Towards the optimal construction of a loss function without spurious local minima for solving quadratic equations

Zhenzhen Li\textsuperscript{*} \hspace{1cm} Jian-Feng Cai\textsuperscript{†} \hspace{1cm} Ke Wei\textsuperscript{‡}

November 7, 2019

Abstract

The problem of finding a vector $\mathbf{x}$ which obeys a set of quadratic equations $|\mathbf{a}_k^\top \mathbf{x}|^2 = y_k$, $k = 1, \cdots, m$, plays an important role in many applications. In this paper we consider the case when both $\mathbf{x}$ and $\mathbf{a}_k$ are real-valued vectors of length $n$. A new loss function is constructed for this problem, which combines the smooth quadratic loss function with an activation function. Under the Gaussian measurement model, we establish that with high probability the target solution $\mathbf{x}$ is the unique local minimizer (up to a global phase factor) of the new loss function provided $m \gtrsim n$. Moreover, the loss function always has a negative directional curvature around its saddle points.

1 Introduction

Many applications in science and engineering, such as X-ray crystallography \cite{23}, diffraction and array imaging \cite{7}, and electron microscopy \cite{30}, are essentially about solving systems of quadratic equations. This paper concerns a real-valued case of the problem. The goal is to find a vector $\mathbf{x} \in \mathbb{R}^n$ which can solve $m$ quadratic equations of the form

$$y_k = |\mathbf{a}_k^\top \mathbf{x}|^2, \quad k = 1, \cdots, m,$$

where $\mathbf{y} = [y_1, \cdots, y_m]^\top \in \mathbb{R}_+^m$ and $\{\mathbf{a}_k \in \mathbb{R}^n\}_{k=1}^m$ are known. Despite the seeming simplicity of (1), solving this problem is computationally intractable. Indeed, a special instance of (1) is the NP-hard stone problem \cite{14}.

The problem of recovering a vector from a set of quadratic measurements, especially from the Fourier type measurements, has long been studied. Moreover, it has received intensive investigations over the past few years largely due to its connection with low rank matrix recovery. Even though the corresponding low rank matrix recovery problem is still nonconvex and computationally intractable, we can approximate it by its nearest convex relaxation, leading to a convex formulation known as PhaseLift. Performance guarantee of PhaseLift has been established in \cite{13, 10, 11, 17} under different measurement models, showing that successful recovery can be achieved when the number

\textsuperscript{*}Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China; Email: zlice@ust.hk.

\textsuperscript{†}Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China; Email: jfcai@ust.hk.

\textsuperscript{‡}School of Data Science, Fudan University, Shanghai, China; Email: kewei@fudan.edu.cn.
of equations is (nearly) proportional to the number of unknowns. There are also other convex relaxation methods for solving systems of quadratic equations; see for example [42, 2, 21, 22].

Though convex approximations usually come with recovery guarantees, they are not computationally desirable for large-scale problems. In contrast, many simple nonconvex algorithms are able to solve (1) both accurately and efficiently. Among them are a family of algorithms with optimal or near-optimal provable guarantees, including alternating projections and its resampled variant [31, 41], Kaczmarz methods [24, 37], and those algorithms which propose to compute the solution of (1) by minimizing certain nonconvex loss functions [12, 14, 43, 9, 49, 44]. Specifically, a gradient descent algorithm known as Wirtinger Flow has been developed in [12] based on the following smooth quadratic loss function

\[
\tilde{f}(z) = \frac{1}{2m} \sum_{k=1}^{m} \left( (a_k^\top z)^2 - y_k \right)^2.
\]  

In [43, 49], gradient descent algorithms were developed based on a loss function similar to (2) but with \((a_k^\top z)^2 - y_k\) replaced by \(|a_k^\top z| - \sqrt{y_k}\), while a Poisson loss function is adopted in [14].

Theoretical guarantees of the aforementioned algorithms typically require that the initial guess is sufficiently close to the true solution. However, numerical simulations show that these algorithms can often achieve successful recovery even with random initialization. To understand this empirical success, Sun et al. [36] investigated the global geometry of the loss function in (2). It has been shown that under the Gaussian measurement model \(\tilde{f}(z)\) does not have any spurious local minima provided \(m \gtrsim n \log^3 n\). Putting it in another way, under this sampling condition, the target signal \(x\) is the only minimizer of \(\tilde{f}(z)\) up to a global phase factor. Moreover, \(\tilde{f}(z)\) possesses a negative directional curvature around each saddle point. Thus, algorithms that can avoid saddle points and converge to a local minimizer are bound to find a global minimizer; see for example [27].

Our work follows this line of research and attempts to construct a loss function with \(x\) being the only minimizer up to a global phase factor when \(m \gtrsim n\). That is, we want to construct a loss function without spurious local minima for (1) conditioned on the optimal sampling complexity.

In recent years, there has been a surge of interest in nonconvex optimization for problems arising from signal processing and machine learning; solving systems of quadratic equations is one of them. For more general low rank matrix recovery, a variety of nonconvex algorithms have been developed and analyzed, including those based on matrix factorization [38, 51] and those based on the embedded manifold of low rank matrices [46, 45]. The reader can refer to the review paper [8] for more details. Geometric landscape of related loss functions for low rank matrix recovery has been investigated in [18, 19, 28, 4, 32]. Similar results have also been established for nonconvex formulations of other problems, for example blind deconvolution [50], dictionary learning [34, 35], tensor completion [1, 20], phase synchronization [5, 29, 6], and deep neural networks [39, 48, 33, 25].

1.1 Motivation and main result

As stated previously, a few of the algorithms for solving Gaussian random systems of quadratic equations are able to achieve successful recovery with high probability provided \(m \gtrsim n\), including TWF [14], TAF [43] and TRGrad [9], just to name a few. In addition, it is also known that a

\[^{1}\text{The notation } m \gtrsim g(n) \text{ means that there exists an absolute constant } C > 0 \text{ such that } m \geq C \cdot g(n). \text{ It is worth noting that } m \gtrsim n \log^3 n \text{ in [36] is a sufficient condition for the well-behaved landscape of } \tilde{f}(z), \text{ so this does not mean that a spurious minimum of } \tilde{f}(z) \text{ exists when } n \lesssim m \lesssim n \log^3 n.\]
unique solution (up to a global phase factor) of (11) can be determined from \( m \geq 2n - 1 \) generic measurements for the real problem or from \( m \geq 4n - 4 \) generic measurements for the complex problem \([3, 16]\). Thus, it is interesting to see whether there exists a loss function for solving random systems of quadratic equations which does not have any spurious local minima when \( m \gg n \), in contrast to \( m \gg n \log^3 n \) for (2) as is established in \([36]\). To the best of our knowledge, this question has not been explored yet. In our work, we will give an affirmative answer for the real-valued problem.

![Figure 1: Two examples of activation functions: \( h_1(u) \) (left) and \( h_2(u) \) (right).](image)

We construct the new loss function \( f(z) \) by coupling (2) with an activation function \( h(u) \),

\[
    f(z) = \frac{1}{2m} \sum_{k=1}^{m} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right),
\]

where the activation function \( h(u) \) satisfies

\[
    \begin{cases}
    h(u) = 1 & \text{if } 0 \leq u \leq \beta, \\
    h(u) \in [0, 1] & \text{if } u \in (\beta, \gamma), \quad \text{and } |h'(u)|, |h''(u)| \text{ exist and are bounded} \\
    h(u) = 0 & \text{if } u \geq \gamma
    \end{cases}
\]

for two predetermined universal parameters \( 1 < \beta < \gamma \) that are sufficiently large. As can be seen later, the activation function has been introduced to control the gradient of the function so that overshooting can be avoided.

For simplicity, we assume \( \gamma = C \cdot \beta \) for some absolute constant \( C > 1 \). Note that the bounds of \( |h'(u)| \) and \( |h''(u)| \) rely on the parameters \( \beta \) and \( \gamma \). Two examples of \( h(u) \) are

\[
    h_1(u) = \begin{cases}
    1 & 0 \leq u \leq \beta \\
    -6 \left( \frac{u-\beta}{\gamma-\beta} \right)^5 + 15 \left( \frac{u-\beta}{\gamma-\beta} \right)^4 - 10 \left( \frac{u-\beta}{\gamma-\beta} \right)^3 + 1 & u \in (\beta, \gamma) \\
    0 & u \geq \gamma.
    \end{cases}
\]
Theorem 1.1

(Main result)

Moreover, at each saddle point \( f_{\text{n}} = 1 \) of

In numerical simulations, a direct examination of the geometric landscape of a loss function seems complicated, for example the Hessian of \( f \).

The introduction of the activation function makes the gradient and Hessian of \( f(z) \) more complicated, for example the Hessian of \( f \) has 12 terms. Thus we postpone the calculations of \( \nabla f(z) \) and \( \nabla^2 f(z) \) to Appendix B. Despite this, the activation function is able to circumvent the effect of the fourth power of Gaussian random variables which is heavy-tailed. To demonstrate this effect, we consider the case when \( n = 1 \) (i.e., both \( z \) and \( x \) are scalars) and then use the Q-Q plot to compare the random variables in the expressions of the gradients (indeed derivative since \( n = 1 \)) of \( \nabla f(z) \) and \( \nabla^2 f(z) \), as well as the random variables in the expressions of the Hessians (indeed second derivative since \( n = 1 \)) of \( \nabla^2 f(z) \) and \( \nabla^2 f(z) \). The plots are presented in Figure 2 from which we can clearly see that the random variables without the activation function are more heavy-tailed.

Assuming that \( a_k, k = 1, \cdots, m \) are independent Gaussian vectors: \( a_k \sim \mathcal{N}(0, I_n) \), our main result for \( f(z) \) is stated as follows:

**Theorem 1.1** (Main result). With probability exceeding \( 1 - e^{-\Omega(m)} \), the function \( f(z) \) defined in (B) with sufficiently large \( 1 < \beta < \gamma \) does not have any spurious local minima provided \( m \geq n \). Moreover, at each saddle point \( f(z) \) has a negative directional curvature.

We would like to note that \( \beta \) and \( \gamma \) in Theorem 1.1 are two absolute positive constants whose values are fixed, and the constant hidden in \( m \geq n \) relies on \( \beta \) and \( \gamma \).

**1.2 Numerical Illustration**

In numerical simulations, a direct examination of the geometric landscape of a loss function seems to be out of reach. Instead, we investigate the performance of the gradient descent iteration

\[
\begin{align*}
z_{t+1} &= z_t - \mu \nabla \tilde{f}(z_t) \\
z_{t+1} &= z_t - \mu \nabla f(z_t)
\end{align*}
\]

Here \( \Omega(m) \) is a value which is greater than \( C \cdot m \) for some numerical constant \( C > 0 \).
Figure 2: Q-Q plots for the random variables in the gradients (left) and Hessians (right) of $f(z)$ and $\tilde{f}(z)$ when $n = 1$. In this simulation, $x = 1$, $z = 2$, and a total of $10^5$ (i.e., $m = 10^5$) standard Gaussian random variables are independently generated. Then we compute the first and second derivatives of each term in (2) and (3), respectively. The activation function $h_1(u)$ with $\beta = 10$ and $\gamma = 20$ is used in $f(z)$.

Figure 3: Recovery performance of gradient descent for functions in (2) and (3).
with three different stepsizes \( \mu \in \{0.1, 0.2, 0.3\}/(\|y\|_1/m) \) when minimizing the loss functions defined in (2) and (3), respectively. We use \( h_2(u) \) with \( \gamma = 1.5\beta \) for the loss function in (3). Different values of \( \beta \) are adopted for different stepsizes, namely, \( \beta = 20 \) when \( \mu = 0.1/(\|y\|_1/m) \), \( \beta = 10 \) when \( \mu = 0.2/(\|y\|_1/m) \), and \( \beta = 5 \) when \( \mu = 0.3/(\|y\|_1/m) \). Roughly speaking, a more stringent activation condition is imposed for the larger stepsize.

Numerical tests are conducted for fixed \( n = 128 \) and \( m/n \) increasing from 4 to 10 by 0.5. For each fixed pair of \((n, m)\), 500 problem instances on randomly generated \( a_k \sim \mathcal{N}(0, I_n) \) and \( x \sim \mathcal{N}(0, I_n) \) are tested. The initial guess for the gradient descent iteration is generated randomly and independently according to the standard Gaussian distribution. We consider the algorithm to have successfully reconstructed a test signal if it returns an estimate with the relative reconstruction error \( \text{dist}(z_1, x)/\|x\| \) being less than or equal to \( 10^{-3} \) under the distance defined by

\[
\text{dist}(z, x) = \min\{\|z - x\|, \|z + x\|\}.
\]

The plots of the successful recovery probability against the sampling ratio for the three different stepsizes are presented in Figures 3a - 3c. We can see that, when \( \mu = 0.1/(\|y\|_1/m) \), the transition curves of the gradient iterations based on the two different loss functions are nearly indistinguishable. However, the advantage of our loss function over the one without the activation function becomes more significant as \( \mu \) increases. In particular, when \( \mu = 0.3/(\|y\|_1/m) \), the gradient iteration based on the new loss function with proper \((\beta, \gamma)\) can achieve more than 80% successful recovery when \( m \geq 6n \), whereas the gradient descent iteration based on the other loss function can hardly succeed even when \( m = 10n \). A close look at the simulation results reveals that the gradient descent method for the vanilla \( \ell_2 \) loss function \( \tilde{f}(z) \) can either diverge or converge to a local minimizer when \( \mu = 0.3/(\|y\|_1/m) \). In contrast, the new loss function can still succeed for larger stepsizes potentially because the activation function can regularize each component of the gradient so that those components that go wild can be avoided.

