1\frac{1}{m}\sum_{k=1}^{m}\frac{-2RZ_k^2X_k^2+X_k^4}{(\beta R+X_k^2)^2}+2\frac{1}{m}\sum_{k=1}^{m}\frac{Z_kX_k^3}{\beta R+X_k^2}}_{=:B_2}.\end{aligned}$$

Without loss of generality we assume $\widehat{u}_1>\frac{1}{10}$. Observe that for $0<R\leq 1$, we have

$$\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{(\beta+X_k^2)^2}\lesssim B_1\lesssim 1+\frac{1}{m}\sum_{k=1}^{m}Z_k^2.$$

Thus with high probability $B_1\sim 1$.

Now by Lemma B.1, for $0<R\ll 1$, we have

$$\frac{1}{m}\sum_{k=1}^{m}\frac{RZ_k^2X_k^2}{(\beta R+X_k^2)^2}\leq\frac{1}{m}\sum_{k=1}^{m}\frac{RZ_k^2}{\beta R+X_k^2}\ll 1.$$

Also for $0<R\leq 1$, we have

$$\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{(\beta+X_k^2)^2}\leq\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{(\beta R+X_k^2)^2}\leq 1.$$

By Lemma B.2, for $0<R\ll 1$, we have

$$c_1\leq\frac{1}{m}\sum_{k=1}^{m}\frac{Z_kX_k^3}{\beta R+X_k^2}\leq c_2,$$

where c_1, $c_2>0$ are constants depending only on β. Thus with high probability we have for $0<R\ll 1$, $B_2\sim 1$.

Now since

$$RA_1=B_1,\quad -RA_2+RA_3=B_2;$$

we obtain

$$A_1+B_3A_2=B_3A_3,$$

where $B_3=B_1/B_2$. Observe that $A_1>0$, $A_2>0$, and

$$A_1\sim\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^4}{\beta R+X_k^2};$$
$$A_3\leq\Big(\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^4}{\beta R+X_k^2}\Big)^{\frac{3}{4}}\Big(\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{\beta R+X_k^2}\Big)^{\frac{1}{4}}.$$

It follows easily that with high probability we have

$$A_1\sim 1.$$

But then it follows from the equation $RA_1=B_1$ that we must have $R\sim 1$. Thus the desired result follows. □

Theorem 3.2 (The regimes $\|u\|_2\ll 1$ and $\|u\|_2\gg 1$ are OK)**.** *For $m\gtrsim n$, with high probability the following hold:*

1. *We have*

$$\partial_R f\geq d_1,\quad \forall R\geq R_1,\quad \forall\hat{u}\in\mathbb{S}^{n-1},$$

where d_1, R_1 are constants depending only on β.

2. *We have*

$$\partial_{u_1u_1}f\leq -2<0,\quad \forall 0<R\leq R_2,\quad \forall\hat{u}\in\mathbb{S}^{n-1}\ \ \textit{with}\ \ |\hat{u}\cdot e_1|\leq\frac{1}{10},$$

where $R_2>0$ is a constant depending only on β.

3. *The loss function $f=f(u)$ has no critical points in the regime*

$$\Big\{u=\sqrt{R}\hat{u}\colon\ 0<R\leq R_3,\ \hat{u}\in\mathbb{S}^{n-1}\ \textit{and}\ \ |\hat{u}\cdot e_1|>\frac{1}{10}\Big\},$$

where $R_3>0$ is a constant depending only on β.

4. *The point $u=0$ is a local maximum point with strictly negative-definite Hessian,*

$$\sum_{k,l=1}^{n}\xi_k\xi_l(\partial_{kl}f)(0)\leq -d_2<0,\quad \forall\xi\in\mathbb{S}^{n-1},$$

where $d_2>0$ is an absolute constant.

Proof. This follows from Lemmas 3.1, 3.2, 3.3 and 3.4. □

3.2 The regime $\|u\|_2 \sim 1$

Lemma 3.5 (The regime $\|u\|_2 \sim 1$ with $\epsilon_0 \leq |\hat{u}\cdot e_1| \leq 1-\epsilon_0$ is OK)**.** *Let $0<\epsilon_0 \ll 1$ be given. Assume $0<c_1<c_2<\infty$ are two given constants. Then for $m \gtrsim n$, the following hold with high probability: The loss function $f=f(u)$ has no critical points in the regime:*

$$\Big\{u=\sqrt{R}\hat{u}\colon\ c_1<R<c_2,\ \epsilon_0 \leq |\hat{u}\cdot e_1| \leq 1-\epsilon_0\Big\}.$$

More precisely, introduce the parametrization $\hat{u}=e_1\cos\theta+e^{\perp}\sin\theta$, where $\theta\in[0,\pi]$ and $e^{\perp}\in\mathbb{S}^{n-1}$ satisfies $e^{\perp}\cdot e_1=0$. Then in the aforementioned regime, we have

$$|\partial_\theta f| \geq \alpha_1 > 0,$$

where α_1 depends only on $(\beta,\ \epsilon_0,\ c_1,\ c_2)$.

Proof. See appendix. □

Lemma 3.6 (The regime $\|u\|_2 \sim 1$ with $|\hat{u}\cdot e_1| \leq \epsilon_1$ is OK)**.** *Let $0<\epsilon_1 \ll 1$ be a sufficiently small constant. Assume $0<c_1<c_2<\infty$ are two given constants. Then for $m \gtrsim n$, the following hold with high probability: Consider the regime*

$$\Big\{u=\sqrt{R}\hat{u}\colon\ c_1<R<c_2,\ |\hat{u}\cdot e_1| \leq \epsilon_1\Big\}.$$

Introduce the parametrization $\hat{u}=e_1\cos\theta+e^{\perp}\sin\theta$, where $\theta\in[0,\pi]$ and $e^{\perp}\in\mathbb{S}^{n-1}$ satisfies $e^{\perp}\cdot e_1=0$. Then in the aforementioned regime, we have

$$\partial_{\theta\theta} f \leq -\alpha_2 < 0,$$

where $\alpha_2>0$ depends only on $(\beta,\ \epsilon_1,\ c_1,\ c_2)$.

Proof. See appendix. □

Theorem 3.3 (The regime $\|u\|_2 \sim 1$, $||\hat{u}\cdot e_1|-1| \leq \epsilon_0$, $|\|u\|_2-1| \geq c(\epsilon_0)$ is OK)**.** *Let $0<R_1<1<R_2<\infty$ be given constants. Let $0<\epsilon_0 \ll 1$ be a given sufficiently small constant and consider the regime $\big||\hat{u}\cdot e_1|-1\big| \leq \epsilon_0$ with $R_1 \leq \|u\|_2^2 \leq R_2$. There exists a constant $c_0=c_0(\epsilon_0,R_1,R_2,\beta)>0$ which tends to zero as $\epsilon_0 \to 0$ such that the following hold: For $m \gtrsim n$, with high probability it holds that (below $u=\sqrt{R}\hat{u}$)*

$$\begin{aligned}
&\partial_R f<0, \quad \forall R_2 \leq R \leq 1-c_0, \quad \forall \hat{u}\in\mathbb{S}^{n-1} \quad \textit{with} \quad ||\hat{u}\cdot e_1|-1| \leq \epsilon_0;\\
&\partial_R f>0, \quad \forall 1+c_0 \leq R \leq R_1, \quad \forall \hat{u}\in\mathbb{S}^{n-1} \quad \textit{with} \quad ||\hat{u}\cdot e_1|-1| \leq \epsilon_0.
\end{aligned}$$

Proof. We first consider the regime $R\geq 1+c$. Let $\phi\in C_c^\infty(\mathbb{R})$ be an even function satisfying $0\leq\phi(x)\leq 1$ for all x, $\phi(x)=1$ for $|x|\leq 1$ and $\phi(x)=0$ for $|x|>2$. By using (3.2), we have

$$\partial_R f \tag{3.4}$$
$$\geq\frac{1}{m}\sum_{k=1}^m\frac{R^2\beta(a_k\cdot\hat{u})^4\phi\Big(\frac{a_k\cdot\hat{u}}{K}\Big)+2R(a_k\cdot\hat{u})^4\phi(\frac{a_k\cdot\hat{u}}{K})(a_k\cdot e_1)^2-\beta(a_k\cdot e_1)^4-2(a_k\cdot e_1)^4(a_k\cdot\hat{u})^2}{(\beta R+(a_k\cdot e_1)^2)^2}.$$

By taking K sufficiently large, we can easily obtain

$$\mathbb{E}\Big(1-\phi\Big(\frac{a\cdot e_1}{K}\Big)\Big)(1+(a\cdot e_1)^2)\ll 1,$$

where $a\sim\mathcal{N}(0,\mathrm{I_n})$. For fixed K, it is not difficult to check that the lower bound (3.4) are OK for union bounds and they can be made close to the expectation with high probability, uniformly in $R\sim 1$ and $\hat{u}\in\mathbb{S}^{n-1}$. The perturbation argument (i.e., estimating the error terms coming from replacing $a_k\cdot\hat{u}$ by $a_k\cdot e_1$ and so on) becomes rather easy after taking the expectation. It is then not difficult to show that

$$\partial_R f>0,$$

for $R\geq 1+c(\epsilon_0)$.

Next we turn to the regime $R_2\leq R\leq 1-c(\epsilon_0)$. Without loss of generality we may assume $|1-\hat{u}\cdot e_1|\leq\epsilon_0$. The idea is to exploit the decomposition used in the proof of Lemma 3.4. Namely using $\partial_R f=0$ and $\partial_{u_1} f=0$, we have

$$\begin{aligned}
&\underbrace{R\frac{1}{m}\sum_{k=1}^m\frac{R\beta Z_k^4+2Z_k^4X_k^2}{(\beta R+X_k^2)^2}}_{=:A_1}\\
=&\underbrace{\frac{1}{m}\sum_{k=1}^m\frac{\beta X_k^4+2X_k^4Z_k^2}{(\beta R+X_k^2)^2}}_{=:B_1};\underbrace{-R\widehat{u}_1\frac{1}{m}\sum_{k=1}^m\frac{\beta RZ_k^4}{(\beta R+X_k^2)^2}}_{=:A_2}+R\cdot 2\underbrace{\frac{1}{m}\sum_{k=1}^m\frac{Z_k^3X_k}{\beta R+X_k^2}}_{=:A_3}\\
=&\underbrace{\beta\widehat{u}_1\frac{1}{m}\sum_{k=1}^m\frac{-2RZ_k^2X_k^2+X_k^4}{(\beta R+X_k^2)^2}+2\frac{1}{m}\sum_{k=1}^m\frac{Z_kX_k^3}{\beta R+X_k^2}}_{=:B_2}.
\end{aligned}$$

It is not difficult to check that with high probability, we have $B_1\sim 1$, $B_2\sim 1$, and

$$\Big|\frac{B_2}{B_1}-1\Big|\leq\eta(\epsilon_0)\ll 1,\quad\forall R_2\leq R\leq 1,\quad\forall\hat{u}\in\mathbb{S}^{n-1}\ \text{ with }\ |\hat{u}\cdot e_1-1|\leq\epsilon_0,$$

where $\eta(\epsilon_0)\to 0$ as $\epsilon_0\to 0$. We then obtain

$$A_1=\Big(1+\mathcal{O}(\eta(\epsilon_0))\Big)(-A_2+A_3).$$

From this it is easy (similar to an argument used in the proof of Lemma 3.4) to derive that

$$A_1+A_2+|A_3|\lesssim 1.$$

Now note that the pre-factor of A_2 is $\hat{u}_1=1+\mathcal{O}(\epsilon_0)$. By using the relation

$$A_1+A_2-A_3=\mathcal{O}(\eta(\epsilon_0)),$$

we obtain

$$\frac{1}{m}\sum_{k=1}^m\frac{Z_k^4}{\beta R+X_k^2}-\frac{1}{m}\sum_{k=1}^m\frac{Z_k^3X_k}{\beta R+X_k^2}=\mathcal{O}(\eta(\epsilon_0)).$$

By using localization, i.e., decomposing

$$Z_k^3X_k=Z_k^3\phi\Big(\frac{Z_k}{M}\Big)X_k+Z_k^3\Big(1-\phi\Big(\frac{Z_k}{M}\Big)\Big)X_k,$$

Hölder and taking M sufficiently large, one can then derive that (with high probability)

$$\begin{aligned}&\Big|\frac{1}{m}\sum_{k=1}^m\frac{Z_k^4-X_k^4}{\beta R+X_k^2}\Big|+\Big|\frac{1}{m}\sum_{k=1}^m\frac{Z_k^3X_k-X_k^4}{\beta R+X_k^2}\Big|\\ =&\mathcal{O}(\eta_1(\epsilon_0)),\quad \forall R_2\le R\le 1,\quad \forall\hat{u}\in\mathbb{S}^{n-1}\ \text{ with }\ |\hat{u}\cdot e_1-1|\le\epsilon_0,\end{aligned}$$

where $\eta_1(\epsilon_0)\to 0$ as $\epsilon_0\to 0$. It then follows easily that (with high probability)

$$\Big|\frac{1}{m}\sum_{k=1}^m\frac{(Z_k-X_k)^4}{\beta R+X_k^2}\Big|=\mathcal{O}(\eta_2(\epsilon_0)),\quad \forall R_2\le R\le 1,\quad \forall\hat{u}\in\mathbb{S}^{n-1}\ \text{ with }\quad |\hat{u}\cdot e_1-1|\le\epsilon_0,$$

where $\eta_2(\epsilon_0)\to 0$ as $\epsilon_0\to 0$.

