Manifold Gradient Descent Solves Multi-Channel Sparse Blind Deconvolution Provably and Efficiently

Laixi Shi Yuejie Chi
Department of Electrical and Computer Engineering
Carnegie Mellon University
Email: {laixis, yuejiec}@andrew.cmu.edu*†

Abstract

Multi-channel sparse blind deconvolution, or convolutional sparse coding, refers to the problem of learning an unknown filter by observing its circulant convolutions with multiple input signals that are sparse. This problem finds numerous applications in signal processing, computer vision, and inverse problems. However, it is challenging to learn the filter efficiently due to the bilinear structure of the observations with respect to the unknown filter and inputs, as well as the sparsity constraint. In this paper, we propose a novel approach based on nonconvex optimization over the sphere manifold by minimizing a smooth surrogate of the sparsity-promoting loss function. It is demonstrated that manifold gradient descent with random initializations will provably recover the filter, up to scaling and shift ambiguity, as soon as the number of observations is sufficiently large under an appropriate random data model. Numerical experiments are provided to illustrate the performance of the proposed method with comparisons to existing ones.

Keywords: nonconvex optimization, multi-channel sparse blind deconvolution, manifold gradient descent

Contents

*The work of L. Shi and Y. Chi is supported in part by ONR under the grants N00014-18-1-2142 and N00014-19-1-2404, by NSF under the grants CAREER ECCS-1818571, CCF-1806154 and CCF-1901199, and the Liang Ji-Dian Graduate Fellowship to L. Shi.

†The results in this paper were presented in part at the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2020.

1 Introduction

In various fields of signal processing, computer vision, and inverse problems, it is of interest to identify the location of sources from traces of responses collected from sensors. For example, neural or seismic recordings can be modeled as the convolution of a pulse shape (i.e. a filter), corresponding to characteristics of neuron or earth wave propagation, with a spike train modeling time of activations (i.e. a sparse input) [1, 2, 3]. Thanks to the advances of sensing technologies, in many applications, one can make multiple observations that share the same filter, but actuated by diverse sparse inputs, either spatially or temporally. Examples include underwater communications [4, 5], neuroscience [6], seismic imaging [7, 8], image deblurring [9, 10], and so on. The goal of this paper is to identify the filter as well as the sparse inputs by leveraging multiple observations in an efficient manner, a problem termed as multi-channel sparse blind deconvolution (MSBD).

Mathematically, we model each observation $\boldsymbol{y}_i \in \mathbb{R}^n$ as a convolution, between a filter $\boldsymbol{g} \in \mathbb{R}^n$, and a sparse input, $\boldsymbol{x}_i \in \mathbb{R}^n$:

$$\boldsymbol{y}_i = \boldsymbol{g} \circledast \boldsymbol{x}_i = \mathcal{C}(\boldsymbol{g})\boldsymbol{x}_i, \quad i = 1, \dots, p, \tag{1}$$

where the total number of observations is given as p. Here, we consider circulant convolution, denoted as $\circledast$, whose operation is expressed equivalently via pre-multiplying a circulant matrix $\mathcal{C}(\boldsymbol{g})$ to the input, defined as

$$\mathcal{C}(\boldsymbol{g}) = \begin{bmatrix} g_1 & g_n & \cdots & g_2 \\ g_2 & g_1 & \cdots & g_3 \\ \vdots & \vdots & \ddots & \vdots \\ g_n & g_{n-1} & \cdots & g_1 \end{bmatrix}. \tag{2}$$

In practice, the circulant convolution is used in situations when the filter $\boldsymbol{g}$ satisfies periodic boundary conditions [11, 12], or as an approximation of the linear convolution when the filter has compact support or decays fast [13, 14]. It is particularly attractive in large-scale problems to accelerate computation by taking advantage of the fast Fourier transform [11, 12].

1.1 Nonconvex Optimization on the Sphere

Our goal is to recover both the filter $\boldsymbol{g}$ and sparse inputs $\{\boldsymbol{x}_i\}_{i=1}^p$ from the observations $\{\boldsymbol{y}_i\}_{i=1}^p$. The problem is challenging due to the bilinear form of the observations with respect to the unknowns, as well as the sparsity constraint. A direct observation tells that the unknowns are not uniquely identifiable, since for any circulant shift $\mathcal{S}_k(\cdot)$ by k entries (defined in Section 1.3) and a non-zero scalar $\beta \neq 0$, we have

$$\boldsymbol{y}_i = (\beta \mathcal{S}_k(\boldsymbol{g})) \circledast \left(\beta^{-1}\mathcal{S}_{-k}(\boldsymbol{x}_i)\right), \tag{3}$$

for $k = 1, \dots, n-1$. Hence, we can only hope to recover $\boldsymbol{g}$ and $\{\boldsymbol{x}_i\}_{i=1}^p$ accurately up to certain circulant shift and scaling factor.

In this paper, we focus on the case that $\mathcal{C}(\boldsymbol{g})$ is invertible, which is equivalent to requiring all Fourier coefficients of $\boldsymbol{g}$ are nonzero. This condition plays a critical role in guaranteeing the identifiability of the model as long as p is large enough [15]. Under this assumption, there exists a unique inverse filter, $\boldsymbol{g}_{\text{inv}} \in \mathbb{R}^n$, such that

$$\mathcal{C}(\boldsymbol{g}_{\text{inv}})\mathcal{C}(\boldsymbol{g}) = \mathcal{C}(\boldsymbol{g})\mathcal{C}(\boldsymbol{g}_{\text{inv}}) = \mathbf{I}. \tag{4}$$

This allows us to convert the bilinear form (1) into a linear form, by multiplying $\mathcal{C}(\boldsymbol{g}_{\text{inv}})$ on both sides:

$$\mathcal{C}(\boldsymbol{g}_{\text{inv}})\boldsymbol{y}_i = \mathcal{C}(\boldsymbol{g}_{\text{inv}})\mathcal{C}(\boldsymbol{g})\boldsymbol{x}_i = \boldsymbol{x}_i, \quad i = 1, \dots, p.$$

Consequently, we can equivalently aim to recover $\boldsymbol{g}_{\text{inv}}$ via exploiting the sparsity of the inputs $\{\boldsymbol{x}_i\}_{i=1}^p$. An immediate thought is to seek a vector $\boldsymbol{h}$ that minimizes the cardinality of $\mathcal{C}(\boldsymbol{h})\boldsymbol{y}_i = \mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}$:

$$\min_{\boldsymbol{h} \in \mathbb{R}^n} \frac{1}{p} \sum_{i=1}^p \left\| \mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h} \right\|_0,$$

where $\|\cdot\|_0$ is the pseudo-ℓ_0 norm that counts the cardinality of the nonzero entries of the input vector. However, this simple formulation is problematic for two obvious reasons:

1) first, due to scaling ambiguity, a trivial solution is $\boldsymbol{h} = \mathbf{0}$;

2) second, the cardinality minimization is computationally intractable.

The first issue can be addressed by adding a *spherical* constraint $\|\boldsymbol{h}\|_2 = 1$ to avoid scaling ambiguity. The second issue can be addressed by relaxing to a convex smooth surrogate that promotes sparsity. In this paper, we consider the function

$$\psi_\mu(z) = \mu \log\cosh(z/\mu), \tag{5}$$

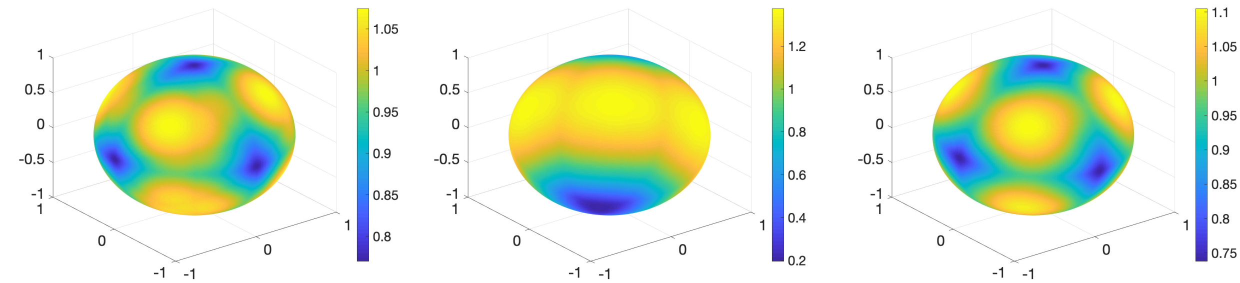
which serves as a convex surrogate of $\|\cdot\|_0$, where $\mu > 0$ controls the smoothness of the surrogate. With slight abuse of notation, we assume $\psi_\mu(\boldsymbol{z}) = \sum_{i=1}^n \psi_\mu(z_i)$ is applied in an entry-wise manner, where $\boldsymbol{z} = [z_i]_{1 \le i \le n}$. Putting them together, we arrive at the following optimization problem:

$$\min_{\boldsymbol{h} \in \mathbb{R}^n} f_o(\boldsymbol{h}) := \frac{1}{p} \sum_{i=1}^p \psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{h}) \quad \text{s.t.} \quad \|\boldsymbol{h}\|_2 = 1, \tag{6}$$

which is a nonconvex optimization problem due to the sphere constraint. As we shall see later, while this approach works well when $\mathcal{C}(\boldsymbol{g})$ is an orthogonal matrix, further care needs to be taken when it is a general invertible matrix in order to guarantee a benign optimization geometry. Following [14, 16], we introduce the following pre-conditioned optimization problem:

$$\min_{\boldsymbol{h} \in \mathbb{R}^n} f(\boldsymbol{h}) = \frac{1}{p} \sum_{i=1}^p \psi_\mu(\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{R}\boldsymbol{h}) \quad \text{s.t.} \quad \|\boldsymbol{h}\|_2 = 1, \tag{7}$$

where $\boldsymbol{R}$ is a pre-conditioning matrix depending only on the observations $\{\boldsymbol{y}_i\}_{i=1}^p$ that we will formally introduce in Section 2.



(a) orthogonal filter no pre-conditioning

(b) general filter no pre-conditioning

(c) general filter with pre-conditioning

Figure 1: An illustration of the landscape of the empirical loss function $f_o(\boldsymbol{h})$ or $f(\boldsymbol{h})$ with or without the pre-conditioning matrix $\boldsymbol{R}$ in $\mathbb{R}^3$, where the sparse inputs are generated according to a Bernoulli-Gaussian model with $p = 30$ observations and activation probability $\theta = 0.3$. (a) orthogonal filter $\mathcal{C}(\boldsymbol{g}) = \mathbf{I}$, no pre-conditioning is applied; (b) a general filter, no pre-conditioning is applied; (c) the same general filter as (b) with pre-conditioning.

1.2 Optimization Geometry and Manifold Gradient Descent

Encouragingly, despite nonconvexity, under a suitable random model of the sparse inputs, the empirical loss functions exhibits benign geometric curvatures as long as the sample size p is sufficiently large. As an illustration, Fig. 1 shows the landscape of $f_o(\boldsymbol{h})$ and $f(\boldsymbol{h})$ when $n = 3$ and $p = 30$, and the sparse inputs $\{\boldsymbol{x}_i\}_{i=1}^p$ follow the standard Bernoulli-Gaussian model (with an activation probability $\theta = 0.3$, see Definition 1). When the filter is orthogonal, e.g. $\mathcal{C}(\boldsymbol{g}) = \mathbf{I}$, it can be seen from Fig. 1 (a) that the function $f_o(\boldsymbol{h})$ in (6) has benign geometry without pre-conditioning, where the local minimizers are approximately all shift and sign-flipped variants of the ground truth (i.e, the basis vectors), and are symmetrically distributed across the sphere. On the other end, for filters that are not orthogonal, the geometry of $f_o(\boldsymbol{h})$ in (6) is less well-posed without pre-conditioning, as illustrated in Fig. 1 (b). By introducing pre-conditioning, which intuitively stretches the loss surface to mirror the orthogonal case, the pre-conditioned loss function $f(\boldsymbol{h})$ given in (7) for the same non-orthogonal filter used in Fig. 1 (b) is much easier to optimize over, as illustrated in Fig. 1 (c).

Motivated by this benign geometry, it is therefore natural to optimize $\boldsymbol{h}$ over the sphere. One simple and low-complexity approach is to minimize $f(\boldsymbol{h})$ over the sphere via (projected) manifold gradient descent (MGD), i.e. for $k = 0, 1, \ldots$

$$\boldsymbol{h}^{(k+1)} := \frac{\boldsymbol{h}^{(k)} - \eta \partial f(\boldsymbol{h}^{(k)})}{\|\boldsymbol{h}^{(k)} - \eta \partial f(\boldsymbol{h}^{(k)})\|_2}, \tag{8}$$

where η is the step size, $\partial f(\boldsymbol{h})$ is the Riemannian manifold gradient with respect to $\boldsymbol{h}$ (defined in Sec. 2.2). Surprisingly, this simple approach works remarkably well even with random initializations for appropriately chosen step sizes. As an illustration, Fig. 2 depicts that MGD converges within a few number of iterations for the problem instance in Fig. 1 (c). Based on such empirical success, our goal is to address the following question: *can we establish theoretical guarantees of MGD to recover the filter for MSBD?*

In this paper, we formally establish the benign geometry of the empirical loss function over the sphere, and prove that MGD, with a small number of random initializations, is guaranteed to recover the filter with high probability in polynomial time. Our result is stated informally below.