We also put the recovery transitions corresponding to the new loss function but with different values of \((\mu, \beta, \gamma)\) in the same plot; see Figure 3d. Competitive performance of the gradient descent iterations corresponding to different triples of \((\mu, \beta, \gamma)\) can be observed when \( m \geq 5n \). This suggests that similar recovery performance can be achieved by trading off appropriately between the stepsize and the parameters in the loss function.

### 1.3 Organization and notation

The rest of this paper is organized as follows. The geometric landscape of the new loss function is presented in Section 2. Section 3 contains the detailed justification, with the proofs for the technical lemmas being presented in Section 4. We conclude this paper with potential future directions in Section 5.

Following the notation above we use bold face lowercase letters to denote column vectors and use normal font letters with subindices for their entries. In particular, we fix \( x \) as the underlying vector to be reconstructed. The \( \ell_1 \)-norm and \( \ell_2 \)-norm of a vector \( z \) are denoted by \( \|z\|_1 \) and \( \|z\| \), respectively. Given two vectors \( z \) and \( x \), their distance, denoted \( \text{dist}(z, x) \), is defined by

\[
\text{dist}(z, x) = \min\{\|z - x\|, \|z + x\|\}.
\]

For a given matrix \( A \), we use \( \|A\| \) to denote the matrix operator norm which is defined by

\[
\|A\| = \sup_{\|z\| = 1} \|Az\|.
\]
When $A$ is symmetric we also have
\[ \|A\| = \sup_{\|z\| = 1} |z^\top Az|. \]

For two symmetric positive semidefinite matrices $A$ and $B$, if $A \preceq B$ then $\|A\| \leq \|B\|$. Recall that the notation $m \gtrsim g(n)$ means that there exists an absolute constant $C > 0$ such that $m \geq C \cdot g(n)$. Similarly, the notation $m \lesssim g(n)$ means that there exists an absolute constant $C > 0$ such that $m \leq C \cdot g(n)$. Throughout the paper, $C$ denotes an absolute constant whose value may change from line to line. In addition, $!!$ means double factorial; that is $n!! = n(n-2)(n-4) \cdots$.

2 Geometric landscape of the new function

In this section we present the detailed geometric landscape of $f(z)$. Differing from the partition in [36], we decompose $\mathbb{R}^n$ into five non-overlapping regions (see Figure 4):

- $\mathcal{R}_1 := \{z : \text{dist}(z, x) \leq \frac{1}{5}\|x\|\}$,
- $\mathcal{R}_{2a} := \{z : \frac{1}{3} - \delta < \frac{\|z\|^2}{\|x\|^2} < \frac{99}{100} \text{ and dist}(z, x) > \frac{1}{3}\|x\|\}$,
- $\mathcal{R}_{2b} := \{z : \frac{99}{100} \leq \frac{\|z\|^2}{\|x\|^2} \leq \frac{101}{100} \text{ and dist}(z, x) > \frac{1}{3}\|x\|\}$,
- $\mathcal{R}_{2c} := \{z : \frac{\|z\|^2}{\|x\|^2} > \frac{101}{100} \text{ and dist}(z, x) > \frac{1}{3}\|x\|\}$,
- $\mathcal{R}_3 := \{z : 0 < \frac{\|z\|^2}{\|x\|^2} \leq \frac{1}{3} - \delta\}$,

where $\delta$ is a fixed constant in $(0, 1/100]$. The properties of $f(z)$ over these five regions are summarized in the following five theorems.

**Theorem 2.1.** With probability at least $1 - e^{-\Omega(m)}$,
\[ \lambda_{\min}(\nabla^2 f(z)) \geq \frac{1}{25}\|x\|^2 \]
holds uniformly for all $z \in \mathcal{R}_1$ provided $m \gtrsim n$.

**Theorem 2.2.** With probability at least $1 - e^{-\Omega(m)}$, all critical points in $\mathcal{R}_{2a}$ must exist in the subregion defined by
\[ \frac{1}{3} - \delta < \frac{\|z\|^2}{\|x\|^2} < \frac{1}{3} + \delta \quad \text{and} \quad |z^\top x| < \delta\|x\|^2 \] (4)
provided $m \gtrsim n$. Moreover, with probability exceeding $1 - e^{-\Omega(m)}$,
\[ x^\top \nabla^2 f(z)x \leq -3\|x\|^4 \quad \text{and} \quad z^\top \nabla^2 f(z)z \geq \|x\|^4. \]
hold uniformly for all $z$ in the subregion (4) provided $m \gtrsim n$. 

7
Theorem 2.3. With probability at least $1 - e^{-\Omega(m)}$,
\[ z^\top \nabla f(z) \geq \frac{9}{1000} \|x\|^4 \]
holds uniformly for all $z \in \mathcal{R}_{2a}$ provided $m \geq n$.

Theorem 2.4. With probability at least $1 - e^{-\Omega(m)}$,
\[ z^\top \nabla f(z) \geq \frac{49}{1000} \|z\|^4 \]
holds uniformly for all $z \in \mathcal{R}_{2c}$ provided $m \geq n$.

Theorem 2.5. With probability at least $1 - e^{-\Omega(m)}$,
\[ z^\top \nabla f(z) \leq -5\delta \|z\|^2 \|x\|^2 \]
holds uniformly for all $z \in \mathcal{R}_3$ provided $m \geq n$.

The proofs of the above theorems are deferred to Section 3. We can check the results in these theorems by conducting a simple numerical test: 1) randomly generate a set of standard Gaussian vectors $\{a_k\}_{k=1}^m \subset \mathbb{R}^n$, 2) fix the measurement vectors and randomly generate different $z$’s for each region, and 3) check whether the result in each theorem holds or not. For conciseness, we report in Table 1 the computational results in regions $\mathcal{R}_1$, $\mathcal{R}_{2c}$ and $\mathcal{R}_3$ (with three randomly generated $z$ in each region).

From the five theorems, it is evident that critical points of $f(z)$ can only occur in $\mathcal{R}_1$ and $\mathcal{R}_{2a}$, since at critical points one has $\nabla f(z) = 0$. Noticing that $\pm x \in \mathcal{R}_1$, $f(z) \geq 0$ and $f(\pm x) = 0$, by Theorem 2.1, we know that $\pm x$ are the local minimizers. Theorem 2.2 implies that at any critical point in $\mathcal{R}_{2a}$, the Hessian of $f(z)$ has a negative directional curvature as well as a positive directional curvature. Thus, critical points in $\mathcal{R}_{2a}$ must be ridable saddle points [36]. Putting it all together, we can establish Theorem 1.1 and show that every local minimizer is a global minimizer. Additionally, though $f(z)$ is singular at $z = 0$, Theorem 2.5 shows that local minimizers of $f(z)$ cannot exist around 0. Moreover, it also implies that searching along the gradient descent direction at any point in $\mathcal{R}_3$ will move the point further away from the origin.
Table 1: This table numerically checks the results in the theorems: \( x = [1, \cdots, 1]^T \in \mathbb{R}^n \) with \( n = 128 \), \( \{a_k\}_{k=1}^m \subset \mathbb{R}^n \) \( (m = 6n) \) are independent standard Gaussian vectors. The vectors \( z \) in \( R_1, R_{2c}, \) and \( R_3 \) are generated uniformly at random (for \( R_{2c} \) we consider its intersection with the ball \( \{ z : \|z\| \leq 2\|x\| \} \)). The parameter \( \delta \) which defines \( R_3 \) is chosen to be \( 1/100 \). The theoretical bounds refer to those given in Thms. 2.1, 2.4, and 2.5.

|                  | \( \lambda_{\text{min}}(\nabla^2 f(z)) \) in \( R_1 \) | \( z^T \nabla f(z) \) in \( R_{2c} \) | \( z^T \nabla f(z) \) in \( R_3 \) |
|------------------|--------------------------|--------------------------|--------------------------|
| Numerical results| 94.43 96.39 108.83   | 2.87e5 2.61e5 6.87e6  | -2184.41 -2203.39 -95.00 |
| Theoretical bounds| 5.12 5.12 5.12     | 2632.14 3233.31 55528.5 | -73.63 -81.98 -2.55     |

(upper bound)

3 Proofs for Section 2

3.1 Technical lemmas

In order to prove the main theorems, we first list several technical lemmas that will be used repeatedly in this section, but defer the proofs to Section 4. Here and throughout this paper, if the expression of a random variable or a random matrix is long we will simply use \( E[\cdot] \) to denote the associated expectation.

**Lemma 3.1.** Let \( h(u) \) be a continuous function defined on \( [0, \infty) \) which obeys

\[
\begin{cases}
  h(u) = 1 & \text{if } 0 \leq u \leq \beta, \\
  h(u) \in [0,1] & \text{if } u \in (\beta, \gamma), \\
  h(u) = 0 & \text{if } u \geq \gamma
\end{cases}
\]

for two absolute numerical constants \( \gamma > \beta \geq 1 \). Assume \( a_k \sim \mathcal{N}(0, I_n) \), \( k = 1, \cdots, m \), are independent. Then for any \( \epsilon \in (0,1) \) and all nonzero vectors \( u, v \in \mathbb{R}^n \),

\[
\left\| \frac{1}{m} \sum_{k=1}^m (a_k^T u)^s (a_k^T v)^t h \left( \frac{|a_k^T u|^2}{\|u\|^2} \right) h \left( \frac{|a_k^T v|^2}{\|v\|^2} \right) a_k a_k^T - E[\cdot] \right\| \\
\lesssim \left( \epsilon \cdot \max\{s,t\} \gamma^{\frac{s+t}{2}} + \gamma^{\frac{s+t}{2}} \epsilon^{-1} e^{-0.49\epsilon^2} + \gamma^{\frac{s+t+1}{2}} e^{-0.49\beta} \right) \|u\|^s \|v\|^t
\]

holds with probability at least \( 1 - e^{-\Omega(m^2)} \) provided \( m \geq \epsilon^{-2} \log \epsilon^{-1} \cdot n \), where the exponents \( s \) and \( t \) are two nonnegative integers.

**Lemma 3.2.** Under the setup of Lemma 3.1,

\[
\left\| E \left[ (a_k^T u)^s (a_k^T v)^t \right] h \left( \frac{|a_k^T u|^2}{\|u\|^2} \right) h \left( \frac{|a_k^T v|^2}{\|v\|^2} \right) - 1 \right. \left. a_k a_k^T \right\| \\
\lesssim ((8s)!!)^{1/8} ((8t)!!)^{1/8} \|u\|^s \|v\|^t \cdot e^{-0.25\beta}
\]

holds for all \( \|u\| \neq 0 \) and \( \|v\| \neq 0 \).

**Lemma 3.3.** Under the setup of Lemma 3.1,

\[
\left\| \frac{1}{m} \sum_{k=1}^m (a_k^T z)^s (a_k^T x)^t h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m |a_k^T x|^2}{\|y\|_1} \right) - h \left( \frac{|a_k^T x|^2}{\|x\|^2} \right) \right. \left. a_k a_k^T \right\|
\]

9
\[
\lesssim 2^t \gamma^{\frac{s+t}{2}} \left( \sqrt{\beta e^{-0.245\beta}} + \epsilon \right) \|z\|^s \|x\|^{t}
\]
holds uniformly for all \(\|z\| \neq 0\) with probability at least \(1 - e^{-\Omega(m^2)}\) provided \(m \gtrsim e^{-2 \log e^{-1} \cdot n}\).