Now observe that for A_1, we have

$$\begin{aligned}|Z_k^4-X_k^4|\le&|Z_k-X_k|(\mathcal{O}(|Z_k|^3)+\mathcal{O}(|X_k|^3))\\ \le&C_\epsilon|Z_k-X_k|^4+\epsilon\cdot(\mathcal{O}(|Z_k|^4)+\mathcal{O}(X_k^4)),\end{aligned}$$

where $C_\epsilon>0$ depends only on ϵ. Clearly by taking $\epsilon>0$ sufficiently small and using the derived quantitative estimates preceding this paragraph, we can guarantee that (with high probability)

$$\Big|A_1-B_1\Big|\ll 1,\quad \forall R_2\leq R\leq 1,\quad \forall \hat{u}\in\mathbb{S}^{n-1}\ \text{ with }\ |\hat{u}\cdot e_1-1|\leq\epsilon_0.$$

It follows that we must have $|R-1|\ll 1$ for a potential critical point. By using (3.2) we have $\partial_R f(R=0)<0$. By using (3.3) we have $\partial_{RR}f>0$. Since we have shown $\partial_R f>0$ for $R>1+c(\epsilon_0)$, it then follows that $\partial_R f=0$ occurs at a unique point $|R-1|\ll 1$ and $\partial_R f<0$ for $R<1-c(\epsilon_0)$ provided $c(\epsilon_0)$ is suitably re-defined. □

We now show restrictive convexity of the loss function $f(u)$ near the global minimizer $u=\pm e_1$.

Theorem 3.4 (Restrictive convexity near the global minimizer)**.** *There exists* $0<\epsilon_0\ll 1$ *sufficiently small such that if* $m\gtrsim n$*, then the following hold with high probability:*

1. *If* $\|u-e_1\|_2\leq\epsilon_0$ *and* $u\neq e_1$*, then for* $\xi=\frac{u-e_1}{\|u-e_1\|_2}\in\mathbb{S}^{n-1}$*, we have*

$$\sum_{i,j=1}^n \xi_i\xi_j(\partial_{ij}f)(u)\geq\gamma>0,$$

   where γ *is a constant depending only on* β*.*

2. *If* $\|u+e_1\|_2\leq\epsilon_0$*, then then for* $\xi=\frac{u+e_1}{\|u+e_1\|_2}\in\mathbb{S}^{n-1}$*, we have*

$$\sum_{i,j=1}^n \xi_i\xi_j(\partial_{ij}f)(u)\geq\gamma>0,$$

   where γ *is a constant depending only on* β*.*

3. *Alternatively we can use the parametrization* $u=\pm e_1+t\xi$*, where* $\xi\in\mathbb{S}^{n-1}$*, and* $|t|\leq\epsilon_0$*. Then with this special parametrization, we have*

$$\sum_{i,j=1}^n \xi_i\xi_j(\partial_{ij}f)(u)\geq\gamma>0.$$

   Note that this includes the global minimizers $u=\pm e_1$*.*

In yet other words, $f(u)$ *is restrictively convex in a sufficiently small neighborhood of* $\pm e_1$*.*

Proof. See appendix. □

Proof of Theorem 3.1. We proceed in several steps as follows.

1. For the regime $\|u\|_2 \ll 1$ and $\|u\|_2 \gg 1$, we use Theorem 3.2. The point $u=0$ is a local maximum point with strictly negative-definite Hessian. All other possible critical points must have negative curvature direction.

2. For the regime $\|u\|_2 \sim 1$, $||\hat{u}\cdot e_1|-1| \geq \epsilon_0$, we use Lemma 3.5 and 3.6. The loss function either has a nonzero gradient, or it is a strict saddle with a negative curvature direction.

3. For the regime $\|u\|_2 \sim 1$, $|\hat{u}\cdot e_1|-1| \leq \epsilon_0$, $|\|u\|_2-1| \geq c(\epsilon_0)$, we apply Theorem 3.3. The loss function has nonzero gradient in this regime.

4. Finally for the regime close to the global minimizers $\pm e_1$, we use Theorem 3.4 to show restrictive convexity. This ensures that $\pm e_1$ are the only minimizers.

This completes the proof. □

4 Quotient intensity model III

Consider for $\beta_1 > 0$, $\beta_2 > 0$,

$$f(u) = \frac{1}{m}\sum_{k=1}^{m} \frac{((a_k\cdot u)^2-(a_k\cdot x)^2)^2}{\|u\|_2^2+\beta_1(a_k\cdot u)^2+\beta_2(a_k\cdot x)^2}. \tag{4.1}$$

Theorem 4.1. *Let $0<\beta_1,\beta_2<\infty$. Assume $\{a_k\}_{k=1}^m$ are i.i.d. standard Gaussian random vectors and $x\neq 0$. There exist positive constants c, C depending only on (β_1,β_2), such that if $m \geq Cn$, then with probability at least $1-e^{-cm}$ the loss function $f=f(u)$ defined by (4.1) has no spurious local minimizers. The only local minimizer is $\pm x$, and the loss function is strongly convex in a neighborhood of $\pm x$. The point $u=0$ is a local maximum point with strictly negative-definite Hessian. All other critical points are strict saddles, i.e., each saddle point has a neighborhood where the function has negative directional curvature.*

Without loss of generality we shall assume $x=e_1$ throughout the rest of the proof. Thus we consider

$$f(u) = \frac{1}{m}\sum_{k=1}^{m} \frac{((a_k\cdot u)^2-(a_k\cdot e_1)^2)^2}{\|u\|_2^2+\beta_1(a_k\cdot u)^2+\beta_2(a_k\cdot e_1)^2}. \tag{4.2}$$

4.1 The regimes $\|u\|_2 \ll 1$ and $\|u\|_2 \gg 1$ are fine

We first investigate the point $u=0$. It is trivial to verify that $\nabla f(0)=0$ since $a_k \cdot e_1 \neq 0$ for all k almost surely.

Lemma 4.1 ($u=0$ has strictly negative-definite Hessian)**.** *We have $u=0$ is a local maximum point with strictly negative-definite Hessian. More precisely, it holds (almost surely) that*

$$\sum_{k,l=1}^{n} \xi_k \xi_l (\partial_{kl} f)(0) \leq -d_1, \quad \forall \xi \in \mathbb{S}^{n-1},$$

where $d_1>0$ is a constant depending only on β_2.

Proof. We begin by noting that since almost surely $a_k \cdot e_1 \neq 0$ for all k, the function f is smooth at $u=0$. It suffices for us to consider (write $u=\sqrt{t}\xi$)

$$G(t)=\frac{1}{m}\sum_{k=1}^{m} \frac{(t(a_k\cdot\xi)^2-(a_k\cdot e_1)^2)^2}{t+t\beta_1(a_k\cdot\xi)^2+\beta_2(a_k\cdot e_1)^2}.$$

By a simple computation, we have

$$G'(0)=-\frac{1}{\beta_2^2}-\frac{\beta_1+2\beta_2}{\beta_2^2}\cdot\frac{1}{m}\sum_{k=1}^{m}(a_k\cdot\xi)^2.$$

The desired conclusion then easily follows. □

Write $u=\sqrt{R}\hat{u}$ where $\hat{u}\in S^{n-1}$ and $R>0$. Denote $X_k=a_k\cdot e_1$. Then

$$f(u)=\frac{1}{m}\sum_{k=1}^{m}\frac{\left(R(a_k\cdot\hat{u})^2-X_k^2\right)^2}{R+\beta_1 R(a_k\cdot\hat{u})^2+\beta_2 X_k^2}.$$

Clearly

$$\partial_R f=\frac{1}{m}\sum_{k=1}^{m}\frac{R^2((a_k\cdot\hat{u})^4+\beta_1(a_k\cdot\hat{u})^6)+2R\beta_2(a_k\cdot\hat{u})^4X_k^2-X_k^4-(\beta_1+2\beta_2)(a_k\cdot\hat{u})^2X_k^4}{(R+R\beta_1(a_k\cdot\hat{u})^2+\beta_2X_k^2)^2}; \quad (4.3)$$

$$\partial_{RR} f=2\frac{1}{m}\sum_{k=1}^{m}\frac{\left(1+(a_k\cdot\hat{u})^2(\beta_1+\beta_2)\right)X_k^4}{(R+R\beta_1(a_k\cdot\hat{u})^2+\beta_2X_k^2)^3}. \quad (4.4)$$

Lemma 4.2 (The regimes $\|u\|_2 \gg 1$ or $\|u\|_2 \ll 1$ are OK)**.** *There exist constants $R_i = R_i(\beta_1,\beta_2) > 0$, $d_i = d_i(\beta_1,\beta_2) > 0$, $i = 1,2$ such that the following hold: For $m \gtrsim n$, with high probability it holds that*

$$\begin{aligned} &\partial_R f \geq d_1, && \forall R \geq R_1, && \forall \hat{u} \in \mathbb{S}^{n-1}; \\ &\partial_R f \leq -d_2 < 0, && \forall 0 < R \leq R_2, && \forall \hat{u} \in \mathbb{S}^{n-1}. \end{aligned}$$

Proof. Denote $Z_k = a_k \cdot \hat{u}$. We first consider the regime $R \gg 1$. Observe that

$$\frac{1}{m}\sum_{k=1}^{m} \frac{R^2 Z_k^4}{(R + R\beta_1 Z_k^2 + \beta_2 X_k^2)^2} \gtrsim \frac{1}{m}\sum_{k=1}^{m} \frac{Z_k^4}{(1 + Z_k^2 + X_k^2)^2} \gtrsim 1, \quad \forall \hat{u} \in \mathbb{S}^{n-1},$$

where the last inequality holds for $m \gtrsim n$ with high probability. On the other hand we note that

$$\begin{aligned} \frac{1}{m}\sum_{k=1}^{m} \frac{X_k^4}{(R + R\beta_1 Z_k^2 + \beta_2 X_k^2)^2} &\lesssim \frac{1}{m}\sum_{k=1}^{m} \frac{X_k^4}{(R + X_k^2)^2} (\chi_{|X_k| \leq R^{\frac{1}{4}}} + \chi_{|X_k| > R^{\frac{1}{4}}}) \\ &\lesssim R^{-1} + \frac{1}{m}\sum_{k=1}^{m} \chi_{|X_k| > R^{\frac{1}{4}}} \ll 1, \quad \forall \hat{u} \in \mathbb{S}^{n-1}, \end{aligned}$$

where again the last inequality holds for R sufficiently large, and for $m \gtrsim n$ with high probability. Similarly we have for R sufficiently large,

$$\begin{aligned} &\frac{1}{m}\sum_{k=1}^{m} \frac{(\beta_1 + 2\beta_2) Z_k^2 X_k^4}{(R + R\beta_1 Z_k^2 + \beta_2 X_k^2)^2} \\ \lesssim &R^{-1} \frac{1}{m}\sum_{k=1}^{m} Z_k^2 + \frac{1}{m}\sum_{k=1}^{m} Z_k^2 \chi_{|X_k| > R^{\frac{1}{4}}} \ll 1, \quad \forall \hat{u} \in \mathbb{S}^{n-1}. \end{aligned}$$

Thus it follows easily that $\partial_R f \gtrsim 1$ for $R \gg 1$.

Now we turn to the regime $0 < R \ll 1$. First we note that the main negative term is OK. This is due to the fact that for $0 < R \leq 1$, we have (for $m \gtrsim n$ and with high probability)

$$\frac{1}{m}\sum_{k=1}^{m} \frac{Z_k^2 X_k^4}{(R + R Z_k^2 + X_k^2)^2} \geq \frac{1}{m}\sum_{k=1}^{m} \frac{Z_k^2 X_k^4}{(1 + Z_k^2 + X_k^2)^2} \gtrsim 1, \quad \forall \hat{u} \in \mathbb{S}^{n-1}.$$

On the other hand, we have (for $m\gtrsim n$ and with high probability)

$$\begin{aligned}
&\frac{1}{m}\sum_{k=1}^{m}\frac{R^2(Z_k^4+Z_k^6)}{(R+RZ_k^2+X_k^2)^2}\\
\leq&\frac{1}{m}\sum_{k=1}^{m}\frac{RZ_k^4}{R+RZ_k^2+X_k^2}\cdot(\chi_{|X_k|\geq R^{\frac{1}{4}}|Z_k|}+\chi_{|X_k|<R^{\frac{1}{4}}|Z_k|})\\
\leq&R^{\frac{1}{2}}\frac{1}{m}\sum_{k=1}^{m}Z_k^2+\frac{1}{m}\sum_{k=1}^{m}Z_k^2\chi_{|X_k|<R^{\frac{1}{4}}|Z_k|}\\
\leq&R^{\frac{1}{2}}\frac{1}{m}\sum_{k=1}^{m}Z_k^2+\frac{1}{m}\sum_{k=1}^{m}Z_k^2\chi_{|Z_k|\geq K}+\frac{1}{m}\sum_{k=1}^{m}K^2\chi_{|X_k|<KR^{\frac{1}{4}}}\\
\ll&1,\quad\forall\hat{u}\in\mathbb{S}^{n-1},
\end{aligned}$$

if we first take K sufficiently large followed by taking R sufficiently small. The estimate of the other term $\frac{RZ_k^4X_k^2}{(R+RZ_k^2+X_k^2)^2}$ is similar and we omit further details.

Collecting the estimates, it is then clear that we can obtain the desired estimate for $\partial_R f$ when $0<R\ll 1$. □

4.2 The regime $\|u\|_2\sim 1$

Lemma 4.3 (The regime $\|u\|_2\sim 1$ with $\epsilon_0\leq|\hat{u}\cdot e_1|\leq 1-\epsilon_0$ is OK)**.** *Let $0<\epsilon_0\ll 1$ be given. Assume $0<c_1<c_2<\infty$ are two given constants. Then for $m\gtrsim n$, the following hold with high probability: The loss function $f=f(u)$ has no critical points in the regime:*

$$\Big\{u=\sqrt{R}\hat{u}\colon\ c_1<R<c_2,\ \epsilon_0\leq|\hat{u}\cdot e_1|\leq 1-\epsilon_0\Big\}.$$

More precisely, introduce the parametrization $\hat{u}=e_1\cos\theta+e^{\perp}\sin\theta$, where $\theta\in[0,\pi]$ and $e^{\perp}\in\mathbb{S}^{n-1}$ satisfies $e^{\perp}\cdot e_1=0$. Then in the aforementioned regime, we have

$$|\partial_\theta f|\geq\alpha_1>0,$$

where α_1 depends only on $(\beta,\ \epsilon_0,\ c_1,\ c_2)$.