Theorem 1 (Informal)**.** *Assume the sparse inputs are generated using a Bernoulli-Gaussian model, where the activation probability $\theta \in (0, 1/3)$. As long as the sample size is sufficiently large, i.e. $p = O(\mathrm{poly}(n))$, manifold gradient descent, initialized from at most $O(\log n)$ independently and uniformly selected points on the sphere, recovers the filter accurately with high probability, for properly chosen μ, and step size η_t.*

Our theorem provides justifications to the empirical success of MGD with random initializations. This result is achieved through an integrated analysis of geometry and optimization. Namely, we identify a union

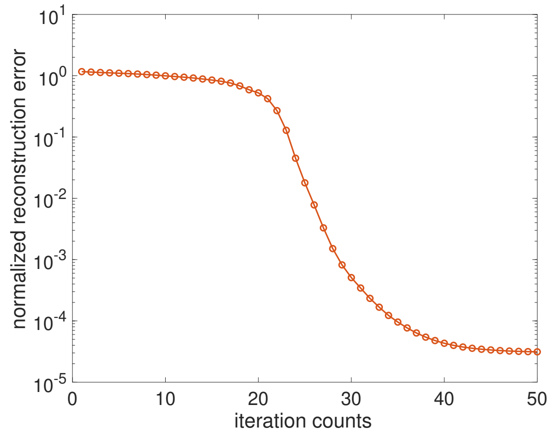


Figure 2: The normalized reconstruction error of MGD with respect to the number of iterations for the problem instance in Fig. 1 (c).

of subsets, corresponding to neighborhoods of equivalent global minimizers, and show that this region has large gradients pointing towards the direction of minimizers. Consequently, if the iterates of MGD lie in this region, and never jump out of it during its execution, we can guarantee that MGD converges to the global minimizers. Luckily, this region is also large enough, so that the probability of a random initialization selected uniformly over the sphere has at least a constant probability falling into the region. By independently initializing a few times, it is guaranteed with high probability at least one of the initializations successfully land into the region of interest and return a faithful estimate of the filter.

1.3 Paper Organization and Notation

The rest of this paper is organized as follows. Section 2 presents the problem formulation and main results, with comparisons to existing approaches. Section 3 outlines the analysis framework and sketches the proof. Section 4 provides numerical experiments on both synthetic and real data with comparisons to existing algorithms. Section 5 further discusses the related literature and we conclude in Section 6 with future directions.

Throughout the paper, we use boldface letters to represent vectors and matrices. Let $\boldsymbol{x}^\top$, $\boldsymbol{x}^{\mathsf{H}}$ denote the transpose and conjugate transpose of $\boldsymbol{x}$, respectively. Let $[n]$ denote the index set $\{1, 2, \cdots, n\}$. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, let x_j denote its jth element. Let $\boldsymbol{x}_{\mathcal{D}}$, $\mathcal{D} \subseteq [n]$ denote the length-$|\mathcal{D}|$ vector composed of the elements in the index set $\mathcal{D}$ of $\boldsymbol{x}$, and let $\boldsymbol{x}_{\setminus \mathcal{D}}$ denote the vector obtained by removing the elements of $\boldsymbol{x}$ in the index set $\mathcal{D}$. For example, $\boldsymbol{x}_{1:j}$ denotes the length-j vector composed of the first j entries of $\boldsymbol{x}$, i.e., the vector $[x_1, x_2, \cdots, x_j]^\top$, and $\boldsymbol{x}_{\setminus\{i\}}$ denotes the length-$(n-1)$ vector composed of all entries of $\boldsymbol{x}$ except the ith one, i.e. the vector $\boldsymbol{x}_{1:i-1,i+1:n}$. If an index $j \notin [n]$ for an n-dimensional vector, then the actual index is computed as in the modulo n sense. $\mathcal{S}_j$ denotes a circular shift by j positions, i.e., $[\mathcal{S}_j(\boldsymbol{x})]_k = x_{k-j}$ for $j, k \in [n]$. Let $\|\cdot\|_p$, $p \in [1, \infty]$ represent the ℓ_p norm of a vector, and $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$ denote the operator norm and the Frobenius norm of a matrix, respectively. Let $\sigma_i(\boldsymbol{A})$ be the ith largest eigenvalue of a matrix $\boldsymbol{A}$.Let $\odot$ denote the Hadamard product for two vector $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ of the same dimension. Let $\mathbf{I}$ denote an identity matrix, and $\boldsymbol{e}_i \in \mathbb{R}^n, i \in [n]$ be the ith standard basis vector. If $\boldsymbol{A} \preceq \boldsymbol{B}$, then $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite. Last, we use $c_1, c_2, C, \dots$ to denote universal constants whose values may change from line to line.

2 Main Results

To begin, we state a few key assumptions. In this paper, we assume that the sparse inputs are generated according to the well-known Bernoulli-Gaussian model, defined below.

Definition 1 (Bernoulli-Gaussian model [17]). *The inputs $\boldsymbol{x}_i$, $i = 1, \cdots, p$, are said to satisfy the Bernoulli-Gaussian model with parameter $\theta \in (0,1)$, i.e. $\boldsymbol{x}_i \sim_{iid} \mathrm{BG}(\theta)$, if $\boldsymbol{x}_i = \boldsymbol{\Omega}_i \odot \boldsymbol{z}_i$, where $\boldsymbol{\Omega}_i$ is an i.i.d. Bernoulli vector with parameter θ, and $\boldsymbol{z}_i$ is a random vector with i.i.d. random Gaussian variables drawn from $\mathcal{N}(0,1)$.*

Furthermore, the geometry of the loss function $f(\boldsymbol{h})$ turns out to be highly related to the condition number of the matrix $\mathcal{C}(\boldsymbol{g})$, which is defined below.

Definition 2 (Condition number). *Let κ be the condition number of $\mathcal{C}(\boldsymbol{g})$, i.e. $\kappa = \sigma_1(\mathcal{C}(\boldsymbol{g}))/\sigma_n(\mathcal{C}(\boldsymbol{g}))$.*

When $\mathcal{C}(\boldsymbol{g})$ is orthogonal, we have $\kappa = 1$. Let the discrete Fourier transform (DFT) of $\boldsymbol{g}$ be $\widehat{\boldsymbol{g}} = \boldsymbol{F}\boldsymbol{g}$, then κ is equivalent to the ratio of the largest and the smallest absolute values of $\widehat{\boldsymbol{g}}$, i.e. $\kappa := |\widehat{\boldsymbol{g}}|_{\max}/|\widehat{\boldsymbol{g}}|_{\min}$. Therefore, κ measures the flatness of the spectrum $\widehat{\boldsymbol{g}}$, which plays a similar role as the coherence introduced in early works of blind deconvolution with a single snapshot [18, 19]. In addition, since $\boldsymbol{g}_{\mathrm{inv}}$ can only be identified up to scaling and shift ambiguities, without loss of generality, we assume $\|\boldsymbol{g}_{\mathrm{inv}}\|_2 = 1$.

2.1 Geometry of the Empirical Loss

We start by describing the geometry of $f_o(\boldsymbol{h})$ when $\mathcal{C}(\boldsymbol{g})$ is an orthonormal matrix, where pre-conditioning is not needed. Without loss of generality, we can assume $\mathcal{C}(\boldsymbol{g}) = \mathbf{I}$,[1] which corresponds to the ground truth $\boldsymbol{g}_{\mathrm{inv}} = \boldsymbol{e}_1$ and $\boldsymbol{y}_i = \boldsymbol{x}_i$. Therefore, the loss function $f_o(\boldsymbol{h})$ in (6) can be equivalently reformulated as

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} \; f_o(\boldsymbol{h}) = \frac{1}{p}\sum_{i=1}^{p} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}) \quad \text{s.t.} \quad \|\boldsymbol{h}\|_2 = 1. \tag{9}$$

Our geometric theorem characterizes benign properties of the curvatures in the local neighborhood of $\{\pm \boldsymbol{e}_i\}_{i=1}^n$, shifted and sign-flipped copies of the ground truth. Inspired by [20, 21], we introduce $2n$ subsets,

$$\mathcal{S}_\xi^{(i\pm)} = \left\{ \boldsymbol{h} : h_i \gtrless 0, \frac{h_i^2}{\|\boldsymbol{h}_{\setminus\{i\}}\|_\infty^2} \geqslant 1 + \xi \right\}, \quad i \in [n], \tag{10}$$

where $\xi \in [0, \infty)$. Clearly, $\boldsymbol{e}_i \in \mathcal{S}_\xi^{(i+)}$ and $-\boldsymbol{e}_i \in \mathcal{S}_\xi^{(i-)}$, for all $i \in [n]$. The quantity ξ captures the size of the local neighborhood — the smaller ξ is, the larger the size of $\mathcal{S}_\xi^{(i\pm)}$.

Due to symmetry, we focus on describing the geometry of $f_o(\boldsymbol{h})$ in one of such subsets, say $\mathcal{S}_\xi^{(n+)}$. For convenience, we introduce a reparametrization trick [16]. Define $\boldsymbol{w} = \boldsymbol{h}_{1:n-1} \in \mathbb{B}^{n-1}$, corresponding to the first $(n-1)$ entries of $\boldsymbol{h}$, where $\mathbb{B}^{n-1} := \{\boldsymbol{w} \in \mathbb{R}^{n-1} : \|\boldsymbol{w}\|_2 \leq 1\}$ is the unit ball in $\mathbb{R}^{n-1}$. Given $\boldsymbol{w}$, the vector $\boldsymbol{h}$ can be written as

$$\boldsymbol{h}(\boldsymbol{w}) = \left(\boldsymbol{w}, \sqrt{1 - \|\boldsymbol{w}\|_2^2}\right), \quad \forall \boldsymbol{w} \in \mathbb{B}^{n-1}. \tag{11}$$

Therefore, $\boldsymbol{w} = \mathbf{0}$ is equivalent to $\boldsymbol{h}(\mathbf{0}) = \boldsymbol{e}_n$, which is the shifted ground truth within $\mathcal{S}_\xi^{(n+)}$. The loss function $f_o(\boldsymbol{h})$ can be rewritten with respect to $\boldsymbol{w}$ as

$$\phi_o(\boldsymbol{w}) = f_o(\boldsymbol{h}(\boldsymbol{w})) = \frac{1}{p}\sum_{i=1}^{p} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})). \tag{12}$$

In addition, a short calculation reveals that,[2]

$$\|\boldsymbol{w}\|_2^2 \leq \frac{n-1}{n+\xi} \quad \text{whenever} \quad \boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_\xi^{(n+)}. \tag{13}$$

[1] Denote $\widetilde{\boldsymbol{h}} = \mathcal{C}(\boldsymbol{g})\boldsymbol{h}$, we have $\|\widetilde{\boldsymbol{h}}\|_2 = \|\mathcal{C}(\boldsymbol{g})\boldsymbol{h}\|_2 = 1$ due to the orthonormality of $\mathcal{C}(\boldsymbol{g})$. Rewriting the loss function with respect to $\widetilde{\boldsymbol{h}}$ confirms this assertion. This does not change the geometry of the objective function that is of primary interest.

[2] When $\boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_\xi^{(n+)}$, we have $h_n^2 \geq (1+\xi)\|\boldsymbol{h}_{\setminus\{i\}}\|_\infty^2$, which leads to $1 = \|\boldsymbol{h}\|_2^2 \leq h_n^2 + (n-1)\|\boldsymbol{h}_{\setminus\{i\}}\|_\infty^2 \leq \left(1 + \frac{n-1}{1+\xi}\right) h_n^2 = \left(1 + \frac{n-1}{1+\xi}\right)(1 - \|\boldsymbol{w}\|_2^2)$.

The theorem below states the geometry of $\phi_o(\boldsymbol{w})$ in the neighborhood $\boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}$ for $\xi_0 \in (0,1)$. In particular, we split the region of interest into two subregions:

$$\mathcal{Q}_1 := \left\{ \boldsymbol{w} : \frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\|_2 \leq \sqrt{\frac{n-1}{n+\xi_0}} \right\}, \quad \mathcal{Q}_2 := \left\{ \boldsymbol{w} : \|\boldsymbol{w}\|_2 \leq \frac{\mu}{4\sqrt{2}} \right\}. \tag{14}$$

Theorem 2 (Geometry in the orthogonal case)**.** *Without loss of generality, suppose $\mathcal{C}(\boldsymbol{g}) = \mathbf{I}$. For any $\xi_0 \in (0,1)$, $\theta \in (0, \frac{1}{3})$, there exist constants $c_1, c_2, c_3, c_4, c_5, C$ such that when $\mu < c_1 \min\{\theta, \xi_0^{1/6} n^{-3/4}\}$ and*

$$p \geq \frac{Cn^4}{\theta^2 \xi_0^2} \log n \log \left(\frac{n^3 \log p}{\mu \theta \xi_0} \right), \tag{15}$$

the following holds with probability at least $1 - c_3 p^{-7} - \exp(-c_4 n)$ for $\boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}$:

$$\text{(large directional gradient)} \quad \frac{\boldsymbol{w}^\top \nabla \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} \geq c_2 \xi_0 \theta, \quad \text{if } \boldsymbol{w} \in \mathcal{Q}_1, \tag{16a}$$

$$\text{(strong convexity)} \quad \nabla^2 \phi_o(\boldsymbol{w}) \succeq \frac{c_2 n \theta}{\mu} \mathbf{I}, \quad \text{if } \boldsymbol{w} \in \mathcal{Q}_2. \tag{16b}$$

Furthermore, the function $\phi_o(\boldsymbol{w})$ has exactly one unique local minimizer $\boldsymbol{w}_o^\star$ near $\mathbf{0}$, such that

$$\|\boldsymbol{w}_o^\star - \mathbf{0}\|_2 \leq \frac{c_5 \mu}{\theta} \sqrt{\frac{\log^2 p}{p}}. \tag{17}$$

Theorem 2 has the following implications when $\boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}$, as long as the sample size p is sufficiently large and satisfies (15):

- The function $\phi_o(\boldsymbol{w})$ either has a large gradient when $\|\boldsymbol{w}\|_2$ is large (cf. (16a)), or is strongly convex when $\|\boldsymbol{w}\|_2$ is small (cf. (16b)), indicating the geometry is rather benign and suitable for optimization using first-order methods such as MGD;
- There are no spurious local minima, and the unique local optimizer is close to the ground truth according to (17) with an error decays at the rate $O\left(\frac{\mu}{\theta}\sqrt{\frac{\log^2 p}{p}}\right)$ as the sample size p increases.