**Lemma 3.4.** Let \(h(u)\) and \(g(u)\) be two continuous functions defined on \([0, \infty)\) satisfying

\[
\begin{cases}
    h(u) = 1 & \text{if } 0 \leq u \leq \beta, \\
    h(u) \in [0, 1] & \text{if } u \in (\beta, \gamma), \\
    h(u) = 0 & \text{if } u \geq \gamma
\end{cases}
\quad \text{and} \quad
\begin{cases}
    g(u) = 0 & \text{if } u \in [0, \beta] \cup [\gamma, \infty], \\
    |g(u)| \leq 1 & \text{if } \beta < u < \gamma
\end{cases}
\]

for two absolute numerical constants \(\gamma > \beta \geq 1\). Assume \(a_k \sim \mathcal{N}(0, I_n), k = 1, \cdots, m,\) are independent. Then for any \(\epsilon \in (0, 1)\) and all nonzero vectors \(z \in \mathbb{R}^n\),

\[
\left\| \frac{1}{m} \sum_{k=1}^m (a_k^\top z)^s (a_k^\top x)^t g \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m |a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top \right\| \lesssim 2^t \gamma^{\frac{s+t}{2}} \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right) \|z\|^s \|x\|^t
\]

holds with probability at least \(1 - e^{-\Omega(m^2)}\) provided \(m \gtrsim e^{-2 \log e^{-1} \cdot n}\), where the exponents \(s\) and \(t\) are two nonnegative integers.

**Lemma 3.5.** Under the setup of Lemma 3.4,

\[
\left\| \frac{1}{m} \sum_{k=1}^m (a_k^\top z)^s (a_k^\top x)^t g \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m |a_k^\top x|^2}{\|y\|_1} \right) z a_k^\top \right\| \lesssim 2^t \gamma^{\frac{s+t+1}{2}} \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) \|z\|^{s+1} \|x\|^t
\]

holds uniformly for all \(\|z\| \neq 0\) with probability at least \(1 - e^{-\Omega(m^2)}\) provided \(m \gtrsim e^{-2 \log e^{-1} \cdot n}\).

**Lemma 3.6.** Under the setup of Lemma 3.4, for \(s \geq 2\)

\[
\left\| \frac{1}{m} \sum_{k=1}^m (a_k^\top z)^s (a_k^\top x)^t g \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m |a_k^\top x|^2}{\|y\|_1} \right) x \right\| \lesssim 2^t \gamma^{\frac{s+t}{2}} \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right) \|z\|^s \|x\|^t
\]

holds uniformly for all \(\|z\| \neq 0\) with probability at least \(1 - e^{-\Omega(m^2)}\) provided \(m \gtrsim e^{-2 \log e^{-1} \cdot n}\).

### 3.2 Proof of Theorem 2.1

Due to symmetry, it suffices to consider the region \(\|z - x\| \leq \frac{1}{2} \|x\|\), from which we have

\[
\frac{4}{5} \|x\| \leq \|z\| \leq \frac{6}{5} \|x\|.
\] (5)

Though there are twelve terms in the expression for \(\nabla^2 f(z)\) (see (22)), it is not difficult to see that the second term through the last term, with their sum denoted by \(I_2\), can be bounded by Lemmas 3.4 to 3.6 (the details are deferred to Appendix B.1), giving

\[
\|I_2\| \lesssim \gamma^2 \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) \max \left\{ \|h'\|_\infty, |h''|\right\} \max \left\{ \|z\|^2, \|x\|^2, \frac{\|x\|^4}{\|z\|^2} \right\}
\] (6)

\[
\lesssim \gamma^2 \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) \max \left\{ \|h'\|_\infty, |h''|\right\} \|x\|^2.
\] (7)
where we have used (5) in the second line. Define
\[
I_1 = \frac{1}{m} \sum_{k=1}^{m} \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k a_k^T,
\]
which is the first term in the Hessian of $\nabla^2 f(z)$. By setting $(s, t)$ to be $(2, 0)$ and $(0, 2)$ respectively in Lemma 3.3 we have
\[
\left\| I_1 - \frac{1}{m} \sum_{k=1}^{m} \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|x\|^2} \right) a_k a_k^T \right\|
\leq \gamma \frac{2}{2} \left( e^{-0.245\beta} + \epsilon \right) \max \{ \|z\|^2, \|x\|^2 \}
\leq \gamma \frac{3}{2} \left( e^{-0.245\beta} + \epsilon \right) \|x\|^2.
\]
(8)
Moreover, letting $(s, t)$ to be $(2, 0)$ and $(0, 2)$ respectively in Lemma 3.1 we have
\[
\lambda_{\min} \left( \frac{1}{m} \sum_{k=1}^{m} \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|x\|^2} \right) a_k a_k^T \right)
\geq \lambda_{\min} \left( E \left[ \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|x\|^2} \right) a_k a_k^T \right] \right)
-C \gamma \frac{3}{2} \left( e + e^{1-0.49e^{-2}} + e^{-0.245\beta} \right) \|x\|^2
\geq \lambda_{\min} \left( E \left[ \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) a_k a_k^T \right] \right)
- \left\| E \left[ \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) \left\{ h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|x\|^2} \right) - 1 \right\} a_k a_k^T \right] \right\|
-C \gamma \frac{3}{2} \left( e + e^{1-0.49e^{-2}} + e^{-0.245\beta} \right) \|x\|^2,
\]
where $C$ is an absolute constant whose value may change from line to line.
For any unit vector $q \in S^{n-1}$, we have
\[
q^T E \left[ \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) a_k a_k^T \right] q
= 6\|q\|^2\|z\|^2 + 12(q^T z)^2 - 2\|q\|^2\|x\|^2 - 4(q^T x)^2
\geq 6\|z\|^2 - 2\|x\|^2 - 4\|q^T(z + x)||q^T(z - x)|
\geq \frac{2}{25}\|x\|^2,
\]
where the second line follows from the standard result
\[
E \left[ (a_k^T u)^2 (a_k^T v)^2 \right] = \|u\|^2\|v\|^2 + 2(u^T v)^2,
\]
the third line follows from $12(q^T z)^2 \geq 4(q^T z)^2$, and the last line follows from (5) which implies $\|z\|^2 \geq \frac{16}{25}\|x\|^2$, $\|z + x\| \leq \frac{14}{15}\|x\|$, and $\|z - x\| \leq \frac{1}{5}\|x\|$. Moreover, the application of Lemma 3.2 yields
\[
\left\| E \left[ \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) \left\{ h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|x\|^2} \right) - 1 \right\} a_k a_k^T \right] \right\|
\]
\[
\lambda_{\min} \left( \frac{1}{m} \sum_{k=1}^{m} \left( 6(a_k^\top z)^2 - 2(a_k^\top x)^2 \right) h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{|a_k^\top x|^2}{\|x\|^2} \right) a_k a_k^\top \right) \\
\geq \left( \frac{2}{25} - Ce^{-0.25\beta} - C\gamma^\frac{3}{2} \left( \epsilon + \epsilon^{-1} e^{-0.49e^{-2}} + e^{-0.49\beta} \right) \right) \|x\|^2 \\
- C\gamma^\frac{3}{2} \left( e^{-0.245\beta} + \epsilon \right) \|x\|^2 - C\gamma^\frac{3}{2} \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) \max \{|h'|_\infty, |h''|_\infty\} \|x\|^2 \\
\geq \frac{1}{25} \|x\|^2
\]

for sufficiently small \( \epsilon \) and sufficiently large \( \beta \) and \( \gamma \) since \( \max \{|h'|_\infty, |h''|_\infty\} = O(1) \) in our construction.

### 3.3 Proof of Theorem 2.2

Due to symmetry we only need to consider the case \( z^\top x \geq 0 \) in \( \mathcal{R}_{2a} \). We will first show that with probability \( 1 - e^{-\Omega(m)} \),

\[
x^\top \nabla f(z) < -\frac{\delta}{100} \|x\|^4
\]

holds uniformly for all \( z \) in the region \( \mathcal{R}_{2a} \cap \{ z \mid z^\top x \geq \delta \|x\|^2 \} \) provided \( m \gtrsim n \), and hence excluding the possibility of any critical points in this region.

Next, we will show that with probability \( 1 - e^{-\Omega(m)} \),

\[
z^\top \nabla f(z) > \delta \|x\|^4
\]

holds uniformly for all \( z \) in the region \( \mathcal{R}_{2a} \cap \{ z \mid 0 \leq z^\top x < \delta \|x\|^2 \} \cap \{ z \mid \|z\|^2/\|x\|^2 \geq \frac{1}{\beta} + \delta \} \) provided \( m \gtrsim n \), and again excluding the possibility of any critical points in this region. Then the first part of Theorem 2.2 follows immediately by combining the above two results together.

**Proof of (10)** Notice that

\[
x^\top \nabla f(z) = \frac{1}{m} \sum_{k=1}^{m} 2 \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) (a_k^\top z)(a_k^\top x) \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) \\
+ \frac{1}{\|z\|^2} \frac{1}{m} \sum_{k=1}^{m} (a_k^\top z)^2 (a_k^\top x)^2 \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h' \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right)
\]
\[-\frac{z^\top x}{\|x\|^4} - \frac{1}{m} \sum_{k=1}^{m} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 (a_k^\top z)^2 h' \left( \frac{a_k^\top z}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) \]

\[:= I_1 + I_2 + I_3.\]

By Lemmas 3.3, 3.1 and 3.2 and noticing \(\|z\|^2 \leq \|x\|^2\) in \(\mathcal{R}_{2a}\), we have

\[I_1 \leq \frac{1}{m} \sum_{k=1}^{m} 2 \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) (a_k^\top z)(a_k^\top x) h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{|a_k^\top x|^2}{\|z\|^2} \right) + C\gamma \left( \sqrt{\beta e^{-0.245\beta}} + \epsilon \right) \|x\|^4 \]

\[\leq \mathbb{E} \left[ 2 \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) (a_k^\top z)(a_k^\top x) h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{|a_k^\top x|^2}{\|z\|^2} \right) \right] + C\gamma \left( \gamma \epsilon + \gamma \epsilon^{-1} e^{-0.49e^{-2}} + \gamma^{1.5} e^{-0.49\beta} \right) \|x\|^4 + C\gamma \left( \sqrt{\beta e^{-0.245\beta}} + \epsilon \right) \|x\|^4 \]

\[\leq \mathbb{E} \left[ 2 \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) (a_k^\top z)(a_k^\top x) \right] + C e^{-0.25\beta} \|x\|^4 + C \left( \gamma \epsilon + \gamma \epsilon^{-1} e^{-0.49e^{-2}} + \gamma^{1.5} e^{-0.49\beta} \right) \|x\|^4 + C\gamma \left( \sqrt{\beta e^{-0.245\beta}} + \epsilon \right) \|x\|^4.\]

On the other hand, by Lemma 3.3 and noticing \(\frac{\beta}{300} \|x\|^2 \leq (\frac{1}{2} - \delta) \|x\|^2 \leq \|z\|^2 \leq \|x\|^2\) in \(\mathcal{R}_{2a}\) since \(\delta \leq \frac{1}{100}\), we have

\[I_2 + I_3 \leq C\|h'\|_{\infty} \gamma^2 \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right) \|x\|^4.\]

Thus, combining the above two inequalities together implies that for all \(z\) in \(\mathcal{R}_{2a} \cap \{ z \mid z^\top x \geq \delta \|x\|^2 \}\) we have

\[x^\top \nabla f(z) \leq \mathbb{E} \left[ 2 \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) (a_k^\top z)(a_k^\top x) \right] + \frac{\delta}{20} \|x\|^4 \]

\[= 6(z^\top x)(\|z\|^2 - \|x\|^2) + \frac{\delta}{20} \|x\|^4 \]

\[\leq -\frac{\delta}{100} \|x\|^4 \]

where the first line can be achieved by choosing \(\epsilon\) sufficiently small and \(\gamma > \beta\) sufficiently large, and in the last line we have used the fact \(\|z\|^2 \leq \frac{\beta}{300} \|x\|^2\) and \(z^\top x \geq \delta \|x\|^2\) in \(\mathcal{R}_{2a} \cap \{ z \mid z^\top x \geq \delta \|x\|^2 \}\).