Proof. We first recall

$$f(u)=\frac{1}{m}\sum_{k=1}^{m}\frac{\Big(R(a_k\cdot\hat{u})^2-X_k^2\Big)^2}{R+\beta_1R(a_k\cdot\hat{u})^2+\beta_2X_k^2}.$$

Clearly $a_k\cdot\hat{u}=X_k\cos\theta+(a_k\cdot e^{\perp})\sin\theta$, and

$$\begin{aligned}&\partial_\theta(a_k\cdot\hat{u})=X_k(-\sin\theta)+(a_k\cdot e^{\perp})\cos\theta;\\&\partial_{\theta\theta}(a_k\cdot\hat{u})=-(a_k\cdot\hat{u}).\end{aligned}$$

In particular, if θ is away from the end-points $0,\pi$, then

$$\partial_\theta(a_k\cdot\hat{u})=(a_k\cdot\hat{u})\cot\theta-X_k\csc\theta.$$

We then obtain (below $Z_k=a_k\cdot\hat{u}$)

$$\begin{aligned}\partial_\theta f=&-\csc\theta\frac{1}{m}\sum_{k=1}^{m}\frac{2RZ_k(-X_k^2+RZ_k^2)\cdot\Big((\beta_1+2\beta_2)X_k^2+R(2+\beta_1Z_k^2)\Big)X_k}{(R+\beta_1RZ_k^2+\beta_2X_k^2)^2}\\&+\cot\theta\frac{1}{m}\sum_{k=1}^{m}\frac{2RZ_k(-X_k^2+RZ_k^2)\cdot\Big((\beta_1+2\beta_2)X_k^2+R(2+\beta_1Z_k^2)\Big)Z_k}{(R+\beta_1RZ_k^2+\beta_2X_k^2)^2}.\end{aligned}$$

Thanks to the strong damping, it is not difficult to check that for any $\epsilon>0$, if $m\gtrsim n$, then with high probability we have

$$|\partial_\theta f-\mathbb{E}\partial_\theta f|\le\epsilon,\quad\forall c_1\le R\le c_2,\quad\forall\hat{u}\in\mathbb{S}^{n-1}.$$

The desired result then follows from Lemma C.1. □

Lemma 4.4 (The regime $\|u\|_2\sim1$ with $|\hat{u}\cdot e_1|\le\epsilon_0$ is OK)**.** *Let $0<\epsilon_1\ll1$ be a sufficiently small constant. Assume $0<c_1<c_2<\infty$ are two given constants. Then for $m\gtrsim n$, the following hold with high probability: Consider the regime*

$$\Big\{u=\sqrt{R}\hat{u}\colon c_1<R<c_2,\ |\hat{u}\cdot e_1|\le\epsilon_1\Big\}.$$

Introduce the parametrization $\hat{u}=e_1\cos\theta+e^{\perp}\sin\theta$, where $\theta\in[0,\pi]$ and $e^{\perp}\in\mathbb{S}^{n-1}$ satisfies $e^{\perp}\cdot e_1=0$. Then in the aforementioned regime, we have

$$\partial_{\theta\theta}f\le-\alpha_2<0,$$

where $\alpha_2>0$ depends only on $(\beta,\ \epsilon_1,\ c_1,\ c_2)$.

Proof. This is similar to the argument in the proof of Lemma 4.3. By a tedious computation, we have

$$\partial_{\theta\theta}f=\frac{1}{m}\sum_{k=1}^{m}\frac{2RG_k}{(R+\beta_2x^2+\beta_1RZ_k^2)^3},$$

where

$$\begin{aligned}G_k=&-8\beta_1RZ_k^2\left(-X_k^2+RZ_k^2\right)\left(R+\beta_2X_k^2+\beta_1RZ_k^2\right)(X_k-Z_k\cos\theta)^2\csc^2\theta\\&-2\left(R+\beta_2X_k^2+\beta_1RZ_k^2\right)^2\Big(X_k^4-3RX_k^2Z_k^2-RZ_k^4-2X_kZ_k\left(X_k^2-3RZ_k^2\right)\cos\theta\\&+Z_k^2(X_k^2-2RZ_k^2)\cos2\theta\Big)\csc^2\theta\\&+\beta_1\left(X_k^2-RZ_k^2\right)^2\Big(Z_k^2\left(R+\beta_2X_k^2+\beta_1RZ_k^2\right)+4\beta_1RZ_k^2(X_k-Z_k\cos\theta)^2\csc^2\theta\\&-\left(R+\beta_2X_k^2+\beta_1RZ_k^2\right)(X_k-Z_k\cos\theta)^2\csc^2\theta\Big).\end{aligned}$$

It is then tedious but not difficult to check that that for any $\epsilon>0$, if $m\gtrsim n$, then with high probability we have

$$|\partial_{\theta\theta}f-\mathbb{E}\partial_{\theta\theta}f|\le\epsilon,\quad\forall c_1\le R\le c_2,\quad\forall\hat{u}\in\mathbb{S}^{n-1}.$$

The desired result then follows from Lemma C.1. □

Theorem 4.2 (The regime $\|u\|_2\sim1$, $||\hat{u}\cdot e_1|-1|\le\epsilon_0$, $|\|u\|_2-1|\ge c(\epsilon_0)$ is OK)**.** *Let $0<c_1<1<c_2<\infty$ be given constants. Let $0<\epsilon_0\ll1$ be a given sufficiently small constant and consider the regime $\left||\hat{u}\cdot e_1|-1\right|\le\epsilon_0$ with $c_1\le\|u\|_2^2\le c_2$. There exists a constant $c_0=c_0(\epsilon_0,c_1,c_2,\beta)>0$ which tends to zero as $\epsilon_0\to0$ such that the following hold: For $m\gtrsim n$, with high probability it holds that (below $u=\sqrt{R}\hat{u}$)*

$$\begin{aligned}&\partial_Rf<0,\qquad\forall c_2\le R\le1-c_0,\quad\forall\hat{u}\in\mathbb{S}^{n-1}\quad\textit{with}\quad||\hat{u}\cdot e_1|-1|\le\epsilon_0;\\&\partial_Rf>0,\qquad\forall1+c_0\le R\le c_1,\quad\forall\hat{u}\in\mathbb{S}^{n-1}\quad\textit{with}\quad||\hat{u}\cdot e_1|-1|\le\epsilon_0.\end{aligned}$$

Proof. We rewrite

$$f(u)=\frac{1}{m}\sum_{k=1}^m g(R,(a_k\cdot\hat{u})^2,X_k^2),$$

where

$$g(R,a,b)=\frac{(Ra-b)^2}{R+\beta_1Ra+\beta_2b}.$$

It is not difficult to check that for $R\sim1$, we have

$$|(\partial_Rg)(R,a,b)-(\partial_Rg)(R,b,b)|\le\|\partial_{Ra}g\|_\infty|b-a|\lesssim|b-a|,\quad\forall a,b\ge0.$$

On the other hand, note that $(\partial_Rg)(1,b,b)=0$, and for $R\sim1$,

$$(\partial_{RR}g)(R,b,b)=\frac{2b^2(1+b(\beta_1+\beta_2))^2}{(R+b(\beta_2+\beta_1R))^3}\sim b.$$

Thus for $R=1+\eta$, $\eta>0$ we have

$$\begin{aligned}(\partial_R g)(R,a,b) &\geq \partial_R g(R,b,b)-\gamma_1|b-a| \\ &\geq \gamma_2\cdot\eta\cdot b-\gamma_1|b-a|,\end{aligned}$$

where $\gamma_1>0$, $\gamma_2>0$ are constants depending only on $(\beta_1,\ \beta_2,\ c_1,\ c_2)$. The desired result (for $\partial_R f>0$ when $R\to 1+$) then follows from this and simple application of Bernstein's inequalities. The estimate for the regime $R\to 1-$ is similar. We omit the details. □

Theorem 4.3 (Strong convexity near the global minimizer)**.** *There exist* $0<\epsilon_0\ll 1$ *and a positive constant* γ *such that if* $m\gtrsim n$*, then the following hold with high probability:*

1. *If* $\|u-e_1\|_2\leq\epsilon_0$*, then*

$$\sum_{i,j=1}^{n}\xi_i\xi_j(\partial_{ij}f)(u)\geq\gamma>0,\quad\forall\xi\in\mathbb{S}^{n-1}.$$

2. *If* $\|u+e_1\|_2\leq\epsilon_0$*, then*

$$\sum_{i,j=1}^{n}\xi_i\xi_j(\partial_{ij}f)(u)\geq\gamma>0,\quad\forall\xi\in\mathbb{S}^{n-1}.$$

In yet other words, $f(u)$ *is strongly convex in a sufficiently small neighborhood of* $\pm e_1$*.*

Proof. See appendix. □

Finally we complete the proof of Theorem 4.1.

Proof of Theorem 4.1. We proceed in several steps. All the statements below hold under the assumption that $m\gtrsim n$ and with high probability.

1. For $u=0$, we use Lemma 4.1. In particular $u=0$ is a local maximum point with strictly negative Hessian.

2. For $\|u\|_2\ll 1$ or $\|u\|_2\gg 1$, we use Lemma 4.2. The loss functions has a nonzero gradient $(\partial_R f\neq 0)$ in this regime.

3. For $\|u\|_2\sim 1$ with $\epsilon_0\leq|\hat{u}\cdot e_1|\leq 1-\epsilon_0$, we use Lemma 4.3 to show that the loss function has a nonzero gradient $(\partial_\theta f\neq 0)$ in this regime.

4. For $\|u\|_2 \sim 1$ with $|\hat{u}\cdot e_1| \leq \epsilon_0$, by Lemma 4.4, the loss function has a negative curvature direction (i.e., $\partial_{\theta\theta} f < 0$) in this regime.

5. For $\|u\|_2 \sim 1$, $||\hat{u}\cdot e_1| - 1| \leq \epsilon_0$, $|\|u\|_2 - 1| \geq c(\epsilon_0)$, Theorem 4.2 shows that the gradient of the loss function does not vanish (i.e., $\partial_R f \neq 0$).

6. For $\|u \pm e_1\| \ll 1$, Theorem 4.3 gives the strong convexity in the full neighborhood.

It is not difficult to check that the above 6 scenarios cover the whole of $\mathbb{R}^n$. We omit further details. □

5 Numerical experiments

In this section, we demonstrate the numerical efficiency of our estimators by simple gradient descent and compare their performance with other competitive algorithms. Our Quotient intensity models are:

QIM1:

$$\min_{u\in\mathbb{R}^n} f(u) = \frac{1}{m}\sum_{k=1}^{m} \frac{((a_k\cdot u)^2 - (a_k\cdot x)^2)^2}{(a_k\cdot x)^2}.$$

QIM2:

$$\min_{u\in\mathbb{R}^n} f(u) = \frac{1}{m}\sum_{k=1}^{m} \frac{((a_k\cdot u)^2 - (a_k\cdot x)^2)^2}{\beta\|u\|_2^2 + (a_k\cdot x)^2}.$$

QIM3:

$$\min_{u\in\mathbb{R}^n} f(u) = \frac{1}{m}\sum_{k=1}^{m} \frac{((a_k\cdot u)^2 - (a_k\cdot x)^2)^2}{\|u\|_2^2 + \beta_1(a_k\cdot u)^2 + \beta_2(a_k\cdot x)^2}.$$

We have shown theoretically that any gradient descent algorithm will not get trapped in a local minimum for the estimators above. Here we present numerical experiments to show that the estimators perform very well with randomized initial guess.

We test the performance of our QIM2 and QIM3 and compare with SAF [5], Trust Region [29], WF [3], TWF [7] and TAF [34]. Here, it is worth emphasizing that random initialization is used for SAF, Trust Region [29] and our QIM2, QIM3 algorithms while all other algorithms have adopted a spectral initialization.

5.1 Recovery of 1D signals

In our numerical experiments, the target vector $x\in\mathbb{R}^n$ is chosen randomly from the standard Gaussian distribution and the measurement vectors a_i, $i=1,\cdots,m$ are generated randomly from standard Gaussian distribution or CDP model. For the real Gaussian case, the signal $x\sim\mathcal{N}(0,I_n)$ and measurement vectors $a_i\sim\mathcal{N}(0,I_n)$ for $i=1,\cdots,m$. For the complex Gaussian case, the signal $x\sim\mathcal{N}(0,I_n)+i\mathcal{N}(0,I_n)$ and measurement vectors $a_i\sim\mathcal{N}(0,I_n/2)+i\mathcal{N}(0,I_n/2)$. For the CDP model, we use masks of octanary patterns as in [3]. For simplicity, our parameters and step size are fixed for all experiments. Specifically, we adopt parameter $\beta=1$ and step size $\mu=0.4$ for QIM2 and choose the parameter $\beta_1=0.1$, $\beta_2=1$, step size $\mu=0.3$ for QIM3. For Trust Region, WF, TWF and TAF, we use the codes provided in the original papers with suggested parameters.