Theorem 2 also suggests that a larger sample size is necessary to guarantee a benign geometry when the subset $\mathcal{S}_{\xi_0}^{(i\pm)}$ gets larger – with the decrease of ξ_0. By a simple union bound, we can ensure a similar geometry applies to all $2n$ subsets $\mathcal{S}_{\xi_0}^{(i\pm)}$ defined in (10).

Extension to the general case. To extend the geometry in Theorem 2 to the general case when $\mathcal{C}(\boldsymbol{g})$ is invertible, we adopt the trick in [14, 16] and introduce the pre-conditioning matrix $\boldsymbol{R}$:

$$\boldsymbol{R} = \left[\frac{1}{\theta n p} \sum_{i=1}^{p} \mathcal{C}(\boldsymbol{y}_i)^\top \mathcal{C}(\boldsymbol{y}_i) \right]^{-1/2}. \tag{18}$$

The main purpose of the pre-conditioning is to convert the loss function to one similar to the orthogonal case studied above. Recognizing that $\mathbb{E}\left[\frac{1}{\theta n p} \sum_{i=1}^{p} \mathcal{C}(\boldsymbol{x}_i)^\top \mathcal{C}(\boldsymbol{x}_i)\right] = \mathbf{I}$, we have $\boldsymbol{R} \approx \left(\mathcal{C}(\boldsymbol{g})^\top \mathcal{C}(\boldsymbol{g})\right)^{-1/2}$ asymptotically as p increases. Plugging $\mathcal{C}(\boldsymbol{y}_i) = \mathcal{C}(\boldsymbol{x}_i)\mathcal{C}(\boldsymbol{g})$ and $\boldsymbol{R} \approx \left(\mathcal{C}(\boldsymbol{g})^\top \mathcal{C}(\boldsymbol{g})\right)^{-1/2}$ into the loss function of (7), we have

$$f(\boldsymbol{h}) \approx \frac{1}{p} \sum_{i=1}^{p} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{U}\boldsymbol{h}), \tag{19}$$

where $\boldsymbol{U}$ is a circulant orthonormal matrix given by

$$\boldsymbol{U} := \mathcal{C}(\boldsymbol{g}) \left(\mathcal{C}(\boldsymbol{g})^\top \mathcal{C}(\boldsymbol{g})\right)^{-\frac{1}{2}}. \tag{20}$$

By the rotation invariance of the loss function over the sphere with respect to the orthonormal transform by $\boldsymbol{U}$, (19) is equivalent to the one studied in the orthogonal case, thus justifying our choice of the pre-conditioning matrix. Returning to the original loss function without approximating $\boldsymbol{R}$ by its population counterpart, we can repeat the same argument performed in (9) and rewrite a rotationed version of (7) as

$$\min_{\boldsymbol{h}\in\mathbb{R}^n} f(\boldsymbol{h}) = \frac{1}{p}\sum_{i=1}^{p} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i)\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{U}^\top \boldsymbol{h}) \quad \text{s.t} \quad \|\boldsymbol{h}\|_2 = 1, \tag{21}$$

where the shifted and sign-flipped ground truth has been rotated to almost $\{\pm \boldsymbol{e}_i\}_{i=1}^n$, which is the same as the orthogonal case. The theorem below suggests that under the same reparameterization $\boldsymbol{h} = \boldsymbol{h}(\boldsymbol{w})$ in (11), a similar geometry as Theorem 2 can be guaranteed for $\phi(\boldsymbol{w}) = f(\boldsymbol{h}(\boldsymbol{w}))$.

Theorem 3 (Geometry in the general case)**.** *Suppose $\mathcal{C}(\boldsymbol{g})$ is invertible with condition number κ. For any $\xi_0 \in (0,1)$, $\theta \in (0,\frac{1}{3})$, there exist constants c_1, c_2, c_3, c_4, C such that when $\mu < c_1 \min\{\theta, \xi_0^{1/6} n^{-3/4}\}$ and*

$$p \geq C\frac{\kappa^8 n^3 \log^4 p \log^2 n}{\theta^4 \mu^2 \xi_0^2}, \tag{22}$$

the geometry (16) holds for $\phi(\boldsymbol{w})$ with probability at least $1 - c_3 p^{-7} - \exp(-c_4 n)$ for $\boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}$. In addition, the function $\phi(\boldsymbol{w})$ has exactly one unique local minimizer $\boldsymbol{w}^\star$ near $\boldsymbol{0}$, such that

$$\|\boldsymbol{w}^\star - \boldsymbol{0}\|_2 \leq \frac{c_2\kappa^4}{\theta^2}\sqrt{\frac{n\log^3 p \log^2 n}{p}}.$$

Theorem 3 demonstrates that a benign geometry similar to that in Theorem 2 can be guaranteed for the general case, as long as a proper pre-conditioning is applied, and the sample size is sufficiently large. In particular, the sample size in (22) increases with the increase of the condition number of $\mathcal{C}(\boldsymbol{g})$.

2.2 Convergence Guarantees of MGD

Owing to the benign geometry in the subsets of interest $\left\{\mathcal{S}_{\xi_0}^{(i\pm)}, i \in [n]\right\}$, a simple MGD algorithm is proposed to optimize (21), by updating

$$\boldsymbol{h}^{(k+1)} = \frac{\boldsymbol{h}^{(k)} - \eta\partial f(\boldsymbol{h}^{(k)})}{\left\|\boldsymbol{h}^{(k)} - \eta\partial f\left(\boldsymbol{h}^{(k)}\right)\right\|_2} \tag{23}$$

for $k = 1, \ldots, T-1$, where $\partial f(\boldsymbol{h}) = (\mathbf{I} - \boldsymbol{h}\boldsymbol{h}^\top)\nabla f(\boldsymbol{h})$ is the Riemannian manifold gradient with respect to $\boldsymbol{h}$, $\nabla f(\boldsymbol{h})$ is the Euclidean gradient of $f(\boldsymbol{h})$, and η is the step size. The next theorem demonstrates that with an initialization in one of the $2n$ subsets $\left\{\mathcal{S}_{\xi_0}^{(i\pm)}, i \in [n]\right\}$, the proposed MGD algorithm, with a proper step size, will recover $\pm\boldsymbol{e}_i$ in that region in a polynomial time.

Theorem 4. *Let $0 < \epsilon < 1$ and instate the assumptions of Theorem 3. If the initialization satisfies $\boldsymbol{h}^{(0)} \in \mathcal{S}_{\xi_0}^{(i\pm)}$ for any $i \in [n]$, then with a step size $\eta \leq \frac{c\mu\xi_0\theta}{n^2\sqrt{\log(np)}}$ for some sufficiently small constant c, the iterates $\boldsymbol{h}^{(k)}, k = 1, 2, \cdots$ stay in $\mathcal{S}_{\xi_0}^{(i\pm)}$ and achieve $\|\boldsymbol{h}^{(T)} \mp \boldsymbol{e}_i\|_2 \lesssim \frac{\kappa^4}{\theta^2}\sqrt{\frac{n\log^3 p\log^2 n}{p}} + \epsilon$ in*

$$T \lesssim \frac{n}{\mu\eta\xi_0\theta} + \frac{\mu}{n\theta\eta}\log\left(\frac{\mu}{\epsilon}\right)$$

iterations.

With Theorem 4 in place, one still needs to address how to find an initialization that satisfies $\boldsymbol{h}^{(0)} \in \mathcal{S}_{\xi_0}^{(i\pm)}$. Fortunately, setting $\xi_0 = 1/(4\log n)$ allows a sufficiently large basin of attraction, such that a random initialization can land into it with a constant probability. A few random initializations guarantee that the MGD algorithm will succeed with high probability. This is made precise in the following corollary.

Algorithm 1: Manifold Gradient Descent for MSBD

Input: Observation $\{\boldsymbol{y}_i\}_{i=1}^p$, step size η, initialization $\boldsymbol{h}^{(0)}$ on the sphere, the loss function $f(\boldsymbol{h})$ in (7);

for $k = 0$ **to** $T-1$ **do**

$$\boldsymbol{h}^{(k+1)} = \frac{\boldsymbol{h}^{(k)} - \eta \partial f(\boldsymbol{h}^{(k)})}{\left\|\boldsymbol{h}^{(k)} - \eta \partial f\left(\boldsymbol{h}^{(k)}\right)\right\|_2};$$

Output: Return $\widehat{\boldsymbol{g}}_{\text{inv}} = \boldsymbol{R}\boldsymbol{h}^{(T)}$, where $\boldsymbol{R}$ is given in (18).

Putting everything together, Alg. 1 summarizes the proposed MGD algorithm for the original loss function in (7), where the pre-conditioning matrix is applied back at the end of the iterations to produce the final estimate $\widehat{\boldsymbol{g}}_{\text{inv}}$ of $\boldsymbol{g}_{\text{inv}}$. To measure the success of recovery, we use the following distance metric that takes into account the ambiguities:

$$\text{dist}(\widehat{\boldsymbol{g}}_{\text{inv}}, \boldsymbol{g}_{\text{inv}}) = \min_{j\in[n]} \|\boldsymbol{g}_{\text{inv}} \pm \mathcal{S}_j(\widehat{\boldsymbol{g}}_{\text{inv}})\|_2. \tag{24}$$

We have the following corollary.

Corollary 1 (Putting everything together)**.** *Suppose $\mathcal{C}(\boldsymbol{g})$ is invertible with condition number κ. For $\theta \in (0, \frac{1}{3})$, there exists some constant c_1 such that when $\mu < c_1 \min\{\theta, (\log n)^{-1/6} n^{-3/4}\}$ and the sample complexity satisfies*

$$p \gtrsim \frac{\kappa^8 n^3 \log^3 n \log^4 p}{\theta^4 \mu^2}, \tag{25}$$

with $O(\log n)$ random initializations selected uniformly over the sphere, the MGD algorithm in Alg. 1 with a proper step size is guaranteed to obtain a vector $\widehat{\boldsymbol{g}}_{\text{inv}}$ that satisfies

$$\text{dist}(\widehat{\boldsymbol{g}}_{\text{inv}}, \boldsymbol{g}_{\text{inv}}) \lesssim \frac{\kappa^4}{\theta^2 \sigma_n(\mathcal{C}(\boldsymbol{g}))} \sqrt{\frac{n \log^3 p \log^2 n}{p}} + \frac{\epsilon}{\sigma_n(\mathcal{C}(\boldsymbol{g}))}$$

for any $0 < \epsilon < 1$, in $O(poly(n))$ iterations.

Corollary 1 provides theoretical footings to the success of MGD for solving the highly nonconvex MSBD problem. In particular, consider the interesting regime when $\theta = O(1)$ and $\kappa = O(1)$, it is sufficient to set $\mu \lesssim (\log n)^{-1/6} n^{-3/4}$, which leads to a sample size $p = O(n^{4.5})$ up to logarithmic factors.

2.3 Comparisons to Existing Approaches

The identifiability of the MSBD problem is established in [15] that says under the Bernoulli-Gaussian model (cf. Definition 1) on the sparse coefficients, the filter is identifiable with high probability, provided that $\boldsymbol{g}$ is invertible, $\theta \in (1/n, 1/4)$ and $p \gtrsim n \log n$. Wang and Chi proposed a linear program in [22] that succeeds when $p \gtrsim n \log^4 n$. However, the success of the linear program therein imposes stringent requirements on the conditioning number of the filter $\boldsymbol{g}$ and the sparsity level θ.

Our approach is most related to the subsequent work of Li and Bresler [14], which runs perturbed MGD with a random initialization, over a spherically constrained loss function based on ℓ_4 norm maximization. Li and Bresler showed that, when the sample complexity is large enough, the landscape of the loss function does not possess spurious local maxima, and all saddle points admit directions that strictly increase the loss function. However, their sample complexity is significantly worse. Specifically, to reach a similar accuracy as ours, [14] requires $O(n^9)$ samples, while we only require $O(n^{4.5})$ samples ignoring logarithmic factors, leading to an order-of-magnitude improvement. One key observation is that the large sample complexity required by [14] is partially due to bounding the global geometry everywhere over the sphere, through uniform concentration of the gradient and the Hessian of the empirical loss function around their population counterparts, which is sufficient but in fact not necessary to ensure the algorithmic success of MGD. Indeed, motivated by [20, 21], to optimize the sample complexity, we only require the uniform concentration of directional gradient over a large region near the global minimizer, which can be guaranteed at a significantly reduced sample complexity. In addition, this region is large enough so that with a logarithmic number of random initializations we are guaranteed to land into this region with high probability and recover the signal of interest via *vanilla* MGD. It is worth pointing out that since we focused on a region without saddle points, no perturbation is needed to ensure the success of MGD, which is another salient difference from [14]. As will be seen in Section 4, the proposed loss function not only theoretically, but also empirically, outperforms the ℓ_4 norm used in [14].