**Proof of (11)** First we have

\[z^\top \nabla f(z) = \frac{1}{m} \sum_{k=1}^{m} 2(|a_k^\top z|^2 - |a_k^\top x|^2)(a_k^\top z)^2 h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) \]

\[\quad = \frac{1}{m} \sum_{k=1}^{m} 2(|a_k^\top z|^2 - |a_k^\top x|^2)(a_k^\top z)^2 h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) \]

By applying Lemmas 3.3, 3.1 and 3.2 in order, we have

\[z^\top \nabla f(z) \geq \frac{1}{m} \sum_{k=1}^{m} 2(|a_k^\top z|^2 - |a_k^\top x|^2)(a_k^\top z)^2 h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) \]

\[- C\gamma \left( \sqrt{\beta e^{-0.245\beta}} + \epsilon \right) (\|z\|^4 + \|z\|^2 \|x\|^2) \]

13
\[ \begin{align*}
&\geq \mathbb{E} \left[ 2(|a_k^T z|^2 - |a_k^T x|^2) (a_k^T z)^2 h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{|a_k^T z|^2}{\|x\|^2} \right) \right] \\
&- C\gamma^{1.5} \left( 2\epsilon + e^{-0.49e^{-2}} \right) \left( \|z\|^4 + \|z\|^2 \|x\|^2 \right) \\
&- C\gamma \left( \sqrt{\beta} e^{-0.245\beta} + \epsilon \right) \left( \|z\|^4 + \|z\|^2 \|x\|^2 \right) \\
&\geq \mathbb{E} \left[ 2(|a_k^T z|^2 - |a_k^T x|^2) (a_k^T z)^2 \right] - C e^{-0.25\beta} \left( \|z\|^4 + \|z\|^2 \|x\|^2 \right) \\
&- C\gamma^{1.5} \left( 2\epsilon + e^{-0.49e^{-2}} \right) \left( \|z\|^4 + \|z\|^2 \|x\|^2 \right) \\
&- C\gamma \left( \sqrt{\beta} e^{-0.245\beta} + \epsilon \right) \left( \|z\|^4 + \|z\|^2 \|x\|^2 \right) .
\end{align*} \]

Noticing that in \( \mathcal{R}_{2\alpha} \cap \{ z \mid 0 \leq z^T x < \delta \|x\|^2 \} \cap \{ z \mid \|z\|^2 / \|x\|^2 \geq \frac{1}{3} + \delta \} \) we have \((\frac{1}{3} + \delta) \|x\|^2 \leq \|z\|^2 \leq \|x\|^2\), and consequently,

\[ z^T f(z) \geq 6\|z\|^4 - 2\|z\|^2 \|x\|^2 - 4(z^T x)^2 \\
- \left( C\gamma \left( \sqrt{\beta} e^{-0.245\beta} + \epsilon \right) + C\gamma^{1.5} \left( 2\epsilon + C e^{-0.49e^{-2}} + e^{-0.49\beta} \right) + C e^{-0.25\beta} \right) \|x\|^4 \\
\geq 6\|z\|^4 - 2\|z\|^2 \|x\|^2 - 4(z^T x)^2 - \delta \|x\|^4 \\
\geq 6 \left( \frac{1}{3} + \delta \right)^2 \|x\|^4 - 2 \left( \frac{1}{3} + \delta \right) \|x\|^4 - 4\delta^2 \|x\|^4 - \delta \|x\|^4 \\
= (\delta + 2\delta^2) \|x\|^4 > \delta \|x\|^4 ,
\]

where the second inequality can be achieved by choosing \( \epsilon \) to be sufficiently small and \( \gamma \geq \beta \) to be sufficiently large.

In the first part we have established that critical points in \( \mathcal{R}_{2\alpha} \) must obey

\[ \frac{1}{3} - \delta < \frac{\|z\|^2}{\|x\|^2} < \frac{1}{3} + \delta \] and \( z^T x < \delta \|x\|^2 \).

Thus, by (7), we have

\[ x^T \nabla^2 f(x) x \leq \frac{1}{m} \sum_{k=1}^{m} \left( 6(a_k^T z)^2 - 2(a_k^T x)^2 \right) (a_k^T z)^2 h \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{|a_k^T z|^2}{\|x\|^2} \right) \\
+ C\gamma^2 \left( e^{-0.245\beta} + \epsilon \right) \max \{ |h'|_\infty, |h''|_\infty \} \|x\|^4 
\]

Applying Lemmas 3.3, 3.1 and 3.2 in order yields

\[ x^T \nabla^2 f(x) x \leq \mathbb{E} \left[ 6(a_k^T z)^2 (a_k^T x)^2 - 2(a_k^T x)^4 \right] \\
+ \left( C\gamma \left( \sqrt{\beta} e^{-0.245\beta} + \epsilon \right) + C\gamma^{1.5} \left( 2\epsilon + C e^{-0.49e^{-2}} + e^{-0.49\beta} \right) + C e^{-0.25\beta} \right) \|x\|^4 \\
+ C\gamma^2 \left( e^{-0.245\beta} + \epsilon \right) \max \{ |h'|_\infty, |h''|_\infty \} \|x\|^4 \\
\leq 6\|z\|^2 \|x\|^2 + 12(z^T x)^2 - 6\|x\|^4 + \delta \|x\|^4 \\
\leq \left( 6 \left( \frac{1}{3} + \delta \right) + 12\delta^2 - 6 + \delta \right) \|x\|^4
\]

14
\[ -3\|x\|^4, \]

where in the second inequality for fixed \( \delta \) we choose \( \epsilon \) to be sufficiently small and \( \beta \) and \( \gamma \) to be properly large. Similarly, but considering a different direction, we have

\[
z^\top \nabla^2 f(z) z \geq 18\|z\|^4 - 2\|z\|^2\|x\|^2 - 4(z^\top x)^2 - \left(C\gamma \sqrt{\beta} e^{-0.245\beta} + \epsilon \right) \|x\|^4
\]

\[
- C\gamma^2 \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) \max \{|h'|_\infty, |h''|_\infty\} \|x\|^4
\]

\[
\geq \left( 18 \left( \frac{1}{3} - \delta \right)^2 - 2 \left( \frac{1}{3} + \delta \right) - 4\delta^2 - \delta \right) \|x\|^4
\]

\[
\geq \|x\|^4.
\]

### 3.4 Proof of Theorem 2.3

We only need to consider the case \( z^\top x \geq 0 \). Since in \( \mathcal{R}_{2b} \), one has

\[
\frac{99}{100} \leq \frac{\|z\|^2}{\|x\|^2} \leq \frac{101}{100} \quad \text{and} \quad \|z - x\| > \frac{1}{5} \|x\|.
\]

Thus,

\[
z^\top x \leq \frac{1}{2} \left( \|x\|^2 + \|z\|^2 - \frac{1}{25} \|x\|^2 \right) \leq 0.985 \|x\|^2.
\]

Noticing \( \|z\|^2 \geq \frac{99}{100} \|x\|^2 \) in \( \mathcal{R}_{2b} \), by choosing \( \epsilon \) to be sufficiently small and \( \beta \) and \( \gamma \) to be properly large in \( [13] \), we have

\[
z^\top \nabla f(z) \geq E\left[2\left(|a_k^\top z|^2 - |a_k^\top x|^2\right)(a_k^\top z)^2 - \delta \|x\|^4 \right]
\]

\[
= 6\|z\|^4 - 2\|z\|^2\|x\|^2 - 4(z^\top x)^2 - \delta \|x\|^4
\]

\[
= \|x\|^4 \left( 6 \left( \frac{\|z\|^2}{\|x\|^2} \right)^2 - 2 \frac{\|z\|^2}{\|x\|^2} \right) - 4(z^\top x)^2 - \delta \|x\|^4
\]

\[
\geq \left( 6 \cdot \frac{99^2}{100^2} - 2 \cdot \frac{99}{100} - 4 \cdot 0.985^2 - \delta \right) \|x\|^4
\]

\[
\geq \frac{9}{1000} \|x\|^4
\]

provided \( \delta \leq \frac{1}{100} \), where in the fourth line we have used the fact that the minimum of \( 6\|z\|^4 - 2\|z\|^2\|x\|^2 \) over \( \frac{99}{100} \leq \frac{\|z\|^2}{\|x\|^2} \) is achieved at \( \|z\|^2 = \frac{99}{100} \|x\|^2 \).

### 3.5 Proof of Theorem 2.4

Similarly to the proof for Theorem 2.3, we have

\[
z^\top \nabla f(z) \geq 6\|z\|^4 - 2\|z\|^2\|x\|^2 - 4(z^\top x)^2 - \delta \|x\|^4
\]
\[ \geq 6\|z\|^2(\|z\|^2 - \|x\|^2) - \delta\|x\|^4 \]
\[ \geq \frac{6}{101}\|z\|^4 - \delta\|z\|^4 \]
\[ \geq \frac{49}{100}\|z\|^4, \]

where in the third line we have used the fact \( \|z\|^2 \geq \frac{101}{100}\|x\|^2 \) in \( \mathcal{R}_3 \), and in the last line we have used the assumption \( \delta \leq \frac{1}{100} \).

### 3.6 Proof of Theorem 2.5

Recall that \( z^T \nabla f(z) \) is given in (12). Thus similar to (13) but applying by Lemmas 3.3, 3.1 and 3.2 in the reverse direction yields

\[
\begin{align*}
z^T \nabla f(z) & \leq \mathbb{E} \left[ 2(\|a_k^T z\|^4 - \|a_k^T z\|2\|a_k^T x\|^2) + Ce^{-0.25\beta}\left(\|z\|^4 + \|z\|^2\|x\|^2\right) \right. \\
& \quad + C\gamma^{1.5} \left(2\epsilon + Ce^{-0.49\epsilon - 2} + e^{-0.49}\right) \left(\|z\|^4 + \|z\|^2\|x\|^2\right) \\
& \quad + C\gamma \left(\sqrt{\beta}e^{-0.245\beta} + \epsilon \right) \left(\|z\|^4 + \|z\|^2\|x\|^2\right) \\
& \left. + Ce^{-0.25\beta}\left(\|z\|^4 + \|z\|^2\|x\|^2\right) \right)
\end{align*}
\]

It follows that

\[
\begin{align*}
z^T \nabla f(z) & \leq 6\|z\|^4 - 2\|z\|^2\|x\|^2 - 4(z^T x)^2 \\
& + \left(C\gamma \left(\sqrt{\beta}e^{-0.245\beta} + \epsilon \right) + C\gamma^{1.5} \left(2\epsilon + Ce^{-0.49\epsilon - 2} + e^{-0.49}\right) + C\gamma \left(\sqrt{\beta}e^{-0.245\beta} \right) \left(\|z\|^4 + \|z\|^2\|x\|^2\right) \right) \\
& \leq 6\|z\|^4 - 2\|z\|^2\|x\|^2 + \left(C\gamma^{1.5} \left(2\epsilon + Ce^{-0.49\epsilon - 2} + e^{-0.49}\right) + C\gamma \left(\sqrt{\beta}e^{-0.245\beta} \right) \left(\|z\|^4 + \|z\|^2\|x\|^2\right) \right) \\
& \leq 6\|z\|^4 - 2\|z\|^2\|x\|^2 + \left(C\gamma^{1.5} \left(2\epsilon + Ce^{-0.49\epsilon - 2} + e^{-0.49}\right) + C\gamma \left(\sqrt{\beta}e^{-0.245\beta} \right) \left(\|z\|^4 + \|z\|^2\|x\|^2\right) \right) \\
& \leq 6\|z\|^4 - 2\|z\|^2\|x\|^2 + \left(C\gamma^{1.5} \left(2\epsilon + Ce^{-0.49\epsilon - 2} + e^{-0.49}\right) + C\gamma \left(\sqrt{\beta}e^{-0.245\beta} \right) \left(\|z\|^4 + \|z\|^2\|x\|^2\right) \right) \\
& \leq -5\|z\|^2\|x\|^2,
\end{align*}
\]

where in the second and the third inequalities we have used the assumption \( \frac{\|z\|^2}{\|x\|^2} \leq \frac{1}{\delta} \), and in the last inequality we choose \( \epsilon \) to be sufficiently small and \( \gamma > \beta \) to be sufficiently large.