Example 5.1. In this example, we test the empirical success rate of QIM2, QIM3 versus the number of measurements. We conduct the experiments for the real Gaussian, complex Gaussian and CDP cases, respectively. We choose $n=128$ and the maximum number of iterations is $T=2500$. For real and complex Gaussian cases, we vary m within the range $[n,10n]$. For CDP case, we set the ratio $m/n=L$ from 2 to 10. For each m, we run 100 times trials to calculate the success rate. Here, we say a trial to have successfully reconstructed the target signal if the relative error satisfies $\text{dist}(u_T-x)/\|x\|\leq 10^{-5}$. The results are plotted in Fig. 1. It can be seen that $6n$ Gaussian phaseless measurement or 7 octanary patterns are enough for exactly recovery for QIM2 and QIM3.

Example 5.2. In this example, we compare the convergence rate of QIM2, QIM3 with those of SAF, WF, TWF, TAF for real Gaussian and complex Gaussian cases. We choose $n=128$ and $m=6n$. The results are presented in Fig. 2. We can see that our algorithms perform well comparing with state-of-the-art algorithms with spectral initialization.

Example 5.3. In this example, we compare the time elapsed and the iteration needed for WF, TWF, TAF, SAF and our QIM2, QIM3 to achieve the relative error 10^{-5} and 10^{-10}, respectively. We choose $n=1000$ with $m=8n$. We adopt the same spectral initialization method for WF, TWF, TAF and the initial guess is obtained by power method with 50 iterations. We run 50 times trials to calculate the average time elapsed and iteration number for those algorithms. The results are shown in Table 1. The numerical results show that QIM3 takes around 27 and 50 iterations to escape the saddle points for the real and complex Gaussian cases, respectively.

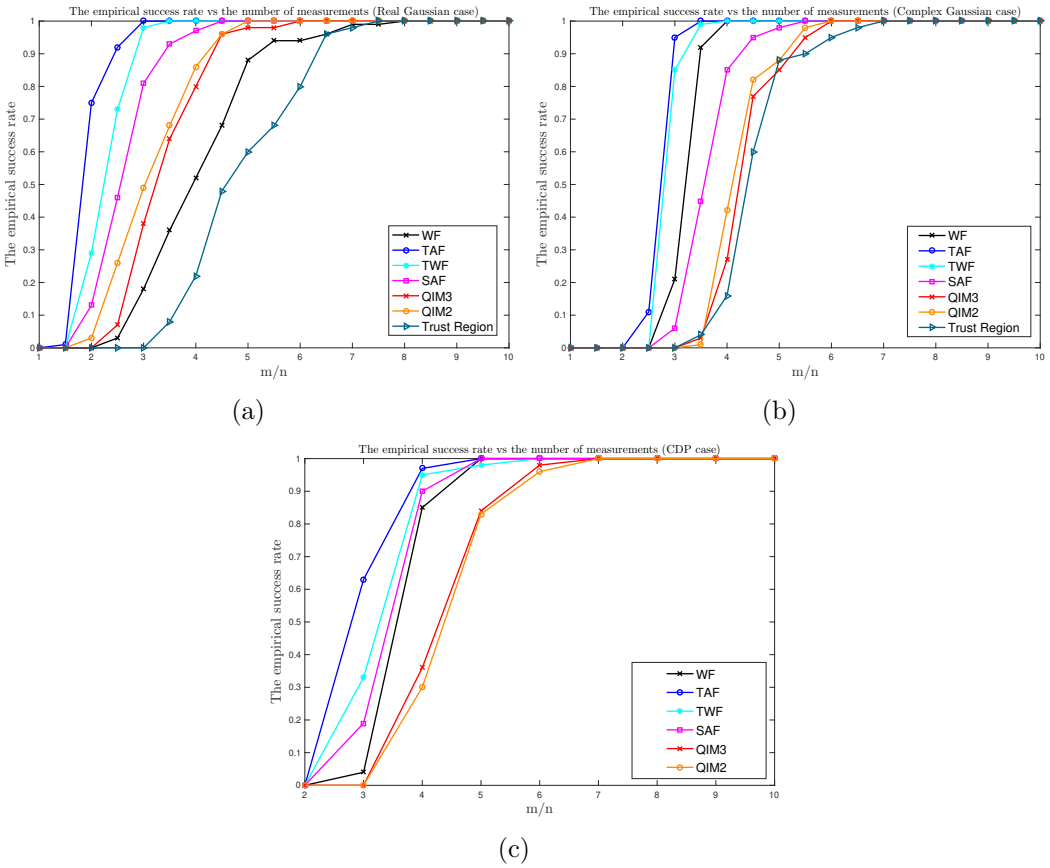
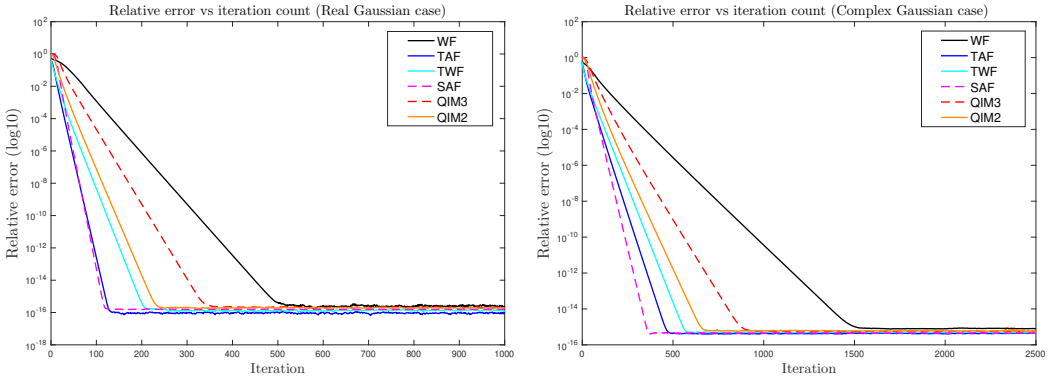


Figure 1: The empirical success rate for different m/n based on 100 random trails. (a) Success rate for real Gaussian case, (b) Success rate for complex Gaussian case, (c) Success rate for CDP case.

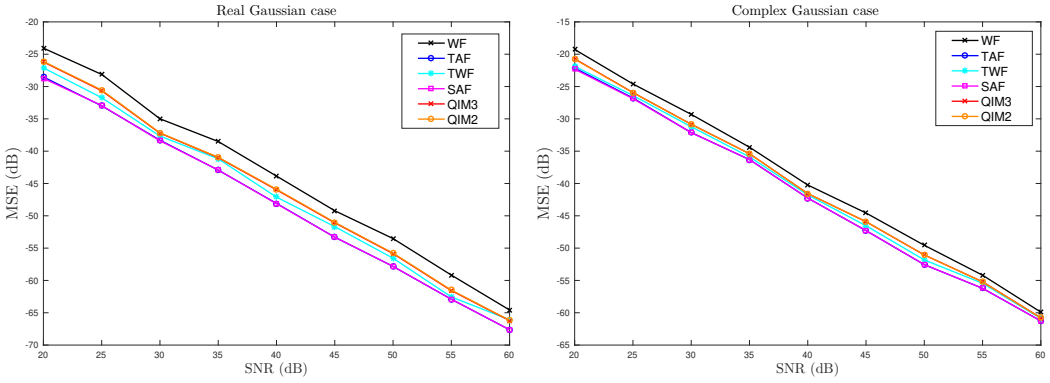
Table 1: Time elapsed and iteration number among algorithms on Gaussian signals with $n=1000$.

Algorithm	Real Gaussian				Complex Gaussian			
	10^{-5}		10^{-10}		10^{-5}		10^{-10}	
	Iter	Time(s)	Iter	Time(s)	Iter	Time(s)	Iter	Time(s)
SAF	44	**0.1556**	68	**0.2276**	113	**1.3092**	190	**2.3596**
QIM2	58	2.0589	117	3.7204	155	21.6235	314	37.1972
QIM3	88	2.4423	161	4.2229	211	30.2235	422	48.1972
WF	125	4.4214	229	6.3176	304	34.6266	655	86.6993
TAF	29	0.2744	60	0.3515	100	1.7704	211	2.7852
TWF	40	0.3181	87	0.4274	112	1.9808	244	3.7432
Trust Region	**21**	2.9832	**29**	4.4683	**33**	19.1252	**42**	29.0338



(a) (b)

Figure 2: Relative error versus number of iterations for QIM, SAF, WF, TWF, and TAF method: (a) Real-valued signals; (b) Complex-valued signals.



(a) (b)

Figure 3: SNR versus relative MSE on a dB-scale under the noisy Gaussian model: (a) Real Gaussian case; (b) Complex Gaussian case.

5.2 Recovery of natural image

We next compare the performance of the above algorithms on recovering a natural image from masked Fourier intensity measurements. The image is the Milky Way Galaxy with resolution 1080×1920. The colored image has RGB channels. We use $L=20$ random octanary patterns to obtain the Fourier intensity measurements for each R/G/B channel as in [3]. Table 2 lists the averaged time elapsed and the iteration needed to achieve the relative error 10^{-5} and 10^{-10} over the three RGB channels. We can see that our algorithms have good performance comparing with state-of-the-art algorithms with spectral initialization.

Table 2: Time elapsed and iteration number among algorithms on recovery of galaxy image.

Algorithm	The Milky Way Galaxy			
	10^{-5}		10^{-10}	
	Iter	Time(s)	Iter	Time(s)
SAF	92	**202.47**	148	**351.21**
QIM2	168	351.32	282	601.68
QIM3	173	371.59	296	709.21
WF	158	381.7	277	621.63
TAF	**65**	223.89	**122**	368.22
TWF	68	315.14	145	566.84

5.3 Recovery of signals with noise

We now demonstrate the robustness of QIM2, QIM3 to noise and compare them with SAF, WF, TWF, TAF. We consider the noisy model $y_i=|\langle a_i,x\rangle|+\eta_i$ and add different level of Gaussian noises to explore the relationship between the signal-to-noise rate (SNR) of the measurements and the mean square error (MSE) of the recovered signal. Specifically, SNR and MSE are evaluated by

$$\mathrm{MSE}:=10\log_{10}\frac{\mathrm{dist}^2(u,x)}{\|x\|^2}\quad\text{and}\quad\mathrm{SNR}=10\log_{10}\frac{\sum_{i=1}^m|a_i^\top x|^2}{\|\eta\|^2},$$

where u is the output of the algorithms given above after 2500 iterations. We choose $n=128$ and $m=8n$. The SNR varies from 20db to 60db. The result is shown in Fig. 3. We can see that our algorithms are stable for noisy phase retrieval.

Appendix A: technical estimates for Section 2

Lemma A.1. *Let $\phi\in C_c^\infty(\mathbb{R})$ satisfies $0\le\phi(x)\le1$ for all x, $\phi(x)=1$ for $|x|\le1$ and $\phi(x)=0$ for $|x|\ge2$. There exist $\epsilon>0$ sufficiently small, and N sufficiently large such that*

$$\mathbb{E}\frac{(a\cdot\xi)^2(a\cdot e_1)^2}{\epsilon+(a\cdot e_1)^2}\phi\Big(\frac{a\cdot\xi}{N}\Big)\ge0.99,\quad\forall\xi\in\mathbb{S}^{n-1},$$

where $a\sim\mathcal{N}(0,\mathrm{I_n})$.

Proof. We first show that there exist $\epsilon>0$, such that

$$\mathbb{E}\frac{(a\cdot\xi)^2(a\cdot e_1)^2}{\epsilon+(a\cdot e_1)^2}\ge0.995,\quad\forall\xi\in\mathbb{S}^{n-1}.\tag{A.1}$$

Clearly it suffices for us to show

$$\sup_{\xi\in\mathbb{S}^{n-1}}\mathbb{E}\frac{\epsilon(a\cdot\xi)^2}{\epsilon+(a\cdot e_1)^2}\to 0 \quad \text{as} \quad \epsilon\to 0. \tag{A.2}$$

Observe that $\xi=se_1+\sqrt{1-s^2}e_1^{\perp}$, $|s|\le 1$, $e^{\perp}\cdot e_1=0$. Thus denoting X and Y as two independent standard Gaussian random variables with mean zero and unit variance, we have

$$\sup_{\xi\in\mathbb{S}^{n-1}}\mathbb{E}\frac{\epsilon(a\cdot\xi)^2}{\epsilon+(a\cdot e_1)^2}\lesssim\mathbb{E}\frac{\epsilon X^2}{\epsilon+X^2}+\mathbb{E}\frac{\epsilon Y^2}{\epsilon+X^2}\lesssim\epsilon+\mathbb{E}\frac{\epsilon}{\epsilon+X^2}\lesssim\sqrt{\epsilon},$$

where in the last inequality we used the fact that

$$\int_{|x|\le 1}\frac{\epsilon}{\epsilon+x^2}dx\sim\sqrt{\epsilon}.$$

Thus (A.2) and (A.1) hold. Now ϵ is fixed. To show the final inequality, we note that

$$\mathbb{E}\frac{(a\cdot\xi)^2(a\cdot e_1)^2}{\epsilon+(a\cdot e_1)^2}\chi_{|a\cdot\xi|\ge N}\le\mathbb{E}(a\cdot\xi)^2\chi_{|a\cdot\xi|\ge N}\le\mathbb{E}X^2\chi_{|X|\ge N}\to 0,$$

as N tend to infinity. Thus the desired inequality easily follows. □

Lemma A.2. *Let $0<\eta_0\ll 1$ be given. Then if $m\gtrsim n$, then the following hold with high probability:*

$$\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^4}{(a_k\cdot e_1)^2}\ge 100,\quad \forall\hat{u}\in\mathbb{S}^{n-1}\quad \textit{with}\quad ||\hat{u}\cdot e_1|-1|\ge\eta_0.$$

Proof. Without loss of generality we write

$$\hat{u}=se^{\perp}\pm\sqrt{1-s^2}e_1,\quad e^{\perp}\in\mathbb{S}^{n-1}\quad\text{with}\quad e^{\perp}\cdot e_1=0.$$