At the time of finishing this paper, we became aware of another concurrent work [23], which optimized a different smooth surrogate of the ℓ_1 norm over the sphere for the same problem. Their work [23] requires a sample complexity on the order of $O(n^5)$, which is slightly worse than ours (i.e. $O(n^{4.5})$), to guarantee the benign geometry in a similar region near the global optimizer. In addition, a refinement procedure is proposed in [23] to allow exact recovery of the filter. Their path to a better sample complexity than [14] is similar to ours as described above. We expect that their method behaves similar to ours in practice.

3 Overview of the Analysis

In this section, we outline the proof of the main results, while leaving the details to the appendix. We first deal with the simpler case when $\mathcal{C}(\boldsymbol{g})$ is an orthonormal matrix employing the objective function $\phi_o(\boldsymbol{w})$ (i.e, $f_o(\boldsymbol{h})$) without pre-conditioning in Section 3.1, and then extend the analysis to the general case where the objective function $\phi(\boldsymbol{w})$ (i.e, $f(\boldsymbol{h})$) is pre-conditioned in Section 3.2. Finally, we discuss the convergence guarantee of MGD in Section 3.3.

3.1 Proof Outline of Theorem 2

The proof of Theorem 2 is divided into several steps.

1. First, we characterize the landscape of the population loss function $\mathbb{E}[\phi_o(\boldsymbol{w})]$;
2. Second, we prove the pointwise concentration of the directional gradient and the Hessian of the empirical loss $\phi_o(\boldsymbol{w})$ around those of the population one $\mathbb{E}[\phi_o(\boldsymbol{w})]$ in the region of interest;
3. Third, we extend such concentrations to the uniform sense, thus the benign geometric properties of $\mathbb{E}[\phi_o(\boldsymbol{w})]$ carry over to the empirical version $\phi_o(\boldsymbol{w})$.

To begin, the lemma below describes the geometry of $\mathbb{E}[\phi_o(\boldsymbol{w})]$, whose proof is given in Appendix B.1.

Lemma 1 (Geometry of the population loss in the orthogonal case)**.** *Without loss of generality, suppose* $\mathcal{C}(\boldsymbol{g}) = \mathbf{I}$. *For any* $\xi_0 \in (0,1)$, $\theta \in (0, \frac{1}{3})$, *there exists some constant* c_1 *such that when* $\mu < c_1 \min\{\theta, \xi_0^{1/6} n^{-3/4}\}$, *we have for* $\boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}$*:*

$$\textit{(large directional gradient)} \quad \frac{\boldsymbol{w}^\top \nabla \mathbb{E}\phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} \geq \frac{\xi_0 \theta}{480\sqrt{10\pi}}, \quad \boldsymbol{w} \in \mathcal{Q}_1, \tag{26a}$$

$$\textit{(strong convexity)} \quad \nabla^2 \mathbb{E}\phi_o(\boldsymbol{w}) \succeq \frac{n\theta}{5\sqrt{2\pi}\mu}\mathbf{I}, \quad \boldsymbol{w} \in \mathcal{Q}_2. \tag{26b}$$

To extend the benign geometry to the empirical loss with a finite sample size p, we first need to prove the pointwise concentration of these quantities around their expectations for a fixed $\boldsymbol{w}$, using the Bernstein's inequality. The next two propositions demonstrate the pointwise concentration results, whose proofs are provided in Appendix B.2 and B.3.

Proposition 1. *For any* $\boldsymbol{w}$ *satisfies* $\|\boldsymbol{w}\|_2 \leq \sqrt{\frac{n-1}{n}}$, *there exist some universal constants* C_1 *and* C_2 *such that for any* $t > 0$*:*

$$\mathbb{P}\left[\left|\frac{\boldsymbol{w}^\top \nabla \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} - \frac{\boldsymbol{w}^\top \nabla \mathbb{E}\phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2}\right| \geq t\right] \leq 2\exp\left(\frac{-pt^2}{C_1 n^3 \log n + C_2 t\sqrt{n^3 \log n}}\right).$$

Proposition 2. *For any $\boldsymbol{w}$ satisfies $\|\boldsymbol{w}\|_2 \le 1/2$, there exist some universal constants C_1 and C_2 such that for any $t > 0$,*

$$\mathbb{P}\left[\left\|\nabla^2 \phi_o(\boldsymbol{w}) - \nabla^2 \mathbb{E}\phi_o(\boldsymbol{w})\right\| \ge t\right] \le 4n \exp\left(\frac{-p\mu^2 t^2}{C_1 n^2 \log^2 n + C_2 \mu n t \log n}\right).$$

The concentration of the Hessian and directional gradient between the empirical and population objective functions at a fixed point suggests that the empirical objective function may inherit the benign geometry of the population one outlined in Lemma 1. However, one needs to carefully extend the pointwise concentrations in Propositions 1 and 2 through a covering argument, which requires bounding the Lipschitz constants of the Hessian and directional gradients. The rest of the proof of Theorem 2 is provided in Appendix B.4.

3.2 Proof Outline of Theorem 3

To extend the benign geometry to the general case, we show that through pre-conditioning, the landscape of $\phi(\boldsymbol{w})$ is not too far from that of $\phi_o(\boldsymbol{w})$. Recall that the pre-conditioned loss function (21) is

$$\begin{aligned}
\phi(\boldsymbol{w}) &= \frac{1}{p}\sum_{i=1}^{p} \psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{U}^\top \boldsymbol{h}(\boldsymbol{w})\right) \\
&= \frac{1}{p}\sum_{i=1}^{p} \psi_\mu\Big(\mathcal{C}(\boldsymbol{x}_i)\big[\mathbf{I} + \underbrace{(\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{U}^{-1} - \mathbf{I})}_{\boldsymbol{\Delta}}\big]\boldsymbol{h}(\boldsymbol{w})\Big),
\end{aligned} \tag{27}$$

where $\boldsymbol{\Delta} := \mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{U}^{-1} - \mathbf{I} = (\boldsymbol{U}' - \boldsymbol{U})\boldsymbol{U}^{-1}$, with $\boldsymbol{U}' = \mathcal{C}(\boldsymbol{g})\boldsymbol{R}$ and $\boldsymbol{U}$ in (20). As was discussed earlier, as $\boldsymbol{R}$ converges to $\left[\mathcal{C}(\boldsymbol{g})^\top \mathcal{C}(\boldsymbol{g})\right]^{-1/2}$ when p increases, it is expected that $\boldsymbol{U}'$ converges to $\boldsymbol{U}$. Therefore, by bounding the size of $\boldsymbol{\Delta}$, we can control the deviation between $\phi_o(\boldsymbol{w})$ and $\phi(\boldsymbol{w})$. To this end, the rest of the proof is divided into the following two steps.

First, we show that the spectral norm of $\boldsymbol{\Delta}$ is bounded when the sample size is sufficiently large in Lemma 2, whose proof is given in Appendix C.1.

Lemma 2 (Spectral norm of $\boldsymbol{\Delta}$)**.** *There exist some constants C_1, c_f, such that when $p \ge \frac{C_1 \kappa^4 n \log^2 n \log p}{\theta^2}$, with probability at least $1 - 2np^{-8}$,*

$$\|\boldsymbol{\Delta}\| \le c_f \kappa^4 \sqrt{\frac{\log^2 n \log p}{\theta^2 p}}. \tag{28}$$

Second, we show that the deviation between the directional gradient and the Hessian of $\phi(\boldsymbol{w})$ and $\phi_o(\boldsymbol{w})$ can be bounded by the spectral norm of $\boldsymbol{\Delta}$, as shown in Lemma 3. The proof can be found in Appendix C.2.

Lemma 3 (Deviation between $\phi_o(\boldsymbol{w})$ and $\phi(\boldsymbol{w})$)**.** *There exist some constants c_g, c_h, C_1, such that when $p \ge \frac{C_1 \kappa^8 n \log^2 n \log p}{\theta^2}$, with probability at least $1 - 2p^{-8}$, we have*

$$\|\nabla \phi_o(\boldsymbol{w}) - \nabla \phi(\boldsymbol{w})\|_2 \le c_g \frac{n^{3/2}\log(np)}{\mu}\|\boldsymbol{\Delta}\|, \quad \boldsymbol{w} \in \mathcal{Q}_1, \tag{29a}$$

$$\left\|\nabla^2 \phi_o(\boldsymbol{w}) - \nabla^2 \phi(\boldsymbol{w})\right\| \le c_h \frac{n^{5/2}\log^{3/2}(np)}{\mu^2}\|\boldsymbol{\Delta}\|, \quad \boldsymbol{w} \in \mathcal{Q}_2. \tag{29b}$$

To complete the proof of Theorem 3, we need to show that the perturbations of the Hessian and the gradient between $\phi_o(\boldsymbol{w})$ and $\phi(\boldsymbol{w})$ are sufficiently small, which hold as long as the sample size is sufficiently large, in view of Lemma 2. Consequently, we can propagate the benign geometry of $\phi_o(\boldsymbol{w})$ in Theorem 2 to $\phi(\boldsymbol{w})$. The complete proof is provided in Appendix C.3.

3.3 Proof Outline of Theorem 4

To capitalize on the benign geometry established in Theorem 3, one of the key arguments is to ensure that the iterates of MGD stay in the $2n$ subsets $\left\{\mathcal{S}_{\xi_0}^{(i\pm)}, i \in [n]\right\}$ implicitly. This requires bounding properties of the directional gradient of $f(\boldsymbol{h})$ in (21), supplied in the following lemma whose proof can be found in Appendix D.1.

Lemma 4 (Uniform concentration of the directional gradient)**.** *Instate the assumptions of Theorem 3. There exist some constants c_a, c_b, C_1, such that with probability at least $1 - 3(np)^{-8} - 2\exp(-c_a n)$,*

$$\partial f(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n} \right) \geq \frac{c_b \xi_0 \theta}{2}, \tag{30}$$

for $\boldsymbol{h} \in \mathcal{H}_k = \left\{ \boldsymbol{h} : \boldsymbol{h} \in \mathcal{S}_{\xi_0}^{(n+)}, h_k \neq 0, h_n^2/h_k^2 < 4 \right\}$, and

$$\|\partial f(\boldsymbol{h})\|_2 \leq \|\nabla f(\boldsymbol{h})\|_2 \leq C_1 n \sqrt{\log(np)} \tag{31}$$

for $\boldsymbol{h} \in \mathbb{S}^{n-1}$.

The following lemma, proved in Appendix D.2, then shows that the iterates of MGD will always stay in one of the subsets $\left\{ \mathcal{S}_{\xi_0}^{(i\pm)}, i \in [n] \right\}$ that it initializes in, as long as the sample complexity p is large enough and the step size is properly chosen.

Lemma 5 (Implicitly staying in the subsets)**.** *Instate the assumptions of Theorem 3. For the MGD algorithm in Alg. 1, if the initialization satisfies that $\boldsymbol{h}^{(0)} \in \mathcal{S}_{\xi_0}^{(i\pm)}$ for any $i \in [n]$, and the step size satisfies $\eta \leq \frac{c}{n^{3/2}\sqrt{\log(np)}}$ for some small enough constant c, then the iterates $\boldsymbol{h}^{(k)}$, $k = 1, 2, \cdots$ will stay in $\mathcal{S}_{\xi_0}^{(i\pm)}$.*

The proof of Theorem 4 then follows by analyzing the convergence in two stages, corresponding to when the iterates lie in the region with large directional gradients, and the region with strong convexity, respectively. The details are given in Appendix D.3.

Till this point, the only left ingredient is to make sure a valid initialization can be obtained efficiently. By setting ξ_0 sufficiently small, it is known from the following lemma [21, Lemma 3] that the union of $\left\{ \mathcal{S}_{\xi_0}^{(i\pm)}, i \in [n] \right\}$ is large enough to ensure a random initialization will land in it with a constant probability.

Lemma 6 ([21, Lemma 3])**.** *When $\xi_0 = \frac{1}{4\log n}$, an initialization selected uniformly at random on the sphere lies in one of these $2n$ subsets $\left\{ \mathcal{S}_{\xi_0}^{(i\pm)}, i \in [n] \right\}$ with probability at least $1/2$.*

Finally, combining Lemma 6 and Theorem 4, by setting $\xi_0 = 1/(4\log n)$, we can guarantee to recover $\boldsymbol{g}_{\mathrm{inv}}$ accurately up to global ambiguity with high probability, as long as Alg. 1 is initialized uniformly at random over the sphere with $O(\log n)$ times. This leads precisely to Corollary 1.