### 4 Proofs of technical lemmas

#### 4.1 Proof of Lemma 3.1

Due to the homogeneity, it suffices to establish the inequality for all \( u \in S^{n-1} \) and \( v \in S^{n-1} \). We will first consider a fixed pair of \( u \) and \( v \) and then use the covering argument. For fixed \( u \) and \( v \) of unit norm, it suffices to establish a uniform bound for

\[
\left| \frac{1}{m} \sum_{k=1}^{m} (a_k^T u)^2 (a_k^T v)^t h(\|a_k^T u\|^2) h(\|a_k^T v\|^2) (a_k^T w)^2 - \mathbb{E} [\cdot] \right| \tag{14}
\]
over all $\mathbf{w} \in \mathcal{N}_{1/4}$, where $\mathcal{N}_{1/4}$ is a $1/4$-net of $\mathcal{S}^{n-1}$. This is because for any symmetric matrix $\mathbf{A}$ one has
\[
\|\mathbf{A}\| \leq 2 \sup_{z \in \mathcal{N}_{1/4}} |\langle \mathbf{A}z, z \rangle|,
\]
see Lemma 5.4 in [40]. Noticing that
\[
\frac{1}{m} \sum_{k=1}^{m} (a_{k}^{\top} \mathbf{u})^s (a_{k}^{\top} \mathbf{v})^t h \left( |a_{k}^{\top} \mathbf{u}|^2 \right) h \left( |a_{k}^{\top} \mathbf{v}|^2 \right) (a_{k}^{\top} \mathbf{w})^2 \]

and $(a_{k}^{\top} \mathbf{w})^2$ is a standard Chi-square, we can see that
\[
\frac{1}{m} \sum_{k=1}^{m} (a_{k}^{\top} \mathbf{u})^s (a_{k}^{\top} \mathbf{v})^t h \left( |a_{k}^{\top} \mathbf{u}|^2 \right) h \left( |a_{k}^{\top} \mathbf{v}|^2 \right) (a_{k}^{\top} \mathbf{w})^2
\]
is sub-exponential with the sub-exponential norm $\| \cdot \|_{\psi_1}$ bounded by an absolute constant times $\gamma^{\frac{s+t}{2}}$. It follows that [40]
\[
\left\| \left( a_{k}^{\top} \mathbf{u} \right)^s (a_{k}^{\top} \mathbf{v})^t h \left( |a_{k}^{\top} \mathbf{u}|^2 \right) h \left( |a_{k}^{\top} \mathbf{v}|^2 \right) (a_{k}^{\top} \mathbf{w})^2 - \mathbb{E} [\cdot] \right\|_{\psi_1} \lesssim \gamma^{\frac{s+t}{2}}.
\]
Thus the application of the Bernstein’s inequality implies that
\[
\left\| \frac{1}{m} \sum_{k=1}^{m} (a_{k}^{\top} \mathbf{u})^s (a_{k}^{\top} \mathbf{v})^t h \left( |a_{k}^{\top} \mathbf{u}|^2 \right) h \left( |a_{k}^{\top} \mathbf{v}|^2 \right) (a_{k}^{\top} \mathbf{w})^2 - \mathbb{E} [\cdot] \right\|_{\psi_1} \lesssim \gamma^{\frac{s+t}{2}} \epsilon
\]
with probability at least $1 - 2e^{-\Omega(m^2)}$ for $\epsilon \in (0, 1)$. By [40] Lemma 5.2 we know that $|\mathcal{N}_{1/4}| \leq 9^n$. Thus the failure probability over all $\mathbf{w} \in \mathcal{N}_{1/4}$ can be bounded by $9^n \cdot 2e^{-\Omega(m^2)} = 2e^{-\Omega(m^2)+n \log 9}$ which is less than $2e^{-\Omega(m^2)}$ (with a different constant hidden in $\Omega(m^2)$) provided $m \gtrsim \epsilon^{-2} \cdot n$. Therefore,
\[
\left\| \frac{1}{m} \sum_{k=1}^{m} (a_{k}^{\top} \mathbf{u})^s (a_{k}^{\top} \mathbf{v})^t h \left( |a_{k}^{\top} \mathbf{u}|^2 \right) h \left( |a_{k}^{\top} \mathbf{v}|^2 \right) (a_{k}^{\top} \mathbf{w})^2 \right\|_{\psi_1} \lesssim \gamma^{\frac{s+t}{2}} \epsilon
\]
\[
\frac{1}{m} \sum_{k=1}^{m} (a_k^T u_0)^s h \left( |a_k^T u_0|^2 \right) \left\{ (a_k^T v)^t h \left( |a_k^T v|^2 \right) - (a_k^T v_0)^t h \left( |a_k^T v_0|^2 \right) \right\} a_k a_k^T \] (17)

Next we will focus on the first term of (17) and the second term can be similarly bounded. We can split the first term into five terms based on the decomposition of \([0, \beta] \times [0, \beta]\) and then provide an upper bound for each term.

**Region** \([0, \beta] \times [0, \beta]\)

\[
\left\| \frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^T u)^s h \left( |a_k^T u|^2 \right) - (a_k^T u_0)^s h \left( |a_k^T u_0|^2 \right) \right\} 1_{\{|a_k^T u|^{2} \leq \beta, |a_k^T u_0|^{2} \leq \beta\}} (a_k^T v)^t h \left( |a_k^T v|^2 \right) a_k a_k^T \right\|
\]

\[
\leq \left\| \frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^T u)^s - (a_k^T u_0)^s \right\} 1_{\{|a_k^T u| \leq \beta, |a_k^T u_0| \leq \beta\}} (a_k^T v)^t h \left( |a_k^T v|^2 \right) a_k a_k^T \right\|
\]

\[
+ \left\| \frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^T u)^s - (a_k^T u_0)^s \right\} 1_{\{|a_k^T u| \leq \beta, |a_k^T u_0| \leq \beta, |a_k^T u - a_k^T u_0| \leq \epsilon\}} (a_k^T v)^t h \left( |a_k^T v|^2 \right) a_k a_k^T \right\|
\]

\[
= \left\| \frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^T u - a_k^T u_0) \left( (a_k^T u)^{s-1} + (a_k^T u)^{s-2}(a_k^T u_0) + \cdots + (a_k^T u_0)^{s-1} \right) \right\} 1_{\{|a_k^T u| \leq \beta, |a_k^T u_0| \leq \beta, |a_k^T u - a_k^T u_0| \leq \epsilon\}} (a_k^T v)^t h \left( |a_k^T v|^2 \right) a_k a_k^T \right\|
\]

\[
\leq \epsilon \cdot s \beta^{s-1} \frac{\gamma^s}{s} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^T \right\| + 2\beta \gamma \frac{\gamma^s}{s} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^T 1_{\{|a_k^T u - a_k^T u_0| > \epsilon\}} \right\|
\]

\[
\leq \epsilon \cdot s \beta^{s-1} \frac{\gamma^s}{s} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^T \right\| + 2\beta \gamma \frac{\gamma^s}{s} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^T 1_{\{|a_k^T u - a_k^T u_0| > \epsilon \|u - u_0\|\}} \right\|
\]

where in the second inequality we have used the fact that because \( h \left( |a_k^T v|^2 \right) = 0 \) when \(|a_k^T v|^2 \geq \gamma\) there holds

\[
(a_k^T v)^t h \left( |a_k^T v|^2 \right) \leq \gamma^{t/2},
\]

and the last equality follows from the assumption \(||u - u_0|| \leq \epsilon\). Note that in the above calculation, it requires \(s \geq 1\). However, when \(s = 0\), we have

\[
\frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^T u)^s h \left( |a_k^T u|^2 \right) - (a_k^T u_0)^s h \left( |a_k^T u_0|^2 \right) \right\} 1_{\{|a_k^T u| \leq \beta, |a_k^T u_0| \leq \beta\}} (a_k^T v)^t h \left( |a_k^T v|^2 \right) a_k a_k^T = 0,
\]

and hence the upper bound still holds.
Region $[0, \beta] \times (\beta, \gamma)$ or $(\beta, \gamma) \times [0, \beta]$

$$\left\| \frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^\top u)^s h \left( |a_k^\top u|^2 \right) - (a_k^\top u_0)^s h \left( |a_k^\top u_0|^2 \right) \right\} \mathbb{1}_{\{ |a_k^\top u|^2 \leq \beta, \beta < |a_k^\top u_0|^2 \leq \gamma \}} (a_k^\top v)^t h \left( |a_k^\top v|^2 \right) a_k a_k^\top \right\| \leq \left( \beta^\frac{s}{2} + \gamma^\frac{s}{2} \right) \gamma^\frac{s}{2} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^\top \mathbb{1}_{\{ |a_k^\top u_0| > \sqrt{\gamma} \}} \right\|$$

and

$$\left\| \frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^\top u)^s h \left( |a_k^\top u|^2 \right) - (a_k^\top u_0)^s h \left( |a_k^\top u_0|^2 \right) \right\} \mathbb{1}_{\{ |a_k^\top u|^2 \leq \gamma, |a_k^\top u_0|^2 \leq \beta \}} (a_k^\top v)^t h \left( |a_k^\top v|^2 \right) a_k a_k^\top \right\| \leq \left( \beta^\frac{s}{2} + \gamma^\frac{s}{2} \right) \gamma^\frac{s}{2} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^\top \mathbb{1}_{\{ |a_k^\top u_0| > \sqrt{\gamma} \}} \right\| .$$

Region $[0, \beta] \times (\gamma, \infty)$ or $(\gamma, \infty) \times [0, \beta]$

$$\left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^\top u)^s h \left( |a_k^\top u|^2 \right) \mathbb{1}_{\{ |a_k^\top u|^2 \leq \beta, |a_k^\top u_0|^2 > \gamma \}} (a_k^\top v)^t h \left( |a_k^\top v|^2 \right) a_k a_k^\top \right\| \leq \beta^\frac{s}{2} \gamma^\frac{t}{2} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^\top \mathbb{1}_{\{ |a_k^\top u_0| > \sqrt{\gamma} \}} \right\|$$

and

$$\left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^\top u)^s h \left( |a_k^\top u|^2 \right) \mathbb{1}_{\{ |a_k^\top u|^2 > \gamma, |a_k^\top u_0|^2 \leq \beta \}} (a_k^\top v)^t h \left( |a_k^\top v|^2 \right) a_k a_k^\top \right\| \leq \beta^\frac{s}{2} \gamma^\frac{t}{2} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^\top \mathbb{1}_{\{ |a_k^\top u_0| > \sqrt{\gamma} \}} \right\| .$$

Region $(\beta, \gamma) \times (\beta, \gamma)$

$$\left\| \frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^\top u)^s h \left( |a_k^\top u|^2 \right) - (a_k^\top u_0)^s h \left( |a_k^\top u_0|^2 \right) \right\} \mathbb{1}_{\{ |a_k^\top u|^2 \leq \beta, \beta < |a_k^\top u_0|^2 \leq \gamma \}} (a_k^\top v)^t h \left( |a_k^\top v|^2 \right) a_k a_k^\top \right\| \leq 2 \gamma^\frac{t+s}{2} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^\top \mathbb{1}_{\{ |a_k^\top u_0| > \sqrt{\gamma} \}} \right\| .$$

Region $(\beta, \gamma) \times (\gamma, \infty)$ or $(\gamma, \infty) \times (\beta, \gamma)$

$$\left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^\top u)^s h \left( |a_k^\top u|^2 \right) \mathbb{1}_{\{ |a_k^\top u|^2 \leq \gamma, |a_k^\top u_0|^2 \geq \gamma \}} (a_k^\top v)^t h \left( |a_k^\top v|^2 \right) a_k a_k^\top \right\| \leq \gamma^\frac{s+t}{2} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^\top \mathbb{1}_{\{ |a_k^\top u_0| > \sqrt{\gamma} \}} \right\| .$$
and

\[
\left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^T u_0)^{s}h \left( |a_k^T u_0|^2 \right) \mathbb{1}_{\{|a_k^T u_0|^2 > \gamma, \beta < |a_k^T u_0|^2 \leq \gamma\}} (a_k^T v)^{t}h \left( |a_k^T v|^2 \right) a_k a_k^T \right\| \\
\leq \gamma^{\frac{s+t}{2}} \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^T \mathbb{1}_{\{|a_k^T u| > \sqrt{\gamma} \|u\| \}} \right\| ,
\]

Combining the bounds from (a) to (e) together and noting that the second term in (17) can be bounded similarly to the first one yields that

\[
\left\| \frac{1}{m} \sum_{k=1}^{m} \left\{ (a_k^T u)^{s}(a_k^T v)^{t}h \left( |a_k^T u|^2 \right) h \left( |a_k^T v|^2 \right) a_k a_k^T - (a_k^T u_0)^{s}(a_k^T v_0)^{t}h \left( |a_k^T u_0|^2 \right) h \left( |a_k^T v_0|^2 \right) a_k a_k^T \right\} \right\|
\leq \epsilon \cdot \max \{|s, t| \gamma^{\frac{s+t}{2}} + \beta^{\frac{s+t}{2}} \gamma^{\frac{s+t}{2}} \left( \epsilon^{-1} e^{-0.49 \epsilon^{-2}} + \epsilon \right) + \left( \beta^{\frac{s+t}{2}} + \gamma^{\frac{s+t}{2}} \right) \left( \sqrt{\beta} e^{-0.49 \gamma} + \epsilon \right),
\]

where each term in the first inequality respectively corresponds the bound for (a) to (e) after applying Lemmas A.1 and A.2 and in the second inequality we have used the fact \( 1 < \beta < \gamma \).

By the same splitting scheme, we can similarly show that

\[
\left\| \mathbb{E} \left[ (a_k^T u)^{s}(a_k^T v)^{t}h \left( |a_k^T u|^2 \right) h \left( |a_k^T v|^2 \right) a_k a_k^T \right] - \mathbb{E} \left[ (a_k^T u_0)^{s}(a_k^T v_0)^{t}h \left( |a_k^T u_0|^2 \right) h \left( |a_k^T v_0|^2 \right) a_k a_k^T \right] \right\|
\leq \epsilon \cdot \max \{|s, t| \gamma^{\frac{s+t}{2}} + \beta^{\frac{s+t}{2}} \gamma^{\frac{s+t}{2}} \left( \epsilon^{-1} e^{-0.49 \epsilon^{-2}} + \epsilon \right) + \left( \beta^{\frac{s+t}{2}} + \gamma^{\frac{s+t}{2}} \right) \left( \sqrt{\beta} e^{-0.49 \gamma} + \epsilon \right),
\]

Then the proof is complete after combining (18), (18) and (19) together and using the triangular inequality.