Clearly $|s|\ge s_0=s_0(\eta_0)>0$, where $s_0(\eta_0)$ is a constant depending only on η_0. Take $a\sim\mathcal{N}(0,\mathrm{I_n})$ and observe that

$$\begin{aligned}
&\mathbb{E}\frac{(a\cdot\hat{u})^4}{\epsilon(1+(a\cdot e^{\perp})^2)+(a\cdot e_1)^2}\\
\ge&\mathbb{E}\frac{s^4(a\cdot e^{\perp})^4}{\epsilon(1+(a\cdot e^{\perp})^2)+(a\cdot e_1)^2}\\
\ge&s_0^4\frac{1}{2\pi}\int_{1\le y\le 2,x\in\mathbb{R}}\frac{y^4}{\epsilon(1+y^2)+x^2}e^{-\frac{x^2+y^2}{2}}dxdy\\
\ge&s_0^4\frac{1}{200}\int_{|x|\le 1}\frac{1}{5\epsilon+x^2}dx\ge s_0^4\cdot\mathcal{O}(\epsilon^{-\frac{1}{2}})\ge 200,
\end{aligned}$$

if $\epsilon>0$ is taken sufficiently small. Now we fix this ϵ. Clearly for $m\gtrsim n$ with high probability it holds that

$$\begin{aligned}\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^4}{(a_k\cdot e_1)^2}\geq&\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^4}{\epsilon(1+(a_k\cdot e^{\perp})^2)+(a_k\cdot e_1)^2}\\ \geq&100,\quad\forall\hat{u}\in\mathbb{S}^{n-1}\quad\text{with}\quad||\hat{u}\cdot e_1|-1|\leq\eta_0.\end{aligned}$$

Thus, we complete the proof. □

Appendix B: technical estimates for Section 3

Lemma B.1. *For any $\epsilon>0$, there exists $R_0=R_0(\beta,\epsilon)>0$ sufficiently small, such that if $m\gtrsim n$, then the following hold with high probability:*

$$\frac{R}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^2}{\beta R+(a_k\cdot e_1)^2}<\epsilon,\quad\forall\hat{u}\in\mathbb{S}^{n-1},\quad\forall 0<R\leq R_0.$$

Proof. Let $\phi\in C_c^{\infty}(\mathbb{R})$ be such that $0\leq\phi(x)\leq 1$ for all x, $\phi(x)=1$ for $|x|\leq 1$ and $\phi(x)=0$ for $|x|\geq 2$. We then split the sum as

$$\frac{R}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})^2}{\beta R+(a_k\cdot e_1)^2}\leq\frac{1}{\beta m}\sum_{k=1}^{m}(a_k\cdot\hat{u})^2\phi\Big(\frac{a_k\cdot e_1}{\eta_0}\Big)+R\cdot\eta_0^{-2}\frac{1}{m}\sum_{k=1}^{m}(a_k\cdot\hat{u})^2.$$

Clearly the first term is amenable to union bounds, and we can make it sufficiently small with high probability by taking η_0 small (depending only on β and ϵ). The second term is trivial since we can take R sufficiently small. Thus we complete the proof. □

Lemma B.2. *There exists $R_1=R_1(\beta)>0$ sufficiently small, such that if $m\gtrsim n$, then the following hold with high probability:*

$$c_1\leq\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})(a_k\cdot e_1)^3}{\beta R+(a_k\cdot e_1)^2}\leq c_2,\quad\forall\hat{u}\in\mathbb{S}^{n-1}\quad\textit{with}\quad\widehat{u}_1\cdot e_1\geq\frac{1}{10},\quad\forall 0<R\leq R_1.$$

In the above c_1, $c_2>0$ are constants depending only on β.

Proof. Denote $X_k=a_k\cdot e_1$. Write $\hat{u}=se_1+\sqrt{1-s^2}e^{\perp}$, where $s\geq\frac{1}{10}$ and $e^{\perp}\in\mathbb{S}^{n-1}$

satisfies $e^{\perp}\cdot e_1=0$. We then write

$$\begin{aligned}&\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot\hat{u})X_k^3}{\beta R+X_k^2}\\ =&s\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{\beta R+X_k^2}+\sqrt{1-s^2}\frac{1}{m}\sum_{k=1}^{m}\frac{(a_k\cdot e^{\perp})X_k^3}{\beta R+X_k^2}\\ =&s\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{\beta R+X_k^2}+\sqrt{1-s^2}\frac{1}{m}\sum_{k=1}^{m}(a_k\cdot e^{\perp})X_k-\sqrt{1-s^2}\frac{1}{m}\sum_{k=1}^{m}(a_k\cdot e^{\perp})\frac{\beta RX_k}{\beta R+X_k^2}.\end{aligned}$$

For the first term we note that for $0<R\leq 1$,

$$\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{\beta+X_k^2}\leq\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{\beta R+X_k^2}\leq\frac{1}{m}\sum_{k=1}^{m}X_k^2.$$

Thus we clearly have for all $0<R\leq 1$, $\frac{1}{10}\leq s\leq 1$, with high probability it holds that

$$2c_1\leq s\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{\beta R+X_k^2}\leq\frac{1}{2}c_2.$$

The second term is clearly OK for union bounds and with high probability it can be made sufficiently small. For the last term, observe that with high probability,

$$\frac{1}{m}\sum_{k=1}^{m}|a_k\cdot e^{\perp}|\frac{\beta R|X_k|}{\beta R+X_k^2}\lesssim\sqrt{\beta R}\frac{1}{m}\sum_{k=1}^{m}|a_k\cdot e^{\perp}|\ll 1,\quad\forall e^{\perp}\in\mathbb{S}^{n-1},$$

if $R\leq R_1$ and R_1 is sufficiently small. The desired result then clearly follows. □

Proof of Lemma 3.5. Without loss of generality we consider the situation $\hat{u}=e_1\cos\theta+e^{\perp}\sin\theta$ with $\epsilon_1\leq\theta\leq\frac{\pi}{2}-\epsilon_2$, where $0<\epsilon_1,\epsilon_2\ll 1$. The point is that θ stays away from the end-points 0 and $\frac{\pi}{2}$. Denote $X_k=a_k\cdot e_1$, $Y_k=a_k\cdot e^{\perp}$ and $Z_k=a_k\cdot\hat{u}$. Then

$$\begin{aligned}&Z_k=\cos\theta X_k+\sin\theta Y_k\quad\Rightarrow\quad Y_k=\frac{1}{\sin\theta}Z_k-\frac{\cos\theta}{\sin\theta}X_k;\\ &\partial_\theta Z_k=-\sin\theta X_k+\cos\theta Y_k=\cot\theta Z_k-\frac{1}{\sin\theta}X_k.\end{aligned}$$

We then obtain

$$\begin{aligned}\partial_\theta f=&4R^2\cot\theta\underbrace{\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^4}{\beta R+X_k^2}}_{=:H_0}-4R^2\csc\theta\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^3X_k}{\beta R+X_k^2}\\ &-4R\cot\theta\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^2X_k^2}{\beta R+X_k^2}+4R\csc\theta\frac{1}{m}\sum_{k=1}^{m}\frac{Z_kX_k^3}{\beta R+X_k^2}.\end{aligned}$$

Since $R\sim 1$, it is not difficult to check that the third and fourth terms above are amenable to union bounds†, i.e., with high probability (for $m\gtrsim n$) we have

$$\Big|\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^2X_k^2}{\beta R+X_k^2}-\text{mean}\Big|+\Big|\frac{1}{m}\sum_{k=1}^{m}\frac{Z_kX_k^3}{\beta R+X_k^2}-\text{mean}\Big|\ll 1,\quad \forall c_1\le R\le c_2,\quad \forall\hat{u}\in\mathbb{S}^{n-1}.$$

Next we treat the second term. Let $\phi\in C_c^\infty(\mathbb{R})$ be such that $0\le\phi(x)\le 1$ for all x, $\phi(x)=1$ for $|x|\le 1$ and $\phi(x)=0$ for $|x|\ge 2$. We have

$$\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^3X_k}{\beta R+X_k^2}=\underbrace{\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^3X_k}{\beta R+X_k^2}\phi\Big(\frac{Z_k}{M\langle X_k\rangle}\Big)}_{=:H_1}+\underbrace{\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^3X_k}{\beta R+X_k^2}\Big(1-\phi\Big(\frac{Z_k}{M\langle X_k\rangle}\Big)\Big)}_{=:H_2},$$

where $\langle z\rangle=(1+|z|^2)^{\frac{1}{2}}$. It is not difficult to check that H_1 is OK for union bounds, and with high probability it holds that

$$\Big|H_1-\mathbb{E}H_1\Big|\ll 1,\quad \forall\hat{u}\in\mathbb{S}^{n-1},\quad \forall c_1\le R\le c_2.$$

For H_2 we have (η_0 will be taken sufficiently small)

$$\begin{aligned}H_2&\le\eta_0\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^4}{\beta R+X_k^2}+\eta_0^{-3}\frac{1}{m}\sum_{k=1}^{m}\frac{X_k^4}{\beta R+X_k^2}\Big(1-\phi\Big(\frac{Z_k}{M\langle X_k\rangle}\Big)\Big)\\&\le\underbrace{\eta_0\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^4}{\beta R+X_k^2}}_{=:H_{2,a}}+\underbrace{\eta_0^{-3}\frac{1}{m}\sum_{k=1}^{m}X_k^2\Big(1-\phi\Big(\frac{Z_k}{M\langle X_k\rangle}\Big)\Big)}_{=:H_{2,b}}.\end{aligned}$$

We first take η_0 sufficiently small so that $H_{2,a}$ can be included in the estimate of H_0 without affecting too much the main order. On the other hand, once η_0 is fixed, we can take M sufficiently large such that

$$|H_{2,b}|+|\mathbb{E}H_{2,b}|\ll 1,\quad \forall\hat{u}\in\mathbb{S}^{n-1},\quad \forall c_1\le R\le c_2.$$

Finally we treat H_0. Clearly

$$H_0\ge\underbrace{\frac{1}{m}\sum_{k=1}^{m}\frac{Z_k^4}{\beta R+X_k^2}\phi\Big(\frac{Z_k}{K}\Big)}_{=:H_{0,a}}.$$

†The union bound includes covering in $\hat{u}$ and R.

By taking K large, it can be easily checked that

$$\sup_{\hat{u}\in\mathbb{S}^{n-1},\ c_1\le R\le c_2}|\mathbb{E}H_0-\mathbb{E}H_{0,a}|\ll 1.$$

On the other hand, for fixed K, clearly $H_{0,a}$ is OK for union bounds. It holds with high probability that

$$|H_{0,a}-\mathbb{E}H_{0,a}|\ll 1.$$

Collecting all the estimates, we obtain

$$\partial_\theta f\ge\mathbb{E}\partial_\theta f+\text{Error},$$

where $|\text{Error}|\ll 1$. The desired lower bound for $\partial_\theta f$ then easily follows from Lemma B.3 below. □

Lemma B.3. *Let $u=\sqrt{R}\hat{u}$ with $0<c_1\le R\le c_2<\infty$ and $\hat{u}\in\mathbb{S}^{n-1}$. Assume $\hat{u}=\cos\theta e_1+\sin\theta e^\perp$, where $\theta\in[0,\pi]$ and $e^\perp\in\mathbb{S}^{n-1}$ satisfies $e^\perp\cdot e_1=0$. We have*

$$\mathbb{E}f(u)=h(\beta,R,\cos^2\theta),$$

where

$$\max_{0\le s\le 1}\partial_s h(\beta,R,s)\le-\gamma_1<0,$$
$$\min_{0\le s\le 1}\partial_{ss} h(\beta,R,s)\ge\gamma_2>0.$$

Here $\gamma_i=\gamma_i(\beta,c_1,c_2)$, $i=1,2$ depend only on $(\beta,\ c_1,\ c_2)$. It follows that

$$\mathbb{E}\partial_\theta f=a_1(\beta,R,\cos^2\theta)\sin(2\theta);$$
$$\mathbb{E}\partial_{\theta\theta}f=2a_1(\beta,R,\cos^2\theta)\cos(2\theta)+a_2(\beta,R,\theta)\sin^2(2\theta),$$

where

$$\gamma_3<a_i(\beta,R,s)\le\gamma_4,\quad\forall s\in[0,1],\quad i=1,2;$$

and $\gamma_3>0$, $\gamma_4>0$ are constants depending only on $(\beta,\ c_1,\ c_2)$.

Proof. We have

$$\begin{aligned}\mathbb{E}f(u)&=\frac{1}{2\pi}\int_{\mathbb{R}^2}\frac{(R(x\cos\theta+y\sin\theta)^2-x^2)^2}{\beta R+x^2}e^{-\frac{x^2+y^2}{2}}dxdy\\&=\frac{1}{\pi}\int_0^\infty\frac{1}{\beta R+x^2}e^{-\frac{x^2}{2}}\cdot\sqrt{2\pi}h_1(R,x,\cos^2\theta)dx,\end{aligned}$$

where

$$h_1(R,x,s)=3R^2-2Rx^2+x^4+s(6R^2-2Rx^2)(-1+x^2)+R^2s^2(3-6x^2+x^4).$$

Integrating further in x then gives

$$\mathbb{E}f(u)=\sqrt{2\pi}\cdot\frac{1}{\pi}\cdot R\Big(c_1s^2+2c_2s+c_3\Big),\quad s=\cos^2\theta,$$

where the value of c_3 is unimportant for us, and

$$c_1=R\int_0^\infty\frac{1}{\beta R+x^2}e^{-\frac{x^2}{2}}(3-6x^2+x^4)dx;$$
$$c_2=\int_0^\infty\frac{1}{\beta R+x^2}e^{-\frac{x^2}{2}}(3R-x^2)(-1+x^2)dx.$$

First we show that $c_2<0$. By a short computation, we have

$$c_2=\frac{3+\beta}{2\beta}\cdot\Big(\beta R\sqrt{2\pi}-e^{\frac{\beta R}{2}}\pi\sqrt{\beta R}(1+\beta R)\mathrm{Erfc}\Big(\sqrt{\frac{\beta R}{2}}\Big)\Big),$$

where

$$\mathrm{Erfc}(y)=\frac{2}{\sqrt{\pi}}\int_y^\infty e^{-t^2}dt.$$

We then reduce the matter to showing

$$y<e^{y^2}(1+2y^2)\int_y^\infty e^{-t^2}dt,\quad\forall y>0. \tag{B.1}$$

This follows easily from the usual bound on $\mathrm{Erfc}(y)$:

$$\frac{1}{y+\sqrt{y^2+2}}<\mathrm{Erfc}(y)\cdot e^{y^2}\cdot\frac{\sqrt{\pi}}{2}\le\frac{1}{y+\sqrt{y^2+\frac{4}{\pi}}},\quad\forall y>0. \tag{B.2}$$

Thus $c_2<0$.