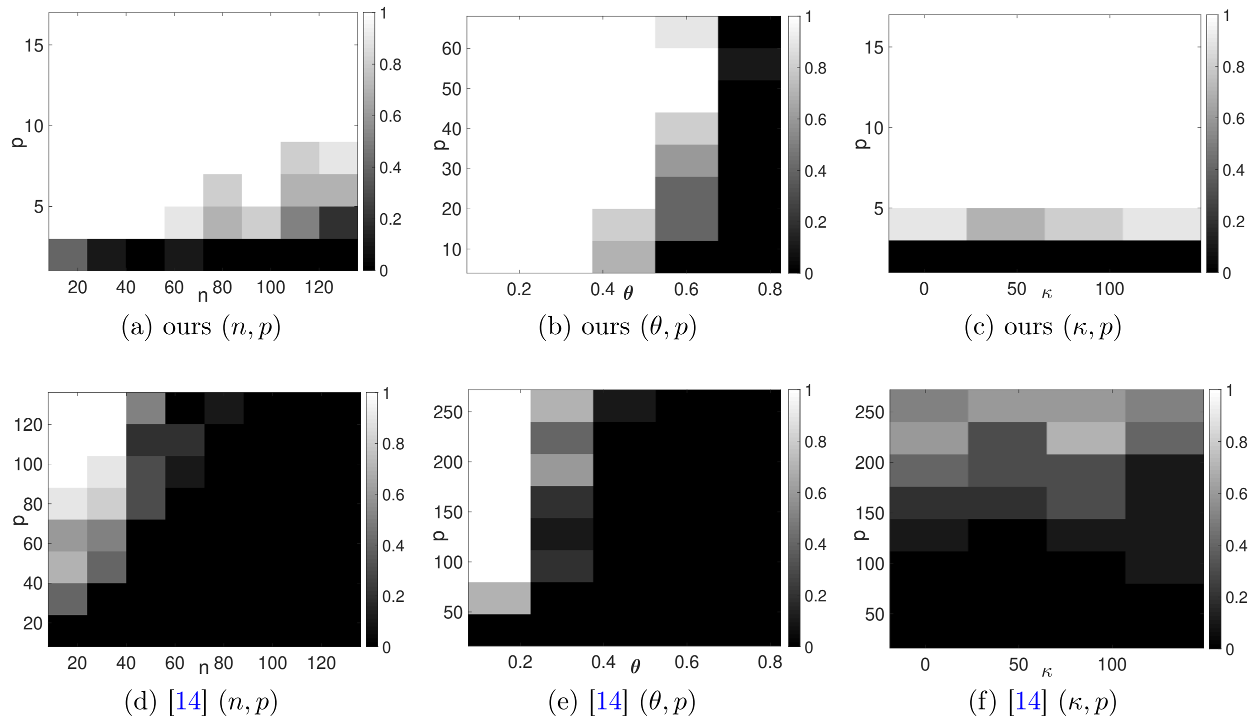
4 Numerical Experiments

In this section, we examine the performance of the proposed approach with comparison to [14], which is also based on MGD but using a different loss function $L(\boldsymbol{h}) = -\frac{1}{4p}\sum_{i=1}^{p} \|\mathcal{C}(\boldsymbol{y}_i)\boldsymbol{R}\boldsymbol{h}\|_4^4$ over the sphere, on both synthetic and real data.

4.1 MSBD with Synthetic Data

We first compare the success rates of the proposed approach and the approach in [14], following a similar simulation setup as in [14]. In each experiment, the sparse inputs are generated following $\mathrm{BG}(\theta)$, and $\mathcal{C}(\boldsymbol{g})$ with specific κ is synthesized by generating the DFT $\widehat{\boldsymbol{g}}$ of $\boldsymbol{g}$ which is random with the following rules: 1) the DFT $\widehat{\boldsymbol{g}}$ is symmetric to ensure that $\boldsymbol{g}$ is real, i.e., $\widehat{\boldsymbol{g}}_j = \widehat{\boldsymbol{g}}_{n+2-j}^*$, where $*$ denotes the conjugate operation; 2) the gains of $\widehat{\boldsymbol{g}}$ follow a uniform distribution on $[1, \kappa]$, and the phases of $\widehat{\boldsymbol{g}}$ follow a uniform distribution on $[0, 2\pi)$.

In all experiments, we run MGD (cf. Alg. 1) for no more than $T = 200$ iterations with a fixed step size of $\eta = 0.1$ and apply backtracking line search for both methods for computational efficiency. For our formulation,



(a) ours (n, p) (b) ours (θ, p) (c) ours (κ, p)

(d) [14] (n, p) (e) [14] (θ, p) (f) [14] (κ, p)

Figure 3: Success rates of the proposed approach (first row) and the approach in [14] (second row) under various parameter settings.

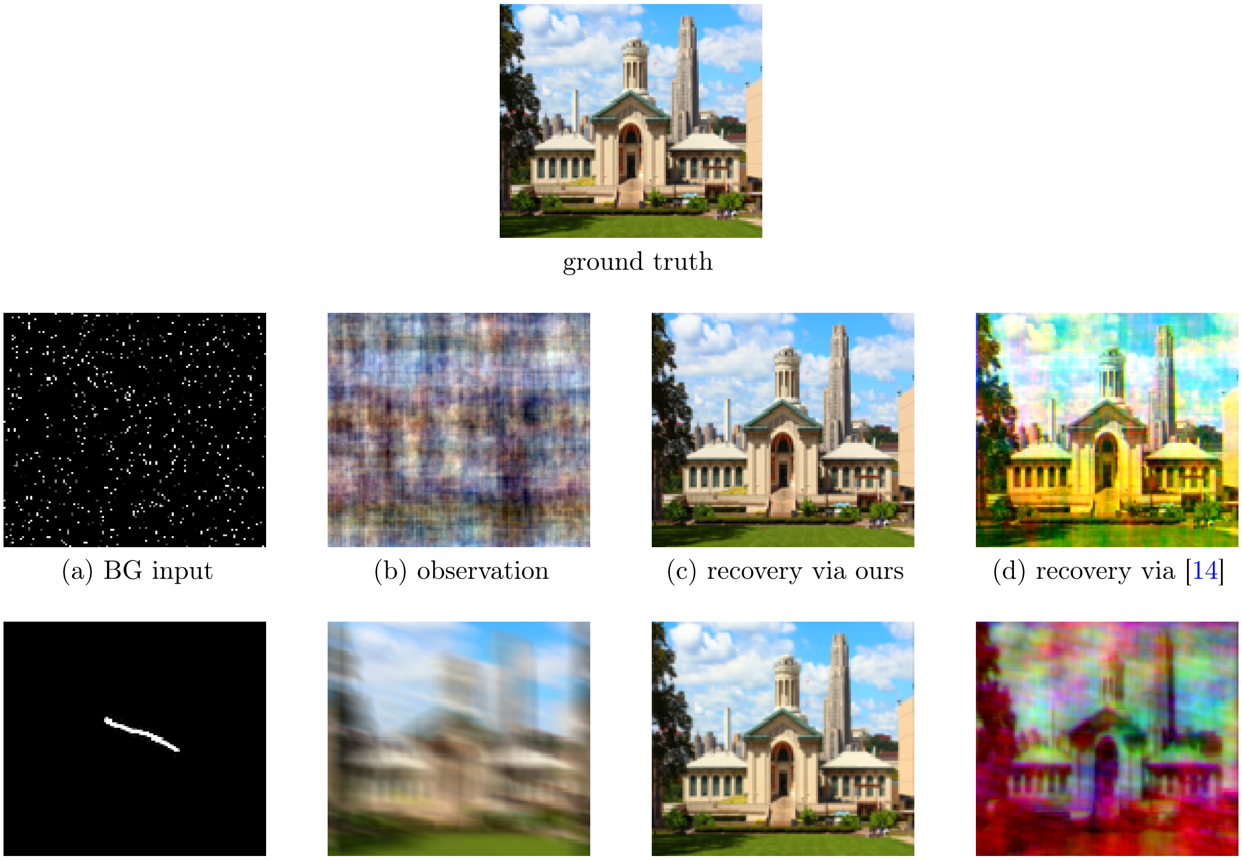
we set $\mu = \min\left(10n^{-5/4}, 0.05\right)$. For each parameter setting, we conduct 10 Monte Carlo simulations to compute the success rate. Recall that the desired estimate $\widehat{\boldsymbol{g}}_{\text{inv}}$ is a signed shifted version of $\boldsymbol{g}_{\text{inv}}$, since $\mathcal{C}(\boldsymbol{g})\boldsymbol{g}_{\text{inv}} = \pm\boldsymbol{e}_j$ $(j \in [n])$. Therefore, to evaluate the accuracy of the output $\widehat{\boldsymbol{g}}_{\text{inv}}$, we compute $\mathcal{C}(\boldsymbol{g})\widehat{\boldsymbol{g}}_{\text{inv}}$ using the ground truth $\boldsymbol{g}$, and declare that the recovery is successful if $\|\mathcal{C}(\boldsymbol{g})\widehat{\boldsymbol{g}}_{\text{inv}}\|_\infty / \|\mathcal{C}(\boldsymbol{g})\widehat{\boldsymbol{g}}_{\text{inv}}\|_2 > 0.99$.

Fig. 3 (a) and (d) show the success rates of the proposed approach and the approach in [14] with respect to n and p, where $\theta = 0.3$ and $\kappa = 8$ are fixed. It can be seen that the proposed approach succeeds at a much smaller sample size, even when p is smaller than n. This indicates room for improvements of our theory. Fig. 3 (b) and (e) shows the success rates of the two approaches with respect to θ and p, where $n = 64$ and $\kappa = 8$ are fixed. The proposed approach continues to work well even at a relatively high value of θ up to around 0.5. Finally, Fig. 3 (c) and (f) shows the success rate of the two approaches with respect to κ and p, where $n = 64$ and $\theta = 0.3$ are fixed. Again, the performance of the proposed approach is quite insensitive to the condition number κ as long as the sample size p is large enough. On the other end, the approach in [14] performs significantly worse than the proposed approach under the examined parameter settings.

4.2 Image Deconvolution and Deblurring

To further evaluate our method, we performance the task of blind image reconstruction and deblurring, and compare with [14]. Firstly, suppose multiple circulant convolutions $\{\boldsymbol{y}_i\}_{i=1}^p$ (illustrated in Fig. 4 (b)) of an unknown 2D image (illustrated in the ground truth figure in Fig. 4, the Hamerschlag Hall on the campus of CMU) and multiple Bernoulli-Gaussian (BG) sparse inputs $\{\boldsymbol{x}_i \sim_{iid} \mathrm{BG}(\theta)\}_{i=1}^p$ (illustrated in Fig. 4 (a)) are observed. Here, the size of the observations is $n = 128 \times 128$, $\theta = 0.1$, and the number of observations $p = 1000$, which is significantly smaller than n.

We apply the proposed reconstruction method to each channel of the image, i.e. R, G, B, respectively using the corresponding channel of the observations $\{\boldsymbol{y}_i\}_{i=1}^p$, and obtain the final recovery by summing up the recovered channels. For each channel, the recovered image is computed as $\widehat{\boldsymbol{g}} = \mathcal{F}^{-1}\left[\mathcal{F}\left(\boldsymbol{R}\widehat{\boldsymbol{h}}\right)^{\odot -1}\right]$, where $\widehat{\boldsymbol{h}}$ denotes the output of the algorithm, $\mathcal{F}$ is the 2D DFT operator, and $\boldsymbol{x}^{\odot -1}$ is the entry-wise inverse of a



ground truth

(a) BG input (b) observation (c) recovery via ours (d) recovery via [14]

(e) blurring input (f) observation (g) recovery via ours (h) recovery via [14]

Figure 4: Multi-channel sparse blind image deconvolution and deblurring. Top row: the ground truth image. Middle row: (a) the sparse input generated from the BG model; (b) the observation; (c) and (d): the recovery via our method and [14]. Bottom row: (e) the sparse input modeling motion blur; (f) the observation; (g) and (h): the recovery via our method and [14].

vector $\boldsymbol{x}$. Fig. 4 (c) and (d) show the final recovered images by our method and [14] (after aligning the shift and sign) respectively. It implies that the proposed approach obtains much better recovery than that in [14] when the sparse inputs $\{\boldsymbol{x}_i\}_{i=1}^p$ are with constant sparsity level θ.

We next consider a more realistic setting and examine the performance of the proposed algorithm when the sparse coefficients do not obey the Bernoulli-Gaussian model. Using the same 2D image, we now generate multiple circulant convolutions $\{\boldsymbol{y}_i\}_{i=1}^p$ (illustrated in Fig. 4 (f)) using realistically-generated motion blur kernels[3] (illustrated in Fig. 4 (e)). Fig. 4 (g) and (h) show the final recovered images by our method and [14] (after aligning the shift and sign) respectively. It can be seen that the proposed approach still obtains a robust recovery and removes the blurring effectively, while the recovery using [14] further degenerates possibly due to the model mismatch.

5 Further Related Work

In this section, we discuss further related literature, emphasizing on algorithms with provable guarantees.

Provable blind deconvolution. The problem of blind deconvolution with a single snapshot (or equivalently, channel) has been studied recently under different geometric priors such as sparsity and subspace assumptions

[3]The nonlinear blur kernels are randomly produced using the tool in https://github.com/LeviBorodenko/motionblur.

on both the filter and the input, using both convex and nonconvex optimization formulations [18, 24, 19, 25, 26, 27, 28, 29, 30, 31, 32]. With the presence of multiple channels, one expects to identify the filter with fewer prior assumptions. Algorithms for multi-channel blind deconvolution include sparse spectral methods [33], linear least squares [34], and nonconvex regularization [35]. A different model called "sparse-and-short" deconvolution is studied in [36, 37].

Provable dictionary learning. Learning a sparsifying invertible transform from data has been extensively studied, e.g. in [38, 17, 20, 21, 39, 40, 41]. In addition, provable algorithms for learning overcomplete dictionaries are also proposed in [42, 43, 44, 45]. Our problem can be regarded as learning a convolutional invertible transform, where the proposed algorithm is inspired by the approach in [20] that characterizes a local region large enough for the success of gradient descent with random initializations. However, the approach in [20] is only applicable to an orthogonal dictionary, while we deal with a general invertible convolutional kernel. Compared to sample complexities required in learning complete dictionaries [16], our result demonstrates the benefit of exploiting convolutional structures in further reducing the sample complexity.

Provable nonconvex statistical estimation. Our work belongs to the recent line of activities of designing provable nonconvex procedures for high-dimensional statistical estimation, see [46, 47, 48] for recent overviews. Our approach interpolates between two popular approaches, namely, global analyses of optimization landscape (e.g. [38, 16, 49, 50, 51, 52, 53, 54, 14]) that are independent of algorithmic choices, and local analyses with careful initializations and local updates (e.g. [55, 56, 57, 58, 59, 60, 26, 61, 62, 63, 64]).