### 4.2 Proof of Lemma 3.2

A direct calculation yields that

\[
\left\| \mathbb{E} \left[ (a_k^T u)^{s}(a_k^T v)^{t}h \left( \frac{|a_k^T u|^2}{\|u\|^2} \right) h \left( \frac{|a_k^T v|^2}{\|v\|^2} \right) a_k a_k^T \right] - \mathbb{E} \left[ (a_k^T u_0)^{s}(a_k^T v_0)^{t}h \left( \frac{|a_k^T u_0|^2}{\|u_0\|^2} \right) h \left( \frac{|a_k^T v_0|^2}{\|v_0\|^2} \right) a_k a_k^T \right] \right\|
\leq \max_{\|q\|=1} \left\| \mathbb{E} \left[ (a_k^T u)^{s}(a_k^T v)^{t}(a_k q)^2 \left( 1 - h \left( \frac{|a_k^T u|^2}{\|u\|^2} \right) \right) \right] \right\|
\leq \max_{\|q\|=1} \mathbb{E} \left[ |a_k^T u|^s |a_k^T v|^t |a_k q|^2 \left( 1 - h \left( \frac{|a_k^T u|^2}{\|u\|^2} \right) \right) \right] \]

\[
\leq \max_{\|q\|=1} \mathbb{E} \left[ |a_k^T u|^s |a_k^T v|^t |a_k q|^2 \left( \mathbb{1}_{\{|a_k^T u| > \sqrt{\gamma} \|u\| \}} + \mathbb{1}_{\{|a_k^T v| > \sqrt{\gamma} \|v\| \}} \right) \right]
\]

20
and where in the third inequality we have used the facts

\[ \mathbb{E}[(a_k^T u)^{2s}|a_k^T v|^{2t}|a_k^T q|^4] \leq \max_{||q||=1} \left( \mathbb{E}[(a_k^T u)^{2s}|a_k^T v|^{2t}|a_k^T q|^4] \right)^{1/2} \left( \sqrt{\mathbb{E} \left[ \mathbb{1}\{a_k^T u > \sqrt{\beta}||u||\} \right]} + \sqrt{\mathbb{E} \left[ \mathbb{1}\{a_k^T v > \sqrt{\beta}||v||\} \right]} \right) \]

\[ \lesssim \left( \max_{||q||=1} \mathbb{E}[a_k^T q] \right)^{1/2} \left( \mathbb{E}[a_k^T u]^{4s}|a_k^T v|^{4t} \right)^{1/2} \sqrt{\frac{2}{\pi\beta}} e^{-\frac{\beta}{\pi}} \]

\[ \lesssim \left( \mathbb{E}[a_k^T u]^{8s} \right)^{1/8} \left( \mathbb{E}[a_k^T v]^{8t} \right)^{1/8} e^{-0.25\beta} \]

\[ \lesssim (8s)!^{1/8} ((8t)!!)^{1/8} ||u||^s ||v||^t \cdot e^{-0.25\beta}, \]

where the fourth inequality follows from Hölder’s inequality and the fact

\[ \mathbb{E} \left[ \mathbb{1}\{||a_k^T u|| > \sqrt{\beta}||u||\} \right] = 2 \int_{\sqrt{\beta}}^{\infty} \frac{1}{t\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \leq \sqrt{\frac{2}{\pi}} \int_{\sqrt{\beta}}^{\infty} t \sqrt{2\pi} e^{-\frac{t^2}{2}} dt \leq \sqrt{\frac{2}{\pi\beta}} e^{-\frac{\beta}{\pi}} \]

as well as \( \beta > 1 \), and the fifth and sixth inequalities hold as the \( 2k \)-th moment of a standard Gaussian variable is \( (2k)!! \).

### 4.3 Proof of Lemma 3.3

Noting that

\[ \frac{1}{m} ||y||_1 = \frac{1}{m} \sum_{k=1}^{m} (a_k^T x)^2 = \left| x^T \left( \frac{1}{m} \sum_{k=1}^{m} a_k a_k^T \right) x \right|, \]

it follows from Lemma [A.1] that \( \frac{1}{m} ||x||^2 \leq \frac{1}{m} ||y||_1 \leq \frac{2}{m} ||x||^2 \) holds with probability \( 1 - e^{\Omega(m)} \) provided \( m \geq n \). Thus, on the same event, we have

\[ \left| \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^s (a_k^T x)^t h \left( \frac{a_k^T z}{||z||^2} \right) \left[ h \left( \frac{m|a_k^T x|^2}{||y||_1} \right) - h \left( \frac{|a_k^T x|^2}{||x||^2} \right) \right] a_k a_k^T \right| \]

\[ \leq \left| \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^s |a_k^T x|^t h \left( \frac{m|a_k^T x|^2}{||y||_1} \right) \left[ h \left( \frac{|a_k^T x|^2}{||y||_1} \right) - h \left( \frac{|a_k^T x|^2}{||x||^2} \right) \right] a_k a_k^T \right| \]

\[ \leq \left| \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^s |a_k^T x|^t h \left( \frac{|a_k^T z|^2}{||z||^2} \right) \cdot \mathbb{1}\{|a_k^T z|^2 < \gamma ||z||^2\} \cdot \mathbb{1}\{|a_k^T x|^2 \leq 2|\gamma||x||^2\} a_k a_k^T \right| \]

\[ \lesssim ||z||^s ||x||^t \cdot 2^{\frac{s}{2} \gamma^{\frac{s}{2} + \frac{t}{2}}} \left( \sqrt{\beta} e^{-0.245\beta} + \epsilon \right), \]

where in the third inequality we have used the facts

\[ |a_k^T z|^s h \left( \frac{|a_k^T z|^2}{||z||^2} \right) \cdot \mathbb{1}\{|a_k^T z|^2 < \gamma ||z||^2\} \leq \gamma^{\frac{s}{2}} ||z||^s \]

and

\[ |a_k^T x|^t \mathbb{1}\{|a_k^T x|^2 \leq 2|\gamma||x||^2\} \leq 2^{\frac{s}{2} \gamma^{\frac{s}{2} + \frac{t}{2}}} ||x||^t \mathbb{1}\{|a_k^T x|^2 \geq \sqrt{\beta} ||x||^2\}, \]

and the last inequality holds with probability exceeding \( 1 - e^{\Omega(m^2)} \) provided \( m \geq \epsilon^2 \log \epsilon^{-1} \cdot n \) (see Lemma [A.2]).
4.4 Proof of Lemma 3.4

It follows from Lemma A.1 that \( \frac{1}{m} \| x \|^2 \leq \frac{1}{m} \| y \|_1 \) holds with probability \( 1 - e^{\Omega(m)} \) provided \( m \gtrsim n \). Thus, on the same event, we have

\[
\left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^\top z)(a_k^\top x)^t g \left( \frac{a_k^\top z}{\| z \|^2} \right) h \left( \frac{m|a_k^\top x|^2}{y_1} \right) a_k a_k^\top \right\|
\leq \left\| \frac{1}{m} \sum_{k=1}^{m} a_k^\top z|a_k^\top x|^t g \left( \frac{a_k^\top z}{\| z \|^2} \right) 1_{\{ \beta \| z \|^2 < |a_k^\top z|^2 < \gamma \| z \|^2 \}} \cdot 1_{\{|a_k^\top x|^2 < 2\gamma \| x \|^2 \}} |a_k a_k^\top| \right\|
\leq \| z \|^2 \| x \|^t \cdot 2^\frac{t}{2} \gamma^\frac{s+1}{2} \left( \sqrt{\beta} e^{-0.49 \beta} + \epsilon \right),
\]

where the last inequality holds with probability exceeding \( 1 - e^{\Omega(m^2)} \) provided \( m \gtrsim \epsilon^{-2} \log \epsilon^{-1} \cdot n \); see Lemma A.2.

4.5 Proof of Lemma 3.5

Firstly, similar to the proof of Lemma 3.4 \( \frac{1}{m} \| x \|^2 \leq \frac{1}{m} \| y \|_1 \) holds with probability \( 1 - e^{\Omega(m)} \) provided \( m \gtrsim n \). Thus, we have

\[
\left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^\top z)^t (a_k^\top x)^t g \left( \frac{a_k^\top z}{\| z \|^2} \right) h \left( \frac{m|a_k^\top x|^2}{y_1} \right) z a_k^\top \right\|
\leq \max_{\| u \|=\| v \|=1} \frac{1}{m} \sum_{k=1}^{m} |a_k^\top z|^t |a_k^\top x|^t g \left( \frac{a_k^\top z}{\| z \|^2} \right) 1_{\{ \beta \| z \|^2 < |a_k^\top z|^2 < \gamma \| z \|^2 \}} \cdot 1_{\{|a_k^\top x|^2 < 2\gamma \| x \|^2 \}} \cdot |u^\top z| \cdot |a_k^\top v|
\leq \| z \|^s+1 \| x \|^t \cdot 2^\frac{t}{2} \gamma^\frac{s+t}{2} \max_{\| v \|=1} \frac{1}{m} \sum_{k=1}^{m} 1_{\{|a_k^\top z| > \sqrt{\beta} \| z \| \}} |a_k^\top v|^2
\leq \| z \|^s+1 \| x \|^t \cdot 2^\frac{t}{2} \gamma^\frac{s+t}{2} \sqrt{\beta} e^{-0.49 \beta} + \epsilon
\leq \| z \|^s+1 \| x \|^t \cdot 2^\frac{t}{2} \gamma^\frac{s+t+1}{2} \left( e^{-0.245 \beta} + \epsilon \right),
\]

where the fourth inequality holds with probability exceeding \( 1 - e^{\Omega(m^2)} \) provided \( m \gtrsim \epsilon^{-2} \log \epsilon^{-1} \cdot n \) which follows from Lemma A.2.

\[
\frac{1}{m} \sum_{k=1}^{m} 1_{\{|a_k^\top z| > \sqrt{\beta} \| z \| \}} |a_k^\top v|^2 = \left\| v^\top \left( \frac{1}{m} \sum_{k=1}^{m} 1_{\{|a_k^\top z| > \sqrt{\beta} \| z \| \}} a_k a_k^\top \right) v \right\|
\leq \left\| \frac{1}{m} \sum_{k=1}^{m} 1_{\{|a_k^\top z| > \sqrt{\beta} \| z \| \}} a_k a_k^\top \right\|
\]
\[ \leq \sqrt{\beta} e^{-0.49\beta} + \epsilon, \]

and the last inequality follows from the fact \( 1 < \beta < \gamma \) and \( \sqrt{a + b} \leq \sqrt{a} + \sqrt{b} \).

### 4.6 Proof of Lemma 3.6

It follows from Lemma [A.1] that \( \frac{1}{m} \|x\|^2 \leq \frac{1}{m} \|y\|_1 \) holds with probability \( 1 - e^{\Omega(m)} \) provided \( m \gtrsim n \).

Then a simple algebra yields that

\[
\begin{align*}
\left| \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^t (a_k^T x)^t g \left( \frac{\|a_k^T z\|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) \right| \\
\leq \frac{1}{m} \sum_{k=1}^{m} |a_k^T z|^t |a_k^T x|^t g \left( \frac{\|a_k^T z\|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) \\
\leq \frac{1}{m} \sum_{k=1}^{m} |a_k^T z|^t |a_k^T x|^t \mathbb{1}_{\{\beta \|z\|^2 \leq |a_k^T z|^2 \leq \gamma \|z\|^2 \}} \mathbb{1}_{\{|a_k^T x|^2 < 2\gamma \|z\|^2 \}} \\
\leq \gamma^{\frac{t-2}{2}} \|z\|^t \cdot 2^{t} \gamma^{\frac{t-2}{2}} \|x\|^t \cdot \frac{1}{m} \sum_{k=1}^{m} |a_k^T z|^2 \mathbb{1}_{\{|a_k^T z| > \sqrt{\beta} \|z\| \}} \\
\lesssim \|z\|^t \cdot 2^{t} \gamma^{\frac{t-2}{2}} \left( \sqrt{\beta} e^{-0.49\beta} + \epsilon \right),
\end{align*}
\]

where the last inequality holds with probability exceeding \( 1 - e^{\Omega(m \epsilon^2)} \) provided \( m \gtrsim \epsilon^{-2} \log \epsilon^{-1} \cdot n \); see Lemma [A.2].

### 5 Conclusion and outlook

A new loss function has been constructed for solving random systems of quadratic equations, which does not have spurious local minima when the sampling complexity is optimal. This paper has focused on the real-valued problem, and we will leave the examination of the complex case to future work. For the complex case, it is interesting to see whether the same loss function is still well-behaved under the optimal sampling complexity, or a more delicate activation function should be adopted. In addition, the technique presented in this paper may apply equally to the problem of reconstructing a general low rank matrix from symmetric rank-1 projections [15, 26, 47].