Next we show that $c_1>0$. We have

$$2\beta c_1=-\sqrt{2\pi}\beta R(5+\beta R)+e^{\frac{\beta R}{2}}\pi\sqrt{\beta R}(3+\beta R(6+\beta R))\cdot\mathrm{Erfc}\Big(\sqrt{\frac{\beta R}{2}}\Big).$$

It amounts to checking

$$e^{y^2}\int_y^\infty e^{-t^2}dt>\frac{y(5+2y^2)}{3+4y^2(3+y^2)},\quad\forall y>0.$$

This follows from Lemma B.4 below.

Finally we show $c_1+c_2<0$. We have

$$\begin{aligned}&2(c_1+c_2)\\&=\sqrt{2\pi}R(-2+\beta-\beta R)-e^{\frac{\beta R}{2}}\pi\cdot(-\beta^{\frac{3}{2}}R^{\frac{5}{2}}+(\beta R)^{\frac{3}{2}}+\sqrt{\beta R}-3R\sqrt{\beta R})\mathrm{Erfc}\Big(\sqrt{\frac{\beta R}{2}}\Big).\end{aligned}$$

Denote $y=\sqrt{\frac{\beta R}{2}}>0$. We then reduce matters to showing

$$2y^2-2R(1+y^2)<e^{y^2}\cdot 2y\cdot(-2y^2R-3R+1+2y^2)\int_y^\infty e^{-t^2}dt.$$

Since we have shown (B.1), we then only need to check

$$1+y^2>e^{y^2}y(2y^2+3)\int_y^\infty e^{-t^2}dt.$$

This in turn follows from Lemma B.4.

Finally we consider the polynomial

$$\tilde{h}(s)=c_1s^2+2c_2s.$$

Since $\tilde{h}'(s)=2c_1s+2c_2$ and $\tilde{h}'(0)=2c_2<0$, $\tilde{h}'(1)=2c_1+2c_2<0$, we have $\tilde{h}'(s)<0$ for all $s\in[0,1]$. Since $c_1>0$, we have $\tilde{h}''(s)>0$. The desired result then easily follows. □

Lemma B.4 (Refined upper and lower bounds on the Complementary Error function)**.** *Let*

$$\mathrm{Erfc}(x)=\frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt \quad \textit{for} \quad x>0.$$

Then

$$\begin{aligned}&e^{x^2}\cdot\mathrm{Erfc}(x)\cdot\frac{\sqrt{\pi}}{2}>\frac{x(5+2x^2)}{3+4x^2(3+x^2)}, &&\forall x>0;\\&e^{x^2}\cdot\mathrm{Erfc}(x)\cdot\frac{\sqrt{\pi}}{2}<\frac{1+x^2}{x(3+2x^2)}, &&\forall x>0.\end{aligned}$$

Remark B.1. In the regime $y\geq 1$, one can check that the upper and lower bounds here are sharper than (B.2). One should also recall that the usual way to derive the lower bound in (B.2) through conditional expectation. Namely one can regard $e^{-y^2}/(\sqrt{\pi}\mathrm{Erfc}(y))$ as the conditional mean $\mu_1(y)=\mathbb{E}(X|X>y)$ where X has the p.d.f. $\frac{1}{\sqrt{\pi}}e^{-x^2}$. Then evaluating the variance $\mathbb{E}((X-\mu_1)^2|X>y)>0$ gives $y\mu_1+\frac{1}{2}-\mu_1^2>0$. This yields the upper bound for μ_1 which in turn is the desired lower bound in (B.2). An interesting question is to derive a sharper two-sided bounds via more careful conditioning. However we shall not dwell on this issue here.

Proof of Lemma B.4. We focus on the regime $x>1$. By performing successive simple change of variables, we have

$$g(x):=e^{x^2}\int_x^\infty e^{-t^2}dt=\int_0^\infty e^{-2xs}e^{-s^2}ds=\frac{1}{2x}\int_0^\infty e^{-s}e^{-(\frac{s}{2x})^2}ds$$
$$\sim\sum_{k=0}^\infty(-1)^k x^{-(2k+1)}\cdot\frac{1}{2}\cdot\frac{(2k)!}{4^k k!}\sim\sum_{k=0}^\infty(-1)^k x^{-(2k+1)}\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)_k,$$

where in the last line we adopted Pochhammer's symbol $(a)_n=a(a+1)\cdots(a+n-1)$. Note that the above is an asymptotic series, and it is not difficult to check that

$$\left|g(x)-\sum_{k=0}^m(-1)^k x^{-(2k+1)}\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)_k\right|\le x^{-2m-3}\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)_{m+1},\quad\forall m\ge1,\quad\forall x>0.$$

Moreover, if m is an even integer, then

$$g(x)<\sum_{k=0}^m(-1)^k x^{-(2k+1)}\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)_k,\quad\forall x>0;$$

and if m is odd, then

$$g(x)>\sum_{k=0}^m(-1)^k x^{-(2k+1)}\cdot\frac{1}{2}\cdot\left(\frac{1}{2}\right)_k,\quad\forall x>0.$$

Now taking $m=4$, we have

$$g(x)<\frac{1}{2}x^{-1}-\frac{1}{4}x^{-3}+\frac{3}{8}x^{-5}-\frac{15}{16}x^{-7}+\frac{105}{32}x^{-9}.$$

For $x\ge3$, it is not difficult to verify that

$$\frac{1}{2}x^{-1}-\frac{1}{4}x^{-3}+\frac{3}{8}x^{-5}-\frac{15}{16}x^{-7}+\frac{105}{32}x^{-9}<\frac{1+x^2}{x(3+2x^2)}.$$

Hence the upper bound is OK for $x\ge3$.

Next taking $m=5$, we have

$$g(x)>\frac{1}{2}x^{-1}-\frac{1}{4}x^{-3}+\frac{3}{8}x^{-5}-\frac{15}{16}x^{-7}+\frac{105}{32}x^{-9}-\frac{945}{64}x^{-11}.$$

It is not difficult to verify that for $x\ge4$, we have

$$\frac{1}{2}x^{-1}-\frac{1}{4}x^{-3}+\frac{3}{8}x^{-5}-\frac{15}{16}x^{-7}+\frac{105}{32}x^{-9}-\frac{945}{64}x^{-11}>\frac{x(5+2x^2)}{3+12x^2+4x^4}.$$

Hence the lower bound is OK for $x \geq 4$.

Finally for the regime $x \in [0,4]$, we use rigorous numerics to verify the inequality. Since we are on a compact interval, this can be done by a rigorous computation with controllable numerical errors. □

Proof of Lemma 3.6. Again denote $X_k = a_k \cdot e_1$ and $Z_k = a_k \cdot \hat{u}$. Without loss of generality we assume $\theta \in [\frac{\pi}{2} - \eta, \frac{\pi}{2} + \eta]$ for some sufficiently small $\eta > 0$. By a tedious computation, we have

$$\begin{aligned}\partial_{\theta\theta} f =& 4R^2(1+2\cos 2\theta)\csc^2\theta \frac{1}{m}\sum_{k=1}^m \frac{Z_k^4}{\beta R + X_k^2} - 24R^2(\cot\theta\csc\theta)\frac{1}{m}\sum_{k=1}^m \frac{X_k Z_k^3}{\beta R + X_k^2} \\ &+ 4R(\csc^2\theta)(3R - \cos 2\theta)\frac{1}{m}\sum_{k=1}^m \frac{Z_k^2 X_k^2}{\beta R + X_k^2} + 8R(\cot\theta\csc\theta)\frac{1}{m}\sum_{k=1}^m \frac{X_k^3 Z_k}{\beta R + X_k^2} \\ &- 4R\csc^2\theta \frac{1}{m}\sum_{k=1}^m \frac{X_k^4}{\beta R + X_k^2}.\end{aligned}$$

Note that the third, fourth and fifth terms are OK for union bounds. The second and the first term can be handled in a similar way as in the proof of Lemma 3.5. The only difference is that the sign is now negative in the regime $\theta \to \frac{\pi}{2}$. Using Lemma B.3 it follows that $\partial_{\theta\theta} f < 0$ in this regime. We omit the repetitive details. □

Proof of Theorem 3.4. Without loss of generality we consider the regime $\|u - e_1\|_2 \ll 1$. Before we work out the needed estimates for the restricted convexity, we explain the main difficulty in connection with the full Hessian matrix. Denote $X_k = a_k \cdot e_1$. Then for any $\xi \in \mathbb{S}^{n-1}$, we have

$$\begin{aligned} H_{\xi\xi} =& \sum_{i,j} \xi_i \xi_j (\partial_{u_i u_j} f)(u) \\ =& 12\frac{1}{m}\sum_{k=1}^m \frac{(a_k\cdot\xi)^2 (a_k\cdot u)^2}{\beta\|u\|_2^2 + X_k^2} && \text{(B.3)} \\ & - 4\frac{1}{m}\sum_{k=1}^m \frac{(a_k\cdot\xi)^2 X_k^2}{\beta\|u\|_2^2 + X_k^2} && \text{(B.4)} \\ & - 16\beta\frac{1}{m}\sum_{k=1}^m \frac{(a_k\cdot u)^3 (a_k\cdot\xi)(u\cdot\xi)}{(\beta\|u\|_2^2 + X_k^2)^2} && \text{(B.5)} \\ & + 16\beta\frac{1}{m}\sum_{k=1}^m \frac{X_k^2 (a_k\cdot u)(a_k\cdot\xi)(u\cdot\xi)}{(\beta\|u\|_2^2 + X_k^2)^2} && \text{(B.6)} \end{aligned}$$

$$-2\beta\frac{1}{m}\sum_{k=1}^{m}\frac{((a_k\cdot u)^2-X_k^2)^2}{(\beta\|u\|_2^2+X_k^2)^2} \tag{B.7}$$

$$+8\beta^2(\xi\cdot u)^2\frac{1}{m}\sum_{k=1}^{m}\frac{((a_k\cdot u)^2-X_k^2)^2}{(\beta\|u\|_2^2+X_k^2)^3}. \tag{B.8}$$

First observe that if $u=e_1$, then the Hessian can be controlled rather easily thanks to the damping $\beta\|u\|_2^2+X_k^2$.

On the other hand, for $u\neq e_1$, as far as the lower bound is concerned, the main difficult terms are (B.7) and (B.5) which are out of control if we do not impose any condition on ξ (i.e., using (B.3) to control it). On the other hand, if we restrict ξ to the direction $u-e_1$, then we can control these difficult terms by using the main good term (B.3). Namely, introduce the decomposition

$$u=e_1+t\xi,$$

where $t=\|u-e_1\|_2\ll 1$. Then for (B.5) we write

$$(a_k\cdot u)^3(a_k\cdot\xi)=(a_k\cdot u)^2(a_k\cdot e_1)(a_k\cdot\xi)+t(a_k\cdot u)^2(a_k\cdot\xi)^2.$$

Since $t\ll 1$, the term $t(a_k\cdot u)^2(a_k\cdot e_1)^2$ (together with the pre-factor term in (B.5)) can be included into (B.3) which still has a good lower bound by using localization. On the other hand, the term $(a_k\cdot u)^2(a_k\cdot e_1)(a_k\cdot\xi)$ can be split as

$$\begin{aligned}&(a_k\cdot u)^2(a_k\cdot e_1)(a_k\cdot\xi)\\ =&(a_k\cdot u)^2(a_k\cdot e_1)(a_k\cdot\xi)\phi\Big(\frac{a_k\cdot u}{K}\Big) &\text{(B.9)}\\ &+(a_k\cdot u)^2(a_k\cdot e_1)(a_k\cdot\xi)\Big(1-\phi\Big(\frac{a_k\cdot u}{K}\Big)\Big), &\text{(B.10)}\end{aligned}$$

where ϕ is a smooth cut-off function satisfying $0\leq\phi(z)\leq 1$ for all $z\in\mathbb{R}$, $\phi(z)=1$ for $|z|\leq 1$ and $\phi(z)=0$ for $|z|\geq 2$. Clearly the contribution of (B.9) in (B.5) is OK for union bounds. On the other hand, for (B.10) we have

$$\begin{aligned}&(a_k\cdot u)^2|a_k\cdot e_1||a_k\cdot\xi|\cdot\Big(1-\phi\Big(\frac{a_k\cdot u}{K}\Big)\Big)\\ \leq&(a_k\cdot u)^2\epsilon(a_k\cdot\xi)^2+\epsilon^{-1}(a_k\cdot u)^2(a_k\cdot e_1)^2\Big(1-\phi\Big(\frac{a_k\cdot u}{K}\Big)\Big).\end{aligned}$$

Clearly this is under control (the first term can again be controlled using (B.3)).