6 Discussions

This paper proposes a novel nonconvex approach for multi-channel sparse blind deconvolution based on manifold gradient descent with random initializations. Under a Bernoulli-Gaussian model for sparse inputs, we demonstrate that the proposed approach succeeds as long as the sample complexity satisfies $p = O(n^{4.5}\text{polylog}p)$, a result significantly improving prior art in [14]. We conclude the paper by some discussions on future directions.

- *Improve sample complexity.* Our numerical experiments indicate that there is still room to further improve the sample complexity of the proposed algorithm, which may require a more careful analysis of the trajectory of the gradient descent iterates, as done in [65].

- *Efficient exploitation of negative curvature.* We remark that it is possible to characterize the global geometry over the sphere, where the remaining region contains saddle points with negative curvatures. However, a direct analysis leads to an increase of sample complexity which is undesirable and therefore not pursued in this paper. On the other end, it seems random initialization *without restarts* also works well in practice, which warrants further investigation.

- *Super-resolution blind deconvolution.* The model studied in this paper assumes the same temporal resolution of the input and the output, while in practice the sparse activations of the input can occur at a much higher resolution. This lead to the consideration of a refined model, where the observation is given as $\boldsymbol{y} = \boldsymbol{F}_{n\times n}^{\mathsf{H}}\text{diag}(\widehat{\boldsymbol{g}})\boldsymbol{F}_{n\times D}\boldsymbol{x}$, where $\boldsymbol{F}_{n\times D}$ is the oversampled DFT matrix of size $n \times D$, $D \geq n$. The approach taken in this paper cannot be applied anymore, and new formulations are needed to address this problem, see [66] for a related problem.

- *Convolutional dictionary learning.* Our work can be regarded as a first step towards developing sample-efficient algorithms for convolutional dictionary learning [67] with performance guarantees. An interesting model for future investigation is when multiple filters are present, and the observation is modeled as $\boldsymbol{y} = \sum_{\ell=1}^{L} \mathcal{C}(\boldsymbol{g}_\ell)\boldsymbol{x}_\ell$, with L the number of filters. The goal is thus to simultaneously learn multiple filters $\{\boldsymbol{g}_\ell\}_{\ell=1}^{L}$ from a number of observations in the form of $\boldsymbol{y}$. See [68] for some recent developments.

Acknowledgment

The authors thank Wenhao Ding of Carnegie Mellon University and Rong Fu of Tsinghua University for useful discussions about the conducted experiments.

References

[1] C. Ekanadham, D. Tranchina, and E. P. Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE transactions on signal processing*, vol. 59, no. 10, pp. 4735–4744, 2011.

[2] D. Donoho, "On minimum entropy deconvolution," in *Applied time series analysis II*. Elsevier, 1981, pp. 565–608.

[3] Y. Lou, B. Zhang, J. Wang, and M. Jiang, "Blind deconvolution for symmetric point-spread functions," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006, pp. 3459–3462.

[4] S.-i. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*. IEEE, 1997, pp. 101–104.

[5] N. Tian, S.-H. Byun, K. Sabra, and J. Romberg, "Multichannel myopic deconvolution in underwater acoustic channels via low-rank recovery," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3337–3348, 2017.

[6] D. R. Gitelman, W. D. Penny, J. Ashburner, and K. J. Friston, "Modeling regional and psychophysiologic interactions in fmri: the importance of hemodynamic deconvolution," *Neuroimage*, vol. 19, no. 1, pp. 200–207, 2003.

[7] T. Kaaresen, Kjetil F a nd Taxt, "Multichannel blind deconvolution of seismic signals," *Geophysics*, vol. 63, no. 6, pp. 2093–2107, 1998.

[8] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet, "Euclid in a taxicab: Sparse blind deconvolution with smoothed ℓ_1/ℓ_2 regularization," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 539–543, 2014.

[9] H. Zhang, D. Wipf, and Y. Zhang, "Multi-observation blind deconvolution with an adaptive sparse prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1628–1643, 2014.

[10] ——, "Multi-image blind deblurring using a coupled adaptive sparse prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1051–1058.

[11] S. Cho and S. Lee, "Fast motion deblurring," in *ACM SIGGRAPH Asia 2009 papers*, 2009, pp. 1–8.

[12] J. Yang, Y. Zhang, and W. Yin, "An efficient tvl1 algorithm for deblurring multichannel images corrupted by impulsive noise," *SIAM Journal on Scientific Computing*, vol. 31, no. 4, pp. 2842–2865, 2009.

[13] T. Strohmer, "Four short stories about Toeplitz matrix calculations," *Linear Algebra and its Applications*, vol. 343, pp. 321–344, 2002.

[14] Y. Li and Y. Bresler, "Global geometry of multichannel sparse blind deconvolution on the sphere," in *Advances in Neural Information Processing Systems*, 2018, pp. 1132–1143.

[15] Y. Li, K. Lee, and Y. Bresler, "A unified framework for identifiability analysis in bilinear inverse problems with applications to subspace and sparsity models," *arXiv preprint arXiv:1501.06120*, 2015.

[16] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere I: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2017.

[17] D. A. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," in *Conference on Learning Theory*, 2012, pp. 1–37.

[18] Y. Chi, "Guaranteed blind sparse spikes deconvolution via lifting and convex optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 782–794, 2016.

[19] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *Information Theory, IEEE Transactions on*, vol. 60, no. 3, pp. 1711–1732, 2014.

[20] Y. Bai, Q. Jiang, and J. Sun, "Subgradient descent learns orthogonal dictionaries," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[21] D. Gilboa, S. Buchanan, and J. Wright, "Efficient dictionary learning with gradient descent," in *International Conference on Machine Learning*, 2019, pp. 2252–2259.

[22] L. Wang and Y. Chi, "Blind deconvolution from multiple sparse inputs," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1384–1388, 2016.

[23] Q. Qu, X. Li, and Z. Zhu, "A nonconvex approach for exact and efficient multichannel sparse blind deconvolution," in *Advances in Neural Information Processing Systems*, 2019, pp. 4015–4026.

[24] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, p. 115002, 2015.

[25] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and computational harmonic analysis*, vol. 47, no. 3, pp. 893–934, 2019.

[26] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution," *Foundations of Computational Mathematics*, pp. 1–182, 2019.

[27] R. Gribonval, G. Chardon, and L. Daudet, "Blind calibration for compressed sensing by convex optimization," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 2713–2716.

[28] C. Bilen, G. Puy, R. Gribonval, and L. Daudet, "Convex optimization approaches for blind sensor calibration using sparsity," *Signal Processing, IEEE Transactions on*, vol. 62, no. 18, pp. 4847–4856, 2014.

[29] Y. Li, K. Lee, and Y. Bresler, "Blind gain and phase calibration via sparse spectral methods," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3097–3123, 2018.

[30] S. Choudhary and U. Mitra, "Sparse blind deconvolution: What cannot be done," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 3002–3006.

[31] Y. Chen, J. Fan, B. Wang, and Y. Yan, "Convex and nonconvex optimization are both minimax-optimal for noisy blind deconvolution," *arXiv preprint arXiv:2008.01724*, 2020.

[32] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy, "Composite optimization for robust blind deconvolution," *arXiv preprint arXiv:1901.01624*, 2019.

[33] K. Lee, N. Tian, and J. Romberg, "Fast and guaranteed blind multichannel deconvolution under a bilinear system model," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 4792–4818, 2018.

[34] S. Ling and T. Strohmer, "Self-calibration and bilinear inverse problems via linear least squares," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 252–292, 2018.

[35] Y. Xia and S. Li, "Identifiability of multichannel blind deconvolution and nonconvex regularization algorithm," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5299–5312, 2018.

[36] H.-W. Kuo, Y. Lau, Y. Zhang, and J. Wright, "Geometry and symmetry in short-and-sparse deconvolution," in *International Conference on Machine Learning*, 2019, pp. 3570–3580.

[37] Y. Zhang, H.-W. Kuo, and J. Wright, "Structured local optima in sparse blind deconvolution," *IEEE Transactions on Information Theory*, 2019.

[38] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Foundations of Computational Mathematics*, vol. 18, no. 5, pp. 1131–1198, 2018.

[39] K. Luh and V. Vu, "Dictionary learning with few samples and matrix concentration," *IEEE Transactions on Information Theory*, vol. 62, no. 3, pp. 1516–1527, March 2016.

[40] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Transactions on Signal Processing*, vol. 61, no. 5, pp. 1072–1086, 2013.

[41] Y. Zhai, Z. Yang, Z. Liao, J. Wright, and Y. Ma, "Complete dictionary learning via ℓ_4-norm maximization over the orthogonal group," *arXiv preprint arXiv:1906.02435*, 2019.

[42] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *Conference on Learning Theory*, 2014, pp. 779–806.

[43] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries," in *Conference on Learning Theory*, 2014, pp. 123–137.

[44] B. Barak, J. A. Kelner, and D. Steurer, "Dictionary learning and tensor decomposition via the sum-of-squares method," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, 2015, pp. 143–151.

[45] S. Arora, R. Ge, T. Ma, and A. Moitra, "Simple, efficient, and neural algorithms for sparse coding," in *Conference on Learning Theory*, 2015, pp. 113–149.

[46] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, Oct 2019.

[47] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 14 – 31, 2018.

[48] Y. Zhang, Q. Qu, and J. Wright, "From symmetry to geometry: Tractable nonconvex problems," *arXiv preprint arXiv:2007.06753*, 2020.

[49] S. Ling, R. Xu, and A. S. Bandeira, "On the landscape of synchronization networks: A perspective from nonconvex optimization," *SIAM Journal on Optimization*, vol. 29, no. 3, pp. 1879–1907, 2019.

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

[51] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the burer-monteiro approach," in *Artificial Intelligence and Statistics*, 2017, pp. 65–74.

[52] 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*, 2017, pp. 1233–1242.

[53] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[54] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3614–3628, 2018.

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

[56] T. T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *The Annals of Statistics*, vol. 44, no. 5, pp. 2221–2251, 2016.

[57] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow," in *International conference on machine learning*, 2016, pp. 1022–1031.

[58] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms," *Journal of Machine Learning Research*, vol. 18, no. 141, pp. 1–35, 2017.

[59] Y. Li, Y. Chi, H. Zhang, and Y. Liang, "Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent," *Information and Inference: A Journal of the IMA*, vol. 9, no. 2, pp. 289–325, 2020.

[60] Y. Chen and E. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Communications on Pure and Applied Mathematics*, vol. 70, no. 5, pp. 822–883, 2017.

[61] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust PCA," in *Advances in Neural Information Processing Systems*, 2014, pp. 1107–1115.

[62] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *International Conference Machine Learning*, 2016, pp. 964–973.

[63] T. Tong, C. Ma, and Y. Chi, "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent," *arXiv preprint arXiv:2005.08898*, 2020.

[64] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," *IEEE Transactions on Information Theory*, vol. 67, no. 3, pp. 1928–1950, 2021.

[65] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval," *Mathematical Programming*, vol. 176, no. 1-2, pp. 5–37, 2019.

[66] M. Ferreira Da Costa and Y. Chi, "Self-calibrated super resolution," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers.* IEEE, 2019, pp. 230–234.

[67] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. LeCun, "Learning convolutional feature hierarchies for visual recognition," in *Advances in neural information processing systems*, 2010, pp. 1090–1098.

[68] Q. Qu, Y. Zhai, X. Li, Y. Zhang, and Z. Zhu, "Geometric analysis of nonconvex optimization landscapes for overcomplete learning," in *International Conference on Learning Representations*, 2019.

[69] R. Vershynin, *High-dimensional probability: An introduction with applications in data science.* Cambridge University Press, 2018, vol. 47.

[70] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[71] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press, 2019, vol. 48.

[72] N. J. Higham, *Functions of matrices: theory and computation.* SIAM, 2008, vol. 104.

A Prerequisites

For convenience, let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the inputs $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p]$. Denote the first-order derivative of $\psi_\mu(x)$ as $\psi'_\mu(x) = \tanh(x/\mu)$ and the second-order derivative as $\psi''_\mu(x) = \left(1 - \tanh^2(x/\mu)\right)/\mu$. The gradient of $\psi_\mu(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h})$ with respect to $\boldsymbol{h}$ can be written as

$$\nabla_{\boldsymbol{h}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}) = \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right), \tag{32}$$

where, with slight abuse of notation, we allow $\tanh(\cdot)$ to take a vector-value in an entry-wise manner.

Recalling the reparameterization $\boldsymbol{h} = \boldsymbol{h}(\boldsymbol{w}) = \left(\boldsymbol{w}, \sqrt{1 - \|\boldsymbol{w}\|_2^2}\right)$, let $\boldsymbol{J_h}(\boldsymbol{w})$ be the Jacobian matrix of $\boldsymbol{h}(\boldsymbol{w})$, i.e.

$$\boldsymbol{J_h}(\boldsymbol{w}) = \left[\mathbf{I}, -\frac{\boldsymbol{w}}{h_n(\boldsymbol{w})}\right] \in \mathbb{R}^{(n-1)\times n}, \tag{33}$$

where $h_n(\boldsymbol{w}) = \sqrt{1 - \|\boldsymbol{w}\|_2^2}$ is the last entry of $\boldsymbol{h}(\boldsymbol{w})$. By the chain rule, the gradient of $\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})\right)$ with respect to $\boldsymbol{w}$ is given as

$$\nabla_{\boldsymbol{w}}\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})\right) = \boldsymbol{J_h}(\boldsymbol{w})\nabla_{\boldsymbol{h}}\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}\right) = \boldsymbol{J_h}(\boldsymbol{w})\mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right). \tag{34}$$

Moreover, the Hessian of $\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})\right)$ is given as

$$\begin{aligned}\nabla^2_{\boldsymbol{w}}\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})\right) = \frac{1}{\mu}\boldsymbol{J_h}(\boldsymbol{w})\mathcal{C}(\boldsymbol{x}_i)^\top\left[\mathbf{I} - \operatorname{diag}\left(\tanh^2\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right)\right]\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{J_h}(\boldsymbol{w})^\top \\ - \frac{1}{h_n}\mathcal{S}_{n-1}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\boldsymbol{J_h}(\boldsymbol{w})\boldsymbol{J_h}(\boldsymbol{w})^\top.\end{aligned} \tag{35}$$

A.1 Useful Concentration Inequalities

We first introduce some notation and properties of sub-Gaussian variables. A random variable X is called sub-Gaussian if its sub-Gaussian norm satisfies $\|X\|_{\psi_2} < \infty$ [69]. Similarly, we have $\|\boldsymbol{x}\|_{\psi_2} < \infty$ for a sub-Gaussian random vector $\boldsymbol{x}$ [69, Definition 3.4.1]. For Bernoulli-Gaussian random variables / vectors, we have the following two facts, which imply that they are also sub-Gaussian.