As stated at the beginning of this paper, the problem of solving systems of quadratic equations can be cast as a rank-1 matrix recovery problem. To see this, let \( A \) be a linear operator from \( n \times n \) symmetric matrices to vectors of length \( m \), defined as

\[
A(Z) = \left\{ \langle Z, a_k a_k^T \rangle \right\}_{k=1}^{m}, \quad \forall \ Z \in \mathbb{R}^{n \times n} \text{ being symmetric}. \tag{20}
\]

Then a simple algebra yields that

\[
y_k = |a_k^T x|^2 = \langle a_k a_k^T, X \rangle,
\]

where \( X = xx^T \) is the lift matrix defined associated with \( x \). Noticing the one to one correspondence between \( X \) and \( x \), instead of reconstructing \( x \), one can attempt to reconstruct \( X \) by seeking a
rank-1 positive semidefinite matrix which fits the measurements as well as possible:

$$\min_Z \frac{1}{2} \| A(Z) - y \|^2 \quad \text{subject to} \quad \text{rank}(Z) = 1 \text{ and } Z \succeq 0.$$ \hfill (21)

Note that the geometric landscape analysis presented in this paper as well as that in [36] are carried out in the vector space. Instead, one can consider the geometric landscape of the loss function $\frac{1}{2} \| A(Z) - y \|^2$ on the embedded manifold of positive semidefinite rank-1 (or general rank-$r$) matrices under the rank-1 measurements. Moreover, it is worth studying whether there exists a loss function on the lift matrix space which is well-behaved under the condition of optimal sampling complexity.

A Auxiliary lemmas

Lemma A.1 ([40, 13]). Assume $a_k \sim \mathcal{N}(0, I_n)$, $k = 1, \ldots, m$, are independent. Then

$$\frac{1}{2} \leq \left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^\top \right\| \leq 2$$

hold with probability at least $1 - 2e^{-\Omega(m)}$ provided $m \gtrsim n$.

Lemma A.2 ([9]). Fix $\eta \geq 1$ and let $\epsilon \in (0, 1)$ be a sufficiently small constant. Assume $a_k \sim \mathcal{N}(0, I_n)$, $k = 1, \ldots, m$, are independent. Then

$$\left\| \frac{1}{m} \sum_{k=1}^{m} a_k a_k^\top 1_{\{|a_k^\top z| > \eta \|z\|\}} \right\| \lesssim \eta e^{-0.49\eta^2} + \epsilon$$

holds uniformly for all $\|z\| \neq 0$ with probability exceeding $1 - 2e^{-\Omega(m^2)}$ provided $m \gtrsim \epsilon^{-2} \log^{-1} \epsilon^{-1} \cdot n$.

B Gradient and Hessian of the loss function

Recall that

$$f(z) = \frac{1}{2m} \sum_{k=1}^{m} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right).$$

By the chain rule we have

$$\nabla f(z) = \frac{1}{m} \sum_{k=1}^{m} 2 \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top z$$

$$+ \frac{1}{m} \sum_{k=1}^{m} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top z$$

$$- \frac{1}{m} \sum_{k=1}^{m} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) \frac{(a_k^\top z)^2}{\|z\|^4}.$$
In order to compute $\nabla^2 f(z)$, let

$$g_{1k} = 2 \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top,$$

$$g_{2k} = \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top,$$

$$g_{3k} = - \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) (a_k^\top z^2) z.$$

Then we have

$$\nabla^2 f(z) = \frac{1}{m} \sum_{k=1}^m J_{g_{1k}} + J_{g_{2k}} + J_{g_{3k}},$$

where $J_{g_{1k}}$, $J_{g_{2k}}$ and $J_{g_{3k}}$ are the Jacobian matrices of $g_{1k}$, $g_{2k}$ and $g_{3k}$ respectively, given by

$$J_{g_{1k}} = 2 \left( 3(a_k^\top z)^2 - (a_k^\top x)^2 \right) h \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top$$

$$+ \frac{4}{\|z\|^2} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) (a_k^\top z)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top$$

$$- \frac{4}{\|z\|^4} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 (a_k^\top z)^3 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k z^\top,$$

$$J_{g_{2k}} = \frac{1}{\|z\|^2} \left( 5(a_k^\top z)^4 - 6(a_k^\top z)^2 (a_k^\top x)^2 + (a_k^\top x)^4 \right) h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top$$

$$+ \frac{2}{\|z\|^4} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 (a_k^\top z)^2 h'' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top$$

$$- \frac{2}{\|z\|^6} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 (a_k^\top z)^3 h'' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k z^\top$$

$$- \frac{2}{\|z\|^4} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 (a_k^\top z)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k z^\top,$$

$$J_{g_{3k}} = - \frac{1}{\|z\|^4} \left( 6(a_k^\top z)^5 - 8(a_k^\top z)^3 (a_k^\top x)^2 + 2(a_k^\top z)(a_k^\top x)^4 \right) h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) z a_k^\top$$

$$- \frac{2}{\|z\|^6} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 (a_k^\top z)^3 h'' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) z a_k^\top$$

$$+ \frac{2}{\|z\|^8} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 (a_k^\top z)^4 h'' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) z z^\top$$

$$+ \frac{4}{\|z\|^6} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 (a_k^\top z)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) z z^\top$$

$$- \frac{1}{\|z\|^4} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) (a_k^\top z)^2 I.$$
Table 2: Computational results from computing \( f'(z) \) and \( f''(z) \) via the formulas and the finite difference schemes. Here, \( x = 1 \), \( \{a_k\}_{k=1}^m \) \((m = 128)\) is a set of standard Gaussian random variables, and the results for three \( z \)'s in \( \{4.3042, 1.7588, 0.5544\} \) are presented. Each \( z \) is uniformly sampled from \([0, 10]\).

|                     | \( f'(z) \) for three \( z \)'s | \( f''(z) \) for three \( z \)'s |
|---------------------|----------------------------------|----------------------------------|
| By formulas         | 787.0769 38.4171 -4.0066         | 569.4601 86.3959 -0.8143         |
| By finite difference| 787.0769 38.4171 -4.0066         | 569.4601 86.3959 -0.8143         |

It is worth noting that even though each Jacobian matrix is not symmetric, their sum is indeed symmetric which satisfies the symmetric property of a Hessian matrix. To see this, adding all the terms involving \( a_kz^\top \) and \( za_k^\top \) together gives

\[
- \frac{2}{\|z\|^6} \left( (a_k^\top z)^2 - (a_k^\top x)^2 \right) (a_k^\top z)^3 h'' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) (a_k^\top z + za_k^\top) \\
- \frac{1}{\|z\|^4} \left( 6(a_k^\top z)^5 - 8(a_k^\top z)^3(a_k^\top x)^2 + 2(a_k^\top z)(a_k^\top x)^4 \right) h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) (a_k^\top z + za_k^\top).
\]

To further check the correctness of our calculations, we consider the special case when \( n = 1 \), and compare the values of \( f'(z) \) and \( f''(z) \) computed using the derived formulas with that computed via the following finite difference schemes:

\[
f'(z) \approx \frac{f(z + \epsilon) - f(z - \epsilon)}{2\epsilon} \quad \text{and} \quad f''(z) \approx \frac{f(z + \epsilon) - 2f(z) + f(z - \epsilon)}{\epsilon^2},
\]

where \( \epsilon > 0 \) is a small constant (here we choose \( \epsilon = 10^{-5} \)). Table 2 includes the computational results for the fixed \( x = 1 \) and a few randomly generated \( z \).

**B.1 Proof of (6)**

Denote by \( B_i, i = 2, \ldots, 12 \) the spectral norm of the \( i \)-th matrix in the Hessian expression (22).

**Bound for** \( B_2 \)

\[
B_2 \leq \frac{4}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^m \left[ (a_k^\top z)^4 - (a_k^\top z)^2(a_k^\top x)^2 \right] h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|
\]

\[
\leq \frac{4}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^m (a_k^\top z)^4 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|
\]

\[
+ \frac{4}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^m (a_k^\top z)^2(a_k^\top x)^2 h' \left( \frac{|a_k^\top z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^\top x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|
\]

\[
\lesssim \frac{4}{\|z\|^2} \cdot \|z\|^4 \cdot |h'|_\infty \cdot \gamma^2 \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right) + \frac{4}{\|z\|^2} \cdot \|z\|^2 \cdot |z\|_\infty \cdot 2\gamma^2 \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right)
\]

\[
\lesssim \gamma^2 \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right) \cdot |h'|_\infty \max \{\|z\|^2, |x|^2\},
\]

where the third inequality follows from Lemma 3.4 with \( (s, t) = (4, 0), (2, 2) \), respectively.
Bound for $B_3$

$$B_3 \leq \frac{4}{\|z\|^4} \left\| \frac{1}{m} \sum_{k=1}^{m} \left[ (a_k^T z)^5 - (a_k^T z)^2 (a_k^T x)^2 + (a_k^T x)^4 \right] h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k z^\top \right\|$$

$$\leq \frac{4}{\|z\|^4} \left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^5 h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k z^\top \right\|$$

$$+ \frac{4}{\|z\|^4} \left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^2 (a_k^T x)^2 h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k z^\top \right\|$$

$$\lesssim \frac{4}{\|z\|^4} \cdot \|z\|^6 \cdot |h'|_\infty \cdot \gamma^3 \left( e^{-0.245 \beta} + \sqrt{\epsilon} \right) + \frac{4}{\|z\|^4} \cdot \|z\|^4 \|x\|^2 \cdot |h'|_\infty \cdot 2\gamma^3 \left( e^{-0.245 \beta} + \sqrt{\epsilon} \right)$$

$$\lesssim \gamma^3 \left( e^{-0.245 \beta} + \sqrt{\epsilon} \right) \cdot |h'|_\infty \max \left\{ \|z\|^2, \|x\|^2 \right\},$$

where the third inequality follows from Lemma 3.5 with $(s, t) = (5, 0), (3, 2)$, respectively.

Bound for $B_4$

$$B_4 \leq \frac{1}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^{m} \left[ 5(a_k^T z)^4 - 6(a_k^T z)^2 (a_k^T x)^2 + (a_k^T x)^4 \right] h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|$$

$$\leq \frac{1}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^{m} 5(a_k^T z)^4 h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|$$

$$+ \frac{1}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^{m} 6(a_k^T z)^2 (a_k^T x)^2 h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|$$

$$+ \frac{1}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^T x)^4 h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|$$

$$\lesssim \frac{1}{\|z\|^2} \cdot 5\|z\|^4 \cdot |h'|_\infty \cdot \gamma^2 \left( \sqrt{\beta e^{-0.49 \beta}} + \epsilon \right) + \frac{1}{\|z\|^2} \cdot 6\|z\|^2 \cdot \|x\|^2 \cdot |h'|_\infty \cdot 2\gamma^2 \left( \sqrt{\beta e^{-0.49 \beta}} + \epsilon \right)$$

$$+ \frac{1}{\|z\|^2} \cdot \|x\|^4 \cdot |h'|_\infty \cdot 4\gamma^2 \left( \sqrt{\beta e^{-0.49 \beta}} + \epsilon \right)$$

$$\lesssim \gamma^2 \left( \sqrt{\beta e^{-0.49 \beta}} + \epsilon \right) \cdot |h'|_\infty \max \left\{ \|z\|^2, \|x\|^2, \|x\|^4 \right\},$$

where the third inequality follows from Lemma 3.4 with $(s, t) = (4, 0), (2, 2), (0, 4)$, respectively.

Bound for $B_5$

$$B_5 \leq \frac{2}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^{m} \left[ (a_k^T z)^6 - 2(a_k^T z)^4 (a_k^T x)^2 + (a_k^T z)^2 (a_k^T x)^4 \right] h'' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|$$

$$\leq \frac{2}{\|z\|^2} \left\| \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^6 h'' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) a_k a_k^\top \right\|$$
where the third inequality follows from Lemma 3.4, with \((s, t) = (6, 0), (4, 2), (2, 4)\), respectively.

**Bound for** \(B_6\)

\[
B_6 \leq \frac{2}{\|z\|^6} \left| \sum_{m=1}^{\infty} \left[ (a_k^T z)^7 - 2(a_k^T z)^5(a_k^T x)^2 + (a_k^T z)^3(a_k^T x)^4 \right] h'' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{|m|a_k^T x|^2}{\|y\|^1} \right) a_k z \right|
\]

where the third inequality follows from Lemma 3.5 with \((s, t) = (7, 0), (5, 2), (3, 4)\), respectively.