Now we turn to (B.7). The main term is $(a_k\cdot u)^4$. We write

$$(a_k\cdot u)^2(a_k\cdot u)^2=(a_k\cdot u)^2(a_k\cdot e_1)^2+t^2(a_k\cdot u)^2(a_k\cdot\xi)^2+2t(a_k\cdot u)^2(a_k\cdot e_1)(a_k\cdot\xi).$$

Clearly then this is also under control.

By further using localization, we can then show that with high probability, it holds that

$$H_{\xi\xi}\geq\mathbb{E}H_{\xi\xi}+\text{Error},$$

where $|\text{Error}|\ll 1$. The desired conclusion then follows from Lemma B.5. □

Remark B.2. Introduce the parametrization $u=\sqrt{R}(e_1\cos\theta+e^{\perp}\sin\theta)$, where $e^{\perp}\in e_1=0$, $|R-1|\ll 1$ and $|\theta|\ll 1$. One might hope to prove that the Hessian matrix

$$\begin{pmatrix}\partial_{RR}f & \partial_{R\theta}f\\ \partial_{R\theta}f & \partial_{\theta\theta}f\end{pmatrix}$$

is positive definite near $u=e_1$ under the mere assume $m\gtrsim n$ and with high probability. However there is a subtle issue which we explain as follows. Consider the main term (write $X=a_k\cdot e_1$ and $Y=a_k\cdot e^{\perp}$)

$$\tilde{f}=\tilde{f}_k=\frac{\Big(R(X\cos\theta+Y\sin\theta)^2-X^2\Big)^2}{\beta R+X^2}.$$

The most troublesome piece come from quartic and cubic terms in Y, and we consider

$$\tilde{h}_1=\frac{R^2Y^4\sin^4\theta}{\beta R+X^2},\quad \tilde{h}_2=\frac{R^2\Big(4Y^3X\sin^3\theta\cos\theta\Big)}{\beta R+X^2}.$$

For $\tilde{h}_2$ we do not have a favorable sign and the only hope is to control it via $\tilde{h}_1$. On the other hand, for $\tilde{h}_1$, we can take $X=Y=1$, $\beta=1$, and compute

$$(\partial_{RR}\tilde{h}_1)\cdot(\partial_{\theta\theta}\tilde{h}_1)-(\partial_{R\theta}\tilde{h}_1)^2=-\frac{8R^2}{(1+R)^4}\sin^6\theta\cdot\Big(3+4R+R^2+(2+4R+R^2)\cos2\theta\Big).$$

In yet other words, the sign is not favorable and this renders the Hessian out of control (before taking the expectation).

Lemma B.5. *Let $u=e_1+t\xi$, where $\xi\in\mathbb{S}^{n-1}$. Then for $|t|\ll 1$, we have*

$$\mathbb{E}\partial_{tt}f(u)\geq c_0>0,\quad \forall\xi\in\mathbb{S}^{n-1},$$

where $c_0>0$ is a constant depending only on β.

Proof. Introduce the parametrization $\xi = se_1 + \sqrt{1-s^2}e^\perp$, where $e^\perp \cdot e_1 = 0$, $|s| \le 1$. Then

$$u = e_1 + t\Big(se_1 + \sqrt{1-s^2}e^\perp\Big) = (1+ts)e_1 + t\sqrt{1-s^2}e^\perp.$$

Thus

$$\mathbb{E}f(u) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \underbrace{\frac{\Big(((1+ts)x + t\sqrt{1-s^2}y)^2 - x^2\Big)^2}{\beta(1+2ts+t^2)+x^2}}_{=:h(t,s,x,y)} e^{-\frac{x^2+y^2}{2}}dxdy.$$

It is not difficult to check that

$$\partial_{tt}h(t,s,x,y)\Big|_{t=0} = \frac{8x^2(sx+\sqrt{1-s^2}y)^2}{\beta+x^2}.$$

Thus it follows that

$$\mathbb{E}\partial_{tt}f(u)\Big|_{t=0,|s|\le 1} \gtrsim 1.$$

The desired result then follows by a simple perturbation argument using the fact that $\mathbb{E}\partial_{ttt}f$ is uniformly bounded and taking $|t|$ sufficiently small. □

Appendix C: technical estimates for Section 4

Lemma C.1. *Let $u = \sqrt{R}\hat{u}$ with $0 < c_1 \le R \le c_2 < \infty$ and $\hat{u} \in \mathbb{S}^{n-1}$. Assume $\hat{u} = \cos\theta e_1 + \sin\theta e^\perp$, where $\theta \in [0,\pi]$ and $e^\perp \in \mathbb{S}^{n-1}$ satisfies $e^\perp \cdot e_1 = 0$. We have*

$$\mathbb{E}\partial_\theta f = a_1(\beta_1,\beta_2,R,\theta)\sin(2\theta);$$

where

$$\gamma_1 < a_1(\beta_1,\beta_2,R,\theta) \le \gamma_2, \quad \forall\theta\in[0,\pi], \quad c_1 \le R \le c_2;$$

and $\gamma_1 > 0$, $\gamma_2 > 0$ are constants depending only on (β_1, β_2, c_1, c_2). Furthermore for some sufficiently small constants $\theta_0 = \theta_0(\beta_1,\beta_2,c_1,c_2) > 0$, $\theta_1 = \theta_1(\beta_1,\beta_2,c_1,c_2) > 0$, we have

$$\begin{aligned} &\gamma_3 < \mathbb{E}\partial_{\theta\theta}f < \gamma_4, && \text{if } \ 0 \le \theta \le \theta_0 \ \text{ or } \ \pi - \theta_0 \le \theta \le \pi, \\ &\gamma_5 < -\mathbb{E}\partial_{\theta\theta}f < \gamma_6, && \text{if } \ \Big|\theta - \frac{\pi}{2}\Big| < \theta_1, \end{aligned}$$

where $\gamma_i > 0$, $i = 3,\cdots,6$ depend only on (β_1, β_2, c_1, c_2).

Proof. We have

$$\mathbb{E}f(u)=\frac{1}{2\pi}\int_{\mathbb{R}^2}\frac{\Big(R(x\cos\theta+y\sin\theta)^2-x^2\Big)^2}{R+\beta_1R(x\cos\theta+y\sin\theta)^2+\beta_2x^2}e^{-\frac{x^2+y^2}{2}}dxdy.$$

Denote

$$h(a,b)=\frac{(Ra^2-b)^2}{R+\beta_1Ra^2+\beta_2b}.$$

Then

$$\begin{aligned}
&\partial_\theta\Big(h(x\cos\theta+y\sin\theta,x^2)\Big)=(-x\sin\theta+y\cos\theta)\partial_a h;\\
&\partial_x\Big(h(x\cos\theta+y\sin\theta,x^2)\Big)=\partial_a h\cdot\cos\theta+2x\partial_b h;\\
&\partial_y\Big(h(x\cos\theta+y\sin\theta,x^2)\Big)=\partial_a h\cdot\sin\theta;\\
&\partial_\theta\Big(h(x\cos\theta+y\sin\theta,x^2)\Big)=(y\partial_x-x\partial_y)\Big(h(x\cos\theta+y\sin\theta,x^2)\Big)-2xy\partial_b h.
\end{aligned}$$

By using integration by parts, we then obtain

$$\begin{aligned}
\mathbb{E}\partial_\theta f=&\frac{1}{\pi}\int_{\mathbb{R}^2}(-xy)(\partial_b h)(x\cos\theta+y\sin\theta,x^2)e^{-\frac{x^2+y^2}{2}}dxdy\\
=&\frac{2}{\pi}\int_{x>0,y>0}\Big((\partial_b h)(x\cos\theta-y\sin\theta,x^2)-(\partial_b h)(x\cos\theta+y\sin\theta,x^2)\Big)xye^{-\frac{x^2+y^2}{2}}dxdy.
\end{aligned}$$

Now denote

$$h_1(a,b)=\frac{(Ra-b)^2}{R+\beta_1Ra+\beta_2b}.$$

It is not difficult to check that for $a\geq0$, $b\geq0$, $\beta_1,\beta_2>0$, $R>0$,

$$\partial_{ab}h_1=-2R^2\frac{(1+a(\beta_1+\beta_2))\cdot(b(\beta_1+\beta_2)+R)}{(\beta_2b+R+\beta_1aR)^3}<0.$$

Observe that

$$(\partial_b h)(a,b)=(\partial_b h_1)(a^2,b).$$

Then if $x,y>0$ and $\theta\in[0,\pi]$, then

$$\begin{aligned}
&(\partial_b h)(x\cos\theta-y\sin\theta,x^2)-(\partial_b h)(x\cos\theta+y\sin\theta,x^2)\\
=&(\partial_b h_1)((x\cos\theta-y\sin\theta)^2,x^2)-(\partial_b h_1)((x\cos\theta+y\sin\theta)^2,x^2)\\
=&-2\int_0^1(\partial_{ab}h_1)\Big((x\cos\theta+y\sin\theta)^2-4\tau xy\cos\theta\sin\theta,x^2\Big)d\tau\cdot xy\cdot\sin(2\theta).
\end{aligned}$$

Integrating in x and y, we then obtain

$$\mathbb{E}\partial_\theta f = a_1(\beta_1,\beta_2,R,\theta)\sin(2\theta),$$

where $a_1 \sim 1$ and is a smooth function of θ. Differentiating in θ then gives

$$\mathbb{E}\partial_{\theta\theta} f = 2a_1(\beta_1,\beta_2,R,\theta)\cos(2\theta) + \partial_\theta a_1(\beta_1,\beta_2,R,\theta)\sin(2\theta).$$

Then second term clearly vanishes near $\theta = 0$, $\frac{\pi}{2}$, π. Thus the desired estimate for $\mathbb{E}\partial_{\theta\theta} f$ follows. □

Lemma C.2 (Strong convexity of $\mathbb{E}f$ when $\|u \pm e_1\| \ll 1$)**.** *Let $h(u) = \mathbb{E}f(u)$. There exists $0 < \epsilon_0 \ll 1$ such that the following hold:*

1. *If $\|u - e_1\|_2 \le \epsilon_0$, then for any $\xi \in \mathbb{S}^{n-1}$, we have*

$$\sum_{i,j=1}^{n} \xi_i \xi_j (\partial_i \partial_j h)(u) \ge \gamma_1 > 0,$$

   where γ_1 is a constant.

2. *If $\|u + e_1\|_2 \le \epsilon_0$, then for any $\xi \in \mathbb{S}^{n-1}$, we have*

$$\sum_{i,j=1}^{n} \xi_i \xi_j (\partial_i \partial_j h)(u) \ge \gamma_1 > 0.$$

Proof. We shall employ the same approach as in the proof of Theorem 2.5 in the second paper of this series of work and sketch only the needed modifications. Without loss of generality consider the regime $\|u - e_1\|_2 \ll 1$ and introduce the change of variables:

$$\begin{aligned} &u = \rho\hat{u}; \\ &\hat{u} = \sqrt{1-s^2}e_1 + se^\perp, \quad e^\perp \cdot e_1 = 0, \quad e^\perp \in \mathbb{S}^{n-1}, \end{aligned}$$

where $|\rho - 1| \ll 1$ and $0 \le s \ll 1$. Denote

$$h_1(\rho, s) = h(u) = h\Big(\rho\Big(\sqrt{1-s^2}e_1 + se^\perp\Big)\Big),$$

where we note that the value of $h(u)$ depends only on (ρ, s). Clearly

$$h_1(\rho,s) = \frac{1}{2\pi}\int_{\mathbb{R}^2} \underbrace{\frac{\Big(\rho^2(\sqrt{1-s^2}x + sy)^2 - x^2\Big)^2}{\rho^2 + \beta_1\rho^2(\sqrt{1-s^2}x + sy)^2 + \beta_2 x^2}}_{=:h_2(\rho,s,x,y)} e^{-\frac{x^2+y^2}{2}} dxdy.$$

It is easy to check that

$$\max_{\frac{1}{2}\leq\rho\leq 2,|s|\leq\frac{1}{2}}\sum_{i,j\leq 4}|\partial_\rho^i\partial_s^j h_1(\rho,s)|\lesssim 1.$$

By a tedious computation, we have

$$\partial_{\rho\rho}h_2(\rho,0,x,y)=\frac{2(3\rho^2+\rho^6)x^4+k_1\cdot x^6+k_2x^8}{(\beta_2x^2+\rho^2(1+\beta_1x^2))^3},$$

where

$$\begin{aligned}k_1&=2(-\beta_2+6\beta_1\rho^2+6\beta_2\rho^2+3\beta_2\rho^4+2\beta_1\rho^6);\\ k_2&=2(-\beta_1\beta_2-2\beta_2^2+3\beta_1^2\rho^2+6\beta_1\beta_2\rho^2+6\beta_2^2\rho^2+3\beta_1\beta_2\rho^4+\beta_1^2\rho^6).\end{aligned}$$

Since $\rho\to 1$, it is clear that $k_1>0$ and $k_2>0$, and thus

$$\partial_{\rho\rho}h_1(1,0)\gtrsim 1.$$

It is not difficult to check that $\partial_s h_1(\rho,0)=0$ for any $\rho>0$. Clearly also $\partial_{\rho s}h_1(\rho,0)=0$ for any $\rho>0$. To compute $\partial_{ss}h_1(1,0)$ we shall use Lemma C.1. Observe that ($s=\sin\theta$ with $\theta\to 0+$)

$$\begin{aligned}&h_1(\rho,\sin\theta)=\mathbb{E}f(u);\\ &\cos\theta\partial_s h_1(\rho,\sin\theta)=\mathbb{E}\partial_\theta f;\\ &-\sin\theta\partial_s h_1(\rho,\sin\theta)+\cos^2\theta\partial_{ss}h_1(\rho,\sin\theta)=\mathbb{E}\partial_{\theta\theta}f.\end{aligned}$$

Clearly it follows that

$$\partial_{ss}h_1(1,0)\gtrsim 1.$$

The rest of the argument is then essentially the same as in the proof of Theorem 2.5 in the second paper. We omit further details. □

Proof of Theorem 4.3. We rewrite

$$f(u)=\frac{1}{m}\sum_{k=1}^m G(\|u\|_2^2,(a_k\cdot u)^2,X_k^2),$$

where

$$G(a,b,c)=\frac{(b-c)^2}{a+\beta_1 b+\beta_2 c}.$$

Clearly for any $\xi\in\mathbb{S}^{n-1}$,

$$\begin{aligned}
&\sum_{i,j=1}^{n}\xi_i\xi_j\partial_{u_iu_j}f\\
=&\frac{1}{m}\sum_{k=1}^{m}\partial_aG\cdot 2\|\xi\|_2^2 &\text{(C.1)}\\
&+\frac{1}{m}\sum_{k=1}^{m}\partial_{aa}G\cdot 4(u\cdot\xi)^2 &\text{(C.2)}\\
&+\frac{1}{m}\sum_{k=1}^{m}\partial_{ab}G\cdot 8(a_k\cdot u)(a_k\cdot\xi)(\xi\cdot u) &\text{(C.3)}\\
&+\frac{1}{m}\sum_{k=1}^{m}\partial_{bb}G\cdot 4(a_k\cdot u)^2(a_k\cdot\xi)^2 &\text{(C.4)}\\
&+\frac{1}{m}\sum_{k=1}^{m}\partial_bG\cdot 2(a_k\cdot\xi)^2. &\text{(C.5)}
\end{aligned}$$

In the above,

$$\partial_aG=(\partial_aG)(\|u\|_2^2,(a_k\cdot u)^2,X_k^2)$$

and similar notation is used for $\partial_{aa}G$, $\partial_{bb}G$, ∂_bG.