Fact 1 ([20, Lemma F.1])**.** *A random variable $X \in \mathrm{BG}(\theta)$ is sub-Gaussian, i.e. $\|X\|_{\psi_2} \le C_a$ for some constant C_a. Similarly, for a random vector $\boldsymbol{x} \sim_{iid} \mathrm{BG}(\theta)$ and any deterministic vector $\boldsymbol{v} \in \mathbb{R}^n$, we have $\left\|\boldsymbol{v}^\top\boldsymbol{x}\right\|_{\psi_2} \le C_b\|\boldsymbol{v}\|_2$ for some constant C_b.*

Fact 2 ([16, Lemma 21])**.** *Assume $\boldsymbol{x}$, $\boldsymbol{y} \in \mathbb{R}^n$ satisfy $\boldsymbol{x} \sim_{iid} \mathrm{BG}(\theta)$ and $\boldsymbol{y} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then for any deterministic vector $\boldsymbol{u} \in \mathbb{R}^n$, we have $\mathbb{E}(\left|\boldsymbol{u}^\top\boldsymbol{x}\right|^m) \le \mathbb{E}(\left|\boldsymbol{u}^\top\boldsymbol{y}\right|^m)$, and $\mathbb{E}(\|\boldsymbol{x}\|_2^m) \le \mathbb{E}(\|\boldsymbol{y}\|_2^m)$ for all integers $m \ge 1$.*

The second fact allows us to bound the moments of a Bernoulli-Gaussian vector via the moments of a Gaussian vector, which are given below.

Fact 3 ([16, Lemma 35])**.** *Let $\boldsymbol{y} \in \mathbb{R}^n$ be $\boldsymbol{y} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{I})$, we have for any $m \ge 1$, $\mathbb{E}\left(\|\boldsymbol{y}\|_2^m\right) \le m! n^{m/2}$.*

In addition, let us list a few more useful facts about sub-Gaussian random variables.

Fact 4 ([69, Lemma 2.6.8])**.** *If X is sub-Gaussian, then $X - \mathbb{E}X$ is also sub-Gaussian with $\|X - \mathbb{E}X\|_{\psi_2} \le C\|X\|_{\psi_2}$ for some constant C.*

Fact 5 ([69, Proposition 2.6.1])**.** *If $X_1, X_2, \cdots, X_n$ are zero-mean independent sub-Gaussian random variables, then there exists some constant C such that $\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \le C\sum_{i=1}^n \|X_i\|_{\psi_2}^2$.*

Fact 6 ([69, Eq. (2.14-2.15)])**.** *If X is sub-Gaussian, it satisfies the following bounds:*

$$\mathbb{P}\left(|X| \ge t\right) \le 2\exp\left(-ct^2/\|X\|_{\psi_2}^2\right) \quad \forall t \ge 0, \quad \left(\mathbb{E}|X|^m\right)^{1/m} \le C\sqrt{m}\|X\|_{\psi_2} \quad \forall m \ge 1,$$

where c, C are some universal constants.

Combining standard tail bounds with the union bound, we have the following facts.

Fact 7. *Let $\{\boldsymbol{x}_i\}_{i=1}^p \in \mathbb{R}^n$ be independent sub-Gaussian vectors with $\|\boldsymbol{x}_i\|_{\psi_2} \le B$ for some constant B. Then there exists some universal constant C such that with probability at least $1 - p^{-8}$, we have*

$$\max_{i\in[p]}\|\boldsymbol{x}_i\|_2 \le CB\sqrt{n\log p}.$$

Fact 8. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be $\boldsymbol{X} \sim_{iid} \mathrm{BG}(\theta)$, where $\theta \in (0, 1/2)$. With probability at least $1 - \theta(np)^{-7}$, we have*

$$\|\boldsymbol{X}\|_\infty = \max_{i,j} |X_{ij}| \le 4\sqrt{\log(np)}.$$

Finally, let us record the useful Bernstein's inequality for random vectors and matrices, which does not require the quantities of interest to be centered. This is a direct consequence of Fact 4 on centering and [70, Theorem 6.2].

Lemma 7 (Moment-controlled Bernstein's inequality)**.** *Let $\{\boldsymbol{X}_k \in \mathbb{R}^{n\times n}\}_{k=1}^p$ be a set of independent random matrices. Assume there exist σ, R such that for all $m \ge 2$, $\mathbb{E}\left(\|\boldsymbol{X}_k\|^m\right) \le \frac{m!}{2}\sigma^2 R^{m-2}$. Denote $\boldsymbol{S} = \frac{1}{p}\sum_{k=1}^p \boldsymbol{X}_k$, then we have for any $t > 0$,*

$$\mathbb{P}\left(\|\boldsymbol{S} - \mathbb{E}\boldsymbol{S}\| > t\right) \le 2n \exp\left(\frac{-pt^2}{2\sigma^2 + 2Rt}\right).$$

Let $\{\boldsymbol{x}_k \in \mathbb{R}^n\}_{k=1}^p$ be a set of independent random vectors. Assume there exist σ, R such that $\mathbb{E}\left(\|\boldsymbol{x}_k\|_2^m\right) \le \frac{m!}{2}\sigma^2 R^{m-2}$. Denote $\boldsymbol{s} = \frac{1}{p}\sum_{k=1}^p \boldsymbol{x}_k$, then we have for any $t > 0$,

$$\mathbb{P}\left(\|\boldsymbol{s} - \mathbb{E}\boldsymbol{s}\|_2 > t\right) \le (n+1) \exp\left(\frac{-pt^2}{2\sigma^2 + 2Rt}\right).$$

A.2 Technical Lemmas

In this section, we provide some technical lemmas that are used throughout the proof. We start with some useful properties about the $\tanh(\cdot)$ function since it appears frequently in our derivation.

Lemma 8. *Let $X \sim \mathcal{N}(0, \sigma_x^2)$, $Y \sim \mathcal{N}(0, \sigma_y^2)$, then we have*

$$\begin{aligned}
\mathbb{E}\left[\tanh(aX)X\right] &= a\sigma_x^2\mathbb{E}\left[1 - \tanh^2(aX)\right], \\
\mathbb{E}\left[\tanh\left(a(X+Y)\right)X\right] &= a\sigma_x^2\mathbb{E}\left[1 - \tanh^2\left(a(X+Y)\right)\right].
\end{aligned}$$

Proof. Using integration by parts, we have

$$\begin{aligned}
&\mathbb{E}\left[\tanh(aX)X\right] = \frac{1}{\sqrt{2\pi}\sigma_x}\int_{-\infty}^{\infty} \tanh(aX) X \exp\left(\frac{-X^2}{2\sigma_x^2}\right) dX \\
&= -\frac{2\sigma_x^2}{\sqrt{2\pi}\sigma_x}\tanh(aX)\exp\left(\frac{-X^2}{2\sigma_x^2}\right)\Big|_0^\infty + \frac{1}{\sqrt{2\pi}\sigma_x}\int_{-\infty}^{\infty} a\sigma_x^2\left(1 - \tanh^2(aX)\right)\exp\left(\frac{-X^2}{2\sigma_x^2}\right) dX \\
&= a\sigma_x^2\mathbb{E}\left[1 - \tanh^2(aX)\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[\tanh\left(a(X+Y)\right)X\right] &= \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{\infty} X\exp\left(\frac{-X^2}{2\sigma_x^2}\right)\int_{-\infty}^{\infty}\tanh(a(X+Y))\exp\left(\frac{-Y^2}{2\sigma_y^2}\right) dY dX \\
&= -\frac{1}{2\pi\sigma_x\sigma_y}\cdot\sigma_x^2\left[\int_{-\infty}^{\infty}\tanh(a(X+Y))\exp\left(\frac{-Y^2}{2\sigma_y^2}\right)dY\right]\exp\left(\frac{-X^2}{2\sigma_x^2}\right)\Big|_{-\infty}^{\infty} \\
&\quad + \frac{1}{2\pi\sigma_x\sigma_y}\iint_{-\infty}^{\infty} a\sigma_x^2\left(1 - \tanh^2(a(X+Y))\right)\exp\left(\frac{-X^2}{2\sigma_x^2}\right)\exp\left(\frac{-Y^2}{2\sigma_y^2}\right) dY dX \\
&= a\sigma_x^2\mathbb{E}\left[1 - \tanh^2(a(X+Y))\right],
\end{aligned}$$

where we used the fact that the first term in the second line is 0. $\square$

Lemma 9. *The functions $\psi_\mu'(x) = \tanh(x/\mu)$ and $\psi_\mu''(x) = \left(1 - \tanh^2\left(x/\mu\right)\right)/\mu$ are Lipschitz continuous with Lipschitz constants $1/\mu$ and $2/\mu^2$, respectively.*

Proof. Since $\psi_\mu(x)$ is continuous and third-order differentiable, we have for any x and x',

$$\left|\psi'_\mu(x) - \psi'_\mu(x')\right| \le \left| \int_x^{x'} \psi''_\mu(z) dz \right| \le |x - x'| \max_z |\psi''_\mu(z)| \le \frac{|x - x'|}{\mu},$$

and

$$\left|\psi''_\mu(x) - \psi''_\mu(x')\right| \le \left| \int_x^{x'} \psi'''_\mu(z) dz \right| \le \int_x^{x'} \left| -\frac{2}{\mu^2} \tanh\left(\frac{z}{\mu}\right) \left(1 - \tanh^2\left(\frac{z}{\mu}\right)\right) \right| dz \le \frac{2\,|x - x'|}{\mu^2},$$

where we use the fact that $|\tanh(x)| \le 1$ and $1 - \tanh^2(x) \le 1$ for all $x \in \mathbb{R}$. $\square$

Lemma 10. *Let $\boldsymbol{x} \sim_{iid} \mathrm{BG}(\theta)$ for $\theta \in (0, 1]$. There exist some constants c_1 and c_2 such that*

$$\mathbb{P}\left(\|\mathcal{C}(\boldsymbol{x})\| \ge t\right) \le 2n \exp\left(\frac{-t^2}{c_1 n}\right), \quad \textit{and} \quad \mathbb{E}\,\|\mathcal{C}(\boldsymbol{x})\|^{2m} \le \frac{m!}{2}\left(c_2 n \log n\right)^m$$

for all $m \ge 1$.

Proof. Since a circulant matrix is diagonalizable by the DFT matrix, the spectral norm of $\mathcal{C}(\boldsymbol{x})$ is the maximum magnitude of the DFT coefficients of $\boldsymbol{x}$, where the ith coefficient is given as $\widehat{x}_i = \boldsymbol{f}_i^{\mathsf{H}} \boldsymbol{x}$, where $\boldsymbol{f}_i = [1, e^{j2\pi i/n}, \cdots, e^{j2\pi i(n-1)/n}]^\top$ is the ith column of the DFT matrix. Since $\boldsymbol{x} \sim_{iid} \mathrm{BG}(\theta)$ is sub-Gaussian, by Fact 1, $\widehat{x}_i$ is also sub-Gaussian with $\|\widehat{x}_i\|_{\psi_2} \le C\|\boldsymbol{f}_i\|_2 = C\sqrt{n}$. Therefore, by the union bound, together with Fact 6, we have

$$\mathbb{P}\left(\|\mathcal{C}(\boldsymbol{x})\| \ge t\right) = \mathbb{P}\left(\max_{i \in [n]} |\widehat{x}_i| \ge t\right) \le 2n \exp\left(\frac{-t^2}{c_1 n}\right),$$

for some constant c_1. Equipped with the above bound, we can bound the moments of $\|\mathcal{C}(\boldsymbol{x})\|^2$ by

$$\mathbb{E}\,\|\mathcal{C}(\boldsymbol{x})\|^{2m} = \int_0^\infty \mathbb{P}(\|\mathcal{C}(\boldsymbol{x})\|^{2m} > u) du = \int_0^\infty \mathbb{P}(\|\mathcal{C}(\boldsymbol{x})\| > t) \cdot 2mt^{2m-1} dt, \tag{36}$$

where the second equality follows by a change of variable $t = u^{1/2m}$. To continue, we break the bound as

$$\begin{aligned}
\mathbb{E}\,\|\mathcal{C}(\boldsymbol{x})\|^{2m} &\le \int_0^{2\sqrt{c_1 n \log n}} 1 \cdot 2mt^{2m-1} dt + \int_{2\sqrt{c_1 n \log n}}^\infty 2n \exp\left(\frac{-t^2}{c_1 n}\right) 2mt^{2m-1} dt \\
&\le \left(4c_1 n \log n\right)^m + \int_0^\infty \exp\left(\frac{-t^2}{2c_1 n}\right) 2mt^{2m-1} dt \\
&= \left(4c_1 n \log n\right)^m + \left(2c_1 n\right)^m m! \le \frac{m!}{2}\left(c_2 n \log n\right)^m,
\end{aligned}$$

where the second line used the fact $\exp\left(\frac{-t^2}{2c_1 n}\right) > 2n \exp\left(\frac{-t^2}{c_1 n}\right)$ when $t \ge 2\sqrt{c_1 n \log n}$, and the third line used the definition of the Gamma function. The proof is completed. $\square$

Lemma 11. *Let $\{\boldsymbol{x}_i\}_{i=1}^p \in \mathbb{R}^n$ be drawn according to $\boldsymbol{x}_i \sim_{iid} \mathrm{BG}(\theta)$, $\theta \in (0, 1/2)$. There exists some constant C such that*

$$\left\| \frac{1}{\theta n p} \sum_{i=1}^p \mathcal{C}(\boldsymbol{x}_i)^\top \mathcal{C}(\boldsymbol{x}_i) - \mathbf{I} \right\| \le C\sqrt{\frac{\log^2 n \log p}{\theta^2 p}}$$

holds with probability at least $1 - 2np^{-8}$.