**Bound for** \(B_7\)

\[
B_7 \leq \frac{2}{\|z\|^4} \left| \sum_{m=1}^{\infty} \left[ (a_k^T z)^5 - 2(a_k^T z)^3(a_k^T x)^2 + (a_k^T z)(a_k^T x)^4 \right] h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{|m|a_k^T x|^2}{\|y\|^1} \right) a_k z \right|
\]
where the third inequality follows from Lemma 3.5 with \((s, t) = (5, 0), (3, 2), (1, 4)\), respectively.

**Bound for \(B_8\)**

\[
B_8 \leq \frac{1}{\|z\|^4} \left[ \frac{1}{m} \sum_{k=1}^{m} \left( 6(a_k^T z)^5 - 8(a_k^T z)^3(a_k^T x)^2 + 2(a_k^T z)(a_k^T x)^4 \right) h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) z a_k^T \right] \
\]

\[
\leq \frac{1}{\|z\|^4} \left[ \frac{1}{m} \sum_{k=1}^{m} 6(a_k^T z)^5 h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) z a_k^T \right] \
+ \frac{1}{\|z\|^4} \left[ \frac{1}{m} \sum_{k=1}^{m} 8(a_k^T z)^3(a_k^T x)^2 h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) z a_k^T \right] \
+ \frac{1}{\|z\|^4} \left[ \frac{1}{m} \sum_{k=1}^{m} 2(a_k^T z)(a_k^T x)^4 h' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) z a_k^T \right] \
\]

\[
\lesssim \frac{1}{\|z\|^4} \cdot 6\|z\|^6 \cdot |h'|_\infty \cdot \gamma^3 \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) + \frac{1}{\|z\|^4} \cdot 8\|z\|^4 \|x\|^2 \cdot |h'|_\infty \cdot 2\gamma^3 \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) \
+ \frac{1}{\|z\|^4} \cdot 2\|z\|^2 \|x\|^4 \cdot |h'|_\infty \cdot 4\gamma^3 \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) \
\lesssim 30\gamma^3 \left( e^{-0.245\beta} + \sqrt{\epsilon} \right) \cdot |h'|_\infty \max \left\{ \|z\|^2, \|x\|^2, \frac{|x|^4}{\|z\|^2} \right\},
\]

where the third inequality follows from Lemma 3.5 with \((s, t) = (5, 0), (3, 2), (1, 4)\), respectively.

**Bound for \(B_9\)**

\[
B_9 \leq \frac{2}{\|z\|^6} \left[ \frac{1}{m} \sum_{k=1}^{m} \left( a_k^T z \right)^7 - 2(a_k^T z)^5(a_k^T x)^2 + (a_k^T z)^3(a_k^T x)^4 \right) h'' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) z a_k^T \right] \
\]

\[
\leq \frac{2}{\|z\|^6} \left[ \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^7 h'' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) z a_k^T \right] \
+ \frac{2}{\|z\|^6} \left[ \frac{1}{m} \sum_{k=1}^{m} 2(a_k^T z)^5(a_k^T x)^2 h'' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) z a_k^T \right] \
+ \frac{2}{\|z\|^6} \left[ \frac{1}{m} \sum_{k=1}^{m} (a_k^T z)^3(a_k^T x)^4 h'' \left( \frac{|a_k^T z|^2}{\|z\|^2} \right) h \left( \frac{m|a_k^T x|^2}{\|y\|_1} \right) z a_k^T \right] \
\]

29
where the third inequality follows from Lemma 3.5 with \((s, t) = (7, 0), (5, 2), (3, 4)\), respectively.

**Bound for \(B_{10}\)**

\[ B_{10} \leq \frac{2}{\|z\|^8} \left\| \frac{1}{m} \sum_{k=1}^{m} \left( (a_k^T z)^8 - 2(a_k^T z)^6(a_k^T x)^2 + (a_k^T z)^4(a_k^T x)^4 \right) \right\| = (7 \cdot \gamma^4) \left\{ \|z\|^2, \|x\|^2, \frac{\|x\|^4}{\|z\|} \right\}, \]

where the third inequality follows from the fact \(\|zz^T\| = \|z\|^2\) and Lemma 3.6 with \((s, t) = (8, 0), (6, 2), (4, 4)\), respectively.

**Bound for \(B_{11}\)**

\[ B_{11} \leq \frac{4}{\|z\|^6} \left\| \frac{1}{m} \sum_{k=1}^{m} \left( (a_k^T z)^6 - 2(a_k^T z)^4(a_k^T x)^2 + (a_k^T z)^2(a_k^T x)^4 \right) \right\| \leq \frac{8}{\|z\|^6} \cdot \|h'|_\infty \cdot \gamma^3 (\sqrt{\beta e^{-0.49\beta}} + \epsilon) + 2 \cdot \|z\|^8 \cdot \|x\|^2 \cdot \|h''|_\infty \cdot 2\gamma^4 (\sqrt{\beta e^{-0.49\beta}} + \epsilon) \]
where the third inequality follows from Lemma 3.6 with $(s, t) = (6, 0), (4, 2), (2, 4)$, respectively.

**Bound for $B_{12}$**

\[
B_{12} \leq \frac{1}{\|z\|^4} \left| \frac{1}{m} \sum_k \left[ (a_k^T z)^6 - 2(a_k^T z)^4(a_k^T x)^2 + (a_k^T z)^2(a_k^T x)^4 \right] h' \left( \frac{\|a_k^T z\|^2}{\|z\|^2} \right) h \left( \frac{m\|a_k^T x\|^2}{\|y\|_1} \right) \right|
\]

\[
+ \frac{1}{\|z\|^4} \left| \frac{1}{m} \sum_k 2(a_k^T z)^4(a_k^T x)^2 h' \left( \frac{\|a_k^T z\|^2}{\|z\|^2} \right) h \left( \frac{m\|a_k^T x\|^2}{\|y\|_1} \right) \right|
\]

\[
+ \frac{1}{\|z\|^4} \left| \frac{1}{m} \sum_k (a_k^T z)^2(a_k^T x)^4 h' \left( \frac{\|a_k^T z\|^2}{\|z\|^2} \right) h \left( \frac{m\|a_k^T x\|^2}{\|y\|_1} \right) \right|
\]

\[
\lesssim \frac{1}{\|z\|^6} \cdot \|h'\|_{\infty} \cdot \gamma^3 \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right) + \frac{2}{\|z\|^4} \cdot \|\|z\|^2 \cdot \|x\|^2 \cdot \|h'\|_{\infty} \cdot 2\gamma^3 \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right)
\]

\[
+ \frac{1}{\|z\|^2} \cdot \|z\|^2 \cdot \|x\|^4 \cdot \|h'\|_{\infty} \cdot 4\gamma^3 \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right)
\]

\[
\lesssim \gamma^3 \left( \sqrt{\beta e^{-0.49\beta}} + \epsilon \right) \cdot \|h'\|_{\infty} \max \left\{ \|z\|^2, \|x\|^2, \left. \frac{\|x\|^4}{\|z\|^2} \right\} \right).
\]

where the third inequality follows from Lemma 3.6 with $(s, t) = (6, 0), (4, 2), (2, 4)$, respectively.

Noting that $1 < \beta < \gamma$, combing the above bounds together yields (6).

**References**

[1] A. Anandkumar, R. Ge, and M. Janzamin. Analyzing tensor power method dynamics in overcomplete regime. *The Journal of Machine Learning Research*, 18(22):1–40, 2017.

[2] S. Bahmani and J. Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. arXiv:1610.04210, 2016.

[3] R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345–356, 2006.

[4] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. arXiv preprint arXiv:1605.07221, 2016.

[5] N. Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377, 2016.
[6] N. Boumal, V. Voroninski, and A. S. Bandeira. The non-convex burermonteiro approach works on smooth semidefinite programs. arXiv preprint arXiv:1606.04970, 2016.

[7] O. Bunk, A. Diaz, F. Pfeiffer, C. David, B. Schmitt, and D. K. Satapathy. Diffractive imaging for periodic samples: Retrieving one-dimensional concentration profiles across microfluidic channels. Acta Crystallographica Section A: Foundations of Crystallography, 63(4):306–314, 2007.

[8] J.-F. Cai and K. Wei. Exploiting the structure effectively and efficiently in low rank matrix recovery. arXiv preprint arXiv:1809.03652, 2018.

[9] J.-F. Cai and K. Wei. Solving systems of phaseless equations via Riemannian optimization with optimal sampling complexity. arXiv preprint arXiv:1809.02773, 2018.

[10] E. J. Candès and X. Li. Solving quadratic equations via phaselift when there are about as many equations as unknowns. Foundations of Computational Mathematics, 14(5):1017–1026, 2014.

[11] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. Applied and Computational Harmonic Analysis, 29(2):277–299, 2015.

[12] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. IEEE Transactions on Information Theory, 61(4):1985–2007, 2015.

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

[14] Y. Chen and E. J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. Communications on Pure and Applied Mathematics, 70(5):822–883, 2017.

[15] Y. Chen, Y. Chi, and A. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. IEEE Transactions on Information Theory, 61(7):4034–4059, 2015.

[16] A. Conca, D. Edidin, M. Herbig, and C. Vinzant. An algebraic characterization of injectivity in phase retrieval. arXiv:1312.0158v1, 2013.

[17] D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. Applied and Computational Harmonic Analysis, 42(1):37–64, 2017.

[18] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In International Conference on Machine Learning, pages 1233–1242, 2017.

[19] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In Advances in Neural Information Processing Systems, pages 2973–2981, 2016.

[20] R. Ge and T. Ma. On the optimization landscape of tensor decompositions. In Advances in Neural Information Processing Systems, pages 3653–3663, 2017.
[21] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. arXiv:1610.07531, 2016.

[22] P. Hand and V. Voroninski. An elementary proof of convex phase retrieval in the natural parameter space via the linear program phasemax. arXiv:1611.03935, 2016.

[23] R. Harrison. Phase problem in crystallography. *Journal of the Optical Society of America A*, 10(5):1046–1055, 1993.

[24] H. Jeong and C. S. Güntürk. Convergence of the randomized Kaczmarz method for phase retrieval. arXiv:1706.10291, 2017.

[25] K. Kawaguchi. Deep learning without poor local minima. arXiv preprint arXiv:1605.07110, 2016.

[26] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. 2014. arXiv preprint arXiv:1410.6913.

[27] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

[28] Q. Li, Z. Zhu, and G. Tang. The non-convex geometry of low-rank matrix optimization. arXiv preprint arXiv:1611.03060, 2016.

[29] H. Liu, M.-C. Yue, and A. M.-C. So. On the estimation performance and convergence rate of the generalized power method for phase synchronization. *SIAM Journal on Optimization*, 27(4):2426–2446, 2017.

[30] J. Miao, T. Ishikawa, Q. Shen, and T. Earnesty. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annual Review of Physical Chemistry*, 59:387–410, 2008.

[31] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015.

[32] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.

[33] D. Soudry and Y. Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. arXiv preprint arXiv:1605.08361, 2016.

[34] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Trans. Info. Theory*, 63(2):853–884, 2017.

[35] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017.

[36] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, pages 1–68, 2018.
[37] Y. S. Tan and R. Vershynin. Phase retrieval via randomized Kaczmarz: Theoretical guarantees. arXiv:1706.09993, 2017.

[38] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In International Conference on Machine Learning, pages 964–973, 2016.

[39] L. Venturi, A. S. Bandeira, and J. Bruna. Neural networks with finite intrinsic dimension have no spurious valleys. arXiv preprint arXiv:1802.06384, 2018.

[40] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, Compressed Sensing: Theory and Applications. Cambridge University Press, 2012.

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

[42] I. Waldspurger, A. d’Aspremont, and S. Mallat. Phase recovery, MaxCut and complex semidefinite programming. Mathematical Programming, Series A, 1(2):47–81, 2015.

[43] G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. IEEE Transactions on Information Theory, 64(2):773–794, 2018.

[44] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen. Phase retrieval via reweighted amplitude flow. IEEE Transactions on Signal Processing, 66(11):2818–2833, 2018.

[45] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix completion. arXiv preprint arXiv:1603.06610, 2016.

[46] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix recovery. SIAM Journal on Matrix Analysis and Applications, 37(3):1198–1222, 2016.

[47] C. D. White, R. Ward, and S. Sanghavi. The local convexity of solving quadratic equations. arXiv preprint arXiv:1506.07868, 2015.

[48] C. Yun, S. Sra, and A. Jadbabaie. A critical view of global optimality in deep learning. arXiv preprint arXiv:1802.03487, 2018.

[49] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. Reshaped wirtinger flow and incremental algorithm for solving quadratic system of equations. arXiv preprint arXiv:1605.07719, 2016.

[50] Y. Zhang, Y. Lau, H.-W. Kuo, S. Cheung, A. Pasupathy, and J. Wright. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, page 43814389, 2017.

[51] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. 2016. arXiv preprint arXiv:1605.07051.