Estimate of (C.1) and (C.2). Clearly these two terms are OK for union bounds, and we have (for $m\gtrsim n$ and with high probability)

$$|(\text{C.1})-\text{mean}|+|(\text{C.2})-\text{mean}|\ll 1,\quad \forall\xi\in\mathbb{S}^{n-1},\quad \forall\frac{1}{2}\le\|u\|_2\le 2.$$

Estimate of (C.3). We have

$$(\partial_{ab}G)(a,b,c)=-\frac{2(b-c)(a+(\beta_1+\beta_2)c)}{(a+\beta_1b+\beta_2c)^3}.$$

Consider the function

$$\tilde{G}_1(a,y,c)=-y\frac{2(y^2-c)(a+(\beta_1+\beta_2)c)}{(a+\beta_1y^2+\beta_2c)^3}.$$

Clearly for $\frac{1}{10}\le a,\tilde{a}\le 10$, $y,\tilde{y}\in\mathbb{R}$, $c\ge 0$, we have $|\tilde{G}_1|\lesssim 1$ and

$$|\tilde{G}_1(a,y,c)-\tilde{G}_1(\tilde{a},\tilde{y},c)|\lesssim|a-\tilde{a}|+|y-\tilde{y}|.$$

Then for any $(u,\,\tilde{u})$ with $\frac{1}{2}\leq\|u\|_2,\|\tilde{u}\|_2\leq 2$ and $(\xi,\,\tilde{\xi})$ with $\xi,\tilde{\xi}\in\mathbb{S}^{n-1}$, we have

$$\begin{aligned}&\Big|(\partial_{ab}G)(\|u\|_2^2,(a_k\cdot u)^2,X_k^2)(a_k\cdot u)(a_k\cdot\xi)\\&\quad-(\partial_{ab}G)(\|\tilde{u}\|_2^2,(a_k\cdot\tilde{u})^2,X_k^2)(a_k\cdot\tilde{u})(a_k\cdot\tilde{\xi})\Big|\\\lesssim&|a_k\cdot(\xi-\tilde{\xi})|+|a_k\cdot\xi|\cdot(|a_k\cdot(u-\tilde{u})|+\|u-\tilde{u}\|_2).\end{aligned}$$

Thus the union bound is also OK for this term, and we have

$$|(\mathrm{C}.3)-\mathrm{mean}|\ll 1,\quad\forall\xi\in\mathbb{S}^{n-1},\quad\forall\frac{1}{2}\leq\|u\|_2\leq 2.$$

Estimate of (C.4) and (C.5). We begin by noting that (C.4) and (C.5) can be combined into one term. Namely, observe that

$$(\partial_{bb}G)(a,b,c)\cdot 2b+(\partial_bG)(a,b,c)=\frac{H_1}{(a+\beta_1b+\beta_2c)^3},$$

where

$$\begin{aligned}H_1=&\beta_1^2b^3+a^2(6b-2c)+3\beta_1\beta_2b^2c+3b(\beta_1^2+2\beta_1\beta_2+2\beta_2^2)c^2-\beta_2(\beta_1+2\beta_2)c^3\\&+a\Big(3\beta_1b^2+6(\beta_1+2\beta_2)bc-(\beta_1+4\beta_2)c^2\Big).\end{aligned}$$

We can then write

$$(\mathrm{C}.4)+(\mathrm{C}.5)=\frac{1}{m}\sum_{k=1}^m(a_k\cdot\xi)^2h_3(u,a_k\cdot u,X_k),$$

where h_3 is a bounded smooth function with bounded derivatives in all of its arguments. Now let $\phi\in C_c^\infty$ be such that $0\leq\phi(x)\leq 1$ for all x, $\phi(x)=1$ for $|x|\leq 1$ and $\phi(x)=0$ for $|x|\geq 2$. We then split the sum as

$$\begin{aligned}&\frac{1}{m}\sum_{k=1}^m(a_k\cdot\xi)^2h_3(u,a_k\cdot u,X_k),\\=&\frac{1}{m}\sum_{k=1}^m(a_k\cdot\xi)^2\phi(\frac{a_k\cdot\xi}{K})h_3(u,a_k\cdot u,X_k)\\&+\frac{1}{m}\sum_{k=1}^m(a_k\cdot\xi)^2(1-\phi(\frac{a_k\cdot\xi}{K}))\cdot h_3(u,a_k\cdot u,X_k),\end{aligned}$$

where K will be taken sufficiently large. Clearly the first term will be OK for union bounds. On the other hand, the second term can be dominated by

$$\text{const}\cdot\frac{1}{m}\sum_{k=1}^{m}(a_k\cdot\xi)^2\Big(1-\phi\Big(\frac{a_k\cdot\xi}{K}\Big)\Big),$$

which can be made small by taking K large. Thus we have

$$|(\text{C.4})+(\text{C.5})-\text{mean}|\ll 1,\quad \forall\xi\in\mathbb{S}^{n-1},\quad \forall\frac{1}{2}\le\|u\|_2\le 2.$$

Collecting the estimates, we have for $m\gtrsim n$ and with high probability,

$$\Big|\sum_{i,j=1}^{n}\xi_i\xi_j\partial_{u_iu_j}f(u)-\text{mean}\Big|\ll 1,\quad \forall\xi\in\mathbb{S}^{n-1},\quad \forall\frac{1}{2}\le\|u\|_2\le 2.$$

The desired result then follows from Lemma C.2. □

Acknowledgements

J. F. Cai was supported in part by Hong Kong Research Grant Council General Research Grant Nos. 16309518, 16309219, 16310620, and 16306821. Y. Wang was supported in part by Hong Kong Research Grant Council General Research Grant Nos. 16306415 and 16308518.

References

[1] S. Bhojanapalli, N. Behnam, and N. Srebro, Global optimality of local search for low rank matrix recovery, Adv. Neural Infor. Proc. Syst., (2016), pp. 3873–3881.

[2] E. J. Candès and X. Li, Solving quadratic equations via PhaseLift when there are about as many equations as unknowns, Found. Comput. Math., 14(5) (2014), pp. 1017–1026.

[3] E. J. Candès, X. Li, and M. Soltanolkotabi, Phase retrieval via Wirtinger flow: Theory and algorithms, IEEE Trans. Inf. Theory, 61(4) (2015), pp. 1985–2007.

[4] E. J. Candès, T. Strohmer, and V. Voroninski, Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming, Commun. Pure Appl. Math., 66(8) (2013), pp. 1241–1274.

[5] J. Cai, M. Huang, D. Li and Y. Wang, Solving phase retrieval with random initial guess is nearly as good as by spectral initialization, Appl. Comput. Harmon. Anal., (2021).

[6] J. Cai, M. Huang, D. Li and Y. Wang, Nearly optimal bounds for the global geometric landscape of phase retrieval, arxiv preprint, in preparation, 2021.
[7] Y. Chen and E. J. Candès, Solving random quadratic systems of equations is nearly as easy as solving linear systems, Commun. Pure Appl. Math., 70(5) (2017), pp. 822–883.
[8] J. C. Dainty and J. R. Fienup, Phase retrieval and image reconstruction for astronomy, Image Recovery: Theory and Application, 231 (1987), 275.
[9] S. S. Du, C. Jin, J. D. Lee, and M. I. Jordan, Gradient descent can take exponential time to escape saddle points, Adv. Neural Infor. Proc. Syst., (2017), pp. 1067–1077.
[10] J. R. Fienup, Phase retrieval algorithms: a comparison, Appl. Opt., 21(15) (1982), pp. 2758–2769.
[11] B. Gao, Y. Wang, and Z. Xu, Solving a perturbed amplitude-based model for phase retrieval, 2019, available: `http://arxiv.org/abs/1904.10307`.
[12] B. Gao and Z. Xu, Phaseless recovery using the Gauss–Newton method, IEEE Trans. Signal Process., 65(22) (2017), pp. 5885–5896.
[13] R. Ge, F. Huang, C. Jin, and Y. Yuan, Escaping from saddle points—online stochastic gradient for tensor decomposition, Conference on Learning Theory, (2015), pp. 797–842.
[14] R. Ge, J. Lee, C. Jin, and T. Ma, Matrix completion has no spurious local minimum, Adv. Neural Infor. Proc. Syst., (2016), pp. 2973–2981.
[15] R. W. Gerchberg, A practical algorithm for the determination of phase from image and diffraction plane pictures, Optik, 35 (1972), pp. 237–246.
[16] R. W. Gerchberg and W. O. Saxton, A practical algorithm for the determination of the phase from image and diffraction plane pictures, Optik, 35 (1972), pp. 237–246.
[17] R. W. Harrison, Phase problem in crystallography, Josa A, 10(5) (1993), pp. 1046–1055.
[18] M. Huang and Y. Wang, Linear convergence of randomized Kaczmarz method for solving complex-valued phaseless equations, 2021, available: `http://arxiv.org/abs/2109.11811`.
[19] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, How to escape saddle points efficiently, Proceedings of the 34th International Conference on Machine Learning-Volume 70, (2017), pp. 1724–1732.
[20] C. Jin, P. Netrapalli, and M. I. Jordan, Accelerated gradient descent escapes saddle points faster than gradient descent, 2017, available: `http://arxiv.org/abs/1711.10456`.
[21] Z. Li, J. F. Cai, and K. Wei, Towards the optimal construction of a loss function without spurious local minima for solving quadratic equations, IEEE Trans. Inf. Theory, 66(5) (2020), pp. 3242–3260.
[22] J. Miao, T. Ishikawa, Q. Shen, and T. Earnest, Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes, Annu. Rev. Phys. Chem., (59) (2008), pp. 387–410.
[23] R. P. Millane, Phase retrieval in crystallography and optics, J. Optical Soc. Am. A,

7(3) (1990), pp. 394–411.
[24] P. Netrapalli, P. Jain, and S. Sanghavi, Phase retrieval using alternating minimization, IEEE Trans. Signal Process., 63(18) (2015), pp. 4814–4826.
[25] D. Park, A. Kyrillidis, and C. Caramanis, Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach, 2016, available: `http://arxiv.org/abs/1609.03240`.
[26] H. Sahinoglou and S. D. Cabrera, On phase retrieval of finite-length sequences using the initial time sample, IEEE Trans. Circuits and Syst., 38(8) (1991), pp. 954–958.
[27] 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, IEEE Signal Process. Mag., 32(3) (2015), pp. 87–109.
[28] J. Sun, Q. Qu, and J, Wright, A geometric analysis of phase retrieval, Found. Comput. Math., 18(5) (2018), pp. 1131–1198.
[29] J. Sun, Q. Qu, and J, Wright, Complete dictionary recovery over the sphere I: Overview and the geometric picture, IEEE Trans. Inf. Theory, 63(2) (2016), pp. 853–884.
[30] Y. S. Tan and R. Vershynin, Phase retrieval via randomized kaczmarz: Theoretical guarantees, Information and Inference: A Journal of the IMA, 8(1) (2019), pp. 97–123.
[31] R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, U.K. Cambridge University Press, 2018.
[32] I. Waldspurger, A. d'Aspremont, and S. Mallat, Phase recovery, maxcut and complex semidefinite programming, Math. Prog., 149(1-2) (2015), pp. 47–81.
[33] A. Walther, The question of phase retrieval in optics, J. Mod. Opt., 10(1) (1963), pp. 41–49.
[34] G. Wang, G. B. Giannakis, and Y. C. Eldar, Solving systems of random quadratic equations via truncated amplitude flow, IEEE Trans. Inf. Theory, 64(2) (2018), pp. 773–794.
[35] K. Wei, Solving systems of phaseless equations via kaczmarz methods: a proof of concept study, Inverse Probl., 31(12) (2015), 125008.
[36] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms, The Journal of Machine Learning Research, 18(1) (2017), pp. 5164–5198.