Proof. By assumption, it is easy to check

$$\mathbb{E}\left[\frac{1}{\theta n p} \sum_{i=1}^p \mathcal{C}(\boldsymbol{x}_i)^\top \mathcal{C}(\boldsymbol{x}_i)\right] = \mathbb{E}\left[\frac{1}{\theta n} \mathcal{C}(\boldsymbol{x}_1)^\top \mathcal{C}(\boldsymbol{x}_1)\right] = \mathbf{I}.$$

The remaining of the proof is to verify the quantities needed to apply Lemma 7. Specifically, we bound the mth moment of $\frac{1}{\theta n}\mathcal{C}(\boldsymbol{x}_i)^\top \mathcal{C}(\boldsymbol{x}_i)$ as

$$\mathbb{E}\left\|\frac{1}{\theta n}\mathcal{C}(\boldsymbol{x}_i)^\top \mathcal{C}(\boldsymbol{x}_i)\right\|^m = \frac{1}{\theta^m n^m}\mathbb{E}\left\|\mathcal{C}(\boldsymbol{x}_i)\right\|^{2m} \le \frac{m!}{2}\left(\frac{c\log n}{\theta}\right)^m,$$

where the last line comes from Lemma 10. Let $\sigma^2 = \frac{c^2\log^2 n}{\theta^2}$, $R = \frac{c\log n}{\theta}$ in Lemma 7, we have

$$\mathbb{P}\left(\left\|\frac{1}{\theta np}\sum_{i=1}^p \mathcal{C}(\boldsymbol{x}_i)^\top \mathcal{C}(\boldsymbol{x}_i) - \mathbf{I}\right\| \ge t\right) \le 2n\exp\left(\frac{-p\theta^2 t^2}{2c^2\log^2 n + 2c\theta t\log n}\right). \tag{37}$$

Setting $t = C\sqrt{\frac{\log^2 n\log p}{\theta^2 p}}$ for some sufficiently large C, we complete the proof. $\square$

B Proofs for Section 3.1

B.1 Proof of Lemma 1

Recall the two regions introduced in (14):

$$\mathcal{Q}_1 := \left\{\boldsymbol{w} : \frac{\mu}{4\sqrt{2}} \le \|\boldsymbol{w}\|_2 \le \sqrt{\frac{n-1}{n+\xi_0}}\right\}, \quad \mathcal{Q}_2 := \left\{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \le \frac{\mu}{4\sqrt{2}}\right\}.$$

We further divide Q_1 into two subregions,

$$\mathcal{R}_0 = \left\{\boldsymbol{w} : \frac{\mu}{4\sqrt{2}} \le \|\boldsymbol{w}\|_2 \le \frac{1}{20\sqrt{5}}\right\}, \qquad \mathcal{R}_1 = \left\{\boldsymbol{w} : \frac{1}{20\sqrt{5}} \le \|\boldsymbol{w}\|_2 \le \sqrt{\frac{n-1}{n+\xi_0}}\right\},$$

which we will prove the desired bound separately.

Note that

$$\mathbb{E}\phi_o(\boldsymbol{w}) = n\cdot\mathbb{E}\psi_\mu\left(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w})\right), \tag{38}$$

since every row of $\mathcal{C}(\boldsymbol{x})$ has the same distribution as $\boldsymbol{x} \sim_{iid} \mathrm{BG}(\theta)$. Therefore, the strong convexity bound (26b) in $\mathcal{Q}_2$ follows directly from the following lemma from [16, Proposition 8] by a multiplicative factor of n.

Lemma 12 ([16, Proposition 8])**.** *For any $\theta \in (0, 1/2)$, if $\mu \le \frac{1}{20\sqrt{n}}$, it holds for all $\boldsymbol{w}$ with $\|\boldsymbol{w}\|_2 \le \frac{\mu}{4\sqrt{2}}$ that $\nabla^2_{\boldsymbol{w}}\mathbb{E}\psi_\mu\left(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w})\right) \succeq \frac{\theta}{5\sqrt{2\pi}\mu}\mathbf{I}$.*

Similarly, by the following lemma from [16, Proposition 7], we have the desired bound (26a) in $\mathcal{R}_0$.

Lemma 13 ([16, Proposition 7])**.** *For any $\theta \in (0, 1/3)$, if $\mu \le 9/50$, it holds for all $\boldsymbol{w} \in \mathcal{R}_0$ such that*

$$\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}}\mathbb{E}\psi_\mu(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w}))}{\|\boldsymbol{w}\|_2} \ge \frac{\theta}{20\sqrt{2\pi}}.$$

Therefore, the remainder of the proof is to show that (26a) also applies to $\mathcal{R}_1$. To ease presentation, we introduce a few short-hand notations. For $\boldsymbol{x} = \boldsymbol{\Omega}\odot\boldsymbol{z} \sim_{iid} \mathrm{BG}(\theta) \in \mathbb{R}^n$, we denote the first $n-1$ dimension of $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{\Omega}$ as $\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{z}}$ and $\bar{\boldsymbol{\Omega}}$, respectively. Denote $\mathcal{I}$ as the support of $\boldsymbol{\Omega}$ and $\mathcal{J}$ as the support of $\bar{\boldsymbol{\Omega}}$.

Note that it is easy to confirm the exchangeability of the expectation and derivatives [16, Lemma 31] as

$$\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}}\mathbb{E}\psi_\mu(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w}))}{\|\boldsymbol{w}\|_2} = \mathbb{E}\left[\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}}\psi_\mu\left(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w})\right)}{\|\boldsymbol{w}\|_2}\right], \tag{39a}$$

$$\nabla^2_{\boldsymbol{w}}\mathbb{E}\psi_\mu\left(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w})\right) = \mathbb{E}\nabla^2_{\boldsymbol{w}}\psi_\mu\left(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w})\right). \tag{39b}$$

Thus, plugging in (34), we rewrite the expectation of the directional gradient as following:

$$\mathbb{E}\left[\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}}\psi_\mu\left(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w})\right)}{\|\boldsymbol{w}\|_2}\right] = \frac{1}{\|\boldsymbol{w}\|_2}\mathbb{E}\left[\tanh\left(\frac{\boldsymbol{x}^\top\cdot \boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\cdot\left(\boldsymbol{w}^\top \bar{\boldsymbol{x}} - \frac{x_n\|\boldsymbol{w}\|_2^2}{h_n}\right)\right]$$

$$= \frac{(1-\theta)}{\|\boldsymbol{w}\|_2} \underbrace{\mathbb{E}_{\bar{\boldsymbol{x}}}\left[\tanh\left(\frac{\boldsymbol{w}^\top \bar{\boldsymbol{x}}}{\mu}\right)\boldsymbol{w}^\top \bar{\boldsymbol{x}}\right]}_{I_1} + \frac{\theta}{\|\boldsymbol{w}\|_2} \underbrace{\mathbb{E}_{\bar{\boldsymbol{x}},z_n}\left[\tanh\left(\frac{\boldsymbol{w}^\top \bar{\boldsymbol{x}} + h_n z_n}{\mu}\right)\left(\boldsymbol{w}^\top \bar{\boldsymbol{x}} - \frac{\|\boldsymbol{w}\|_2^2}{h_n} z_n\right)\right]}_{I_2}, \tag{40}$$

where the second line is expanded over the distribution of $\Omega_n \sim \text{Bernoulli}(\theta)$. Conditioned on the support of $\bar{\boldsymbol{\Omega}}$, we have $X = \boldsymbol{w}^\top \bar{\boldsymbol{x}} | \bar{\boldsymbol{\Omega}} \sim \mathcal{N}(0, \|\boldsymbol{w}_{\mathcal{J}}\|_2^2)$. Moreover, denote $Y = h_n z_n \sim \mathcal{N}(0, h_n^2)$. Therefore, invoking Lemma 8, we can express I_1 and I_2 respectively as

$$I_1 = \mathbb{E}_{\bar{\boldsymbol{\Omega}}}\left[\mathbb{E}_X\left(\tanh\left(\frac{X}{\mu}\right)X\right)\right] = \frac{1}{\mu}\mathbb{E}_{\bar{\boldsymbol{\Omega}}}\left[\|\boldsymbol{w}_{\mathcal{J}}\|_2^2\,\mathbb{E}_X\left(1-\tanh^2\left(\frac{X}{\mu}\right)\right)\right],$$

$$I_2 = \mathbb{E}_{\bar{\boldsymbol{\Omega}}}\left[\mathbb{E}_{X,Y}\left(\tanh\left(\frac{X+Y}{\mu}\right)\left(X - \frac{\|\boldsymbol{w}\|_2^2}{h_n^2}Y\right)\right)\right] = \frac{1}{\mu}\mathbb{E}_{\bar{\boldsymbol{\Omega}}}\left[\left(\|\boldsymbol{w}_{\mathcal{J}}\|_2^2 - \|\boldsymbol{w}\|_2^2\right)\mathbb{E}_{X,Y}\left(1-\tanh^2\left(\frac{X+Y}{\mu}\right)\right)\right].$$

Plugging the above quantities back into (40), and using $\|\boldsymbol{w}_{\mathcal{J}}\|_2^2 = \sum_{i=1}^{n-1} w_i^2 \mathbb{1}\{\Omega_i = 1\}$, $\|\boldsymbol{w}_{\mathcal{J}^c}\|_2^2 = \sum_{i=1}^{n-1} w_i^2 \mathbb{1}\{\Omega_i = 0\}$, we arrive at

$$\begin{aligned}
\mathbb{E}\left[\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \psi_\mu\left(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w})\right)}{\|\boldsymbol{w}\|_2}\right] &= \frac{(1-\theta)}{\mu\|\boldsymbol{w}\|_2}\mathbb{E}_{\bar{\boldsymbol{\Omega}}}\left[\sum_{i=1}^{n-1} w_i^2 \mathbb{1}\{\Omega_i = 1\}\cdot \mathbb{E}_{\boldsymbol{z}}\left(1-\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i\}} + w_i z_i \mathbb{1}\{\Omega_i = 1\}}{\mu}\right)\right)\right] \\
&\quad - \frac{\theta}{\mu\|\boldsymbol{w}\|_2}\mathbb{E}_{\bar{\boldsymbol{\Omega}}}\left[\sum_{i=1}^{n-1} w_i^2 \mathbb{1}\{\Omega_i = 0\}\cdot \mathbb{E}_{\boldsymbol{z}}\left(1-\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i\}} + w_i z_i \mathbb{1}\{\Omega_i = 1\} + h_n z_n}{\mu}\right)\right)\right] \\
&= \frac{1}{\mu\|\boldsymbol{w}\|_2}\sum_{i=1}^{n-1} w_i^2 Q_i,
\end{aligned} \tag{41}$$

where Q_i is written as

$$\begin{aligned}
Q_i = (1-\theta)\mathbb{E}_{\bar{\boldsymbol{\Omega}}}&\left[\mathbb{1}\{\Omega_i = 1\}\cdot \mathbb{E}_{\boldsymbol{z}}\left(1-\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i\}} + w_i z_i \mathbb{1}\{\Omega_i = 1\}}{\mu}\right)\right)\right] \\
&- \theta\mathbb{E}_{\bar{\boldsymbol{\Omega}}}\left[\mathbb{1}\{\Omega_i = 0\}\cdot \mathbb{E}_{\boldsymbol{z}}\left(1-\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i\}} + w_i z_i \mathbb{1}\{\Omega_i = 1\} + h_n z_n}{\mu}\right)\right)\right].
\end{aligned}$$

Evaluating $\mathbb{E}_{\bar{\boldsymbol{\Omega}}}$ over $\bar{\boldsymbol{\Omega}} \setminus \{i\}$ and Ω_i sequentially, and combining terms, we can rewrite Q_i as,

$$\begin{aligned}
Q_i &= (1-\theta)\theta\mathbb{E}_{\bar{\boldsymbol{\Omega}}\setminus\{i\}}\left[\mathbb{E}_{\boldsymbol{z}}\left(1-\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i\}} + w_i z_i}{\mu}\right)\right) - \mathbb{E}_{\boldsymbol{z}}\left(1-\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i\}} + h_n z_n}{\mu}\right)\right)\right] \\
&= (1-\theta)\theta\mathbb{E}_{\bar{\boldsymbol{\Omega}}\setminus\{i\}}\Bigg[\underbrace{\mathbb{E}_{\boldsymbol{z}}\left(\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i\}} + h_n z_n}{\mu}\right) - \tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i\}} + w_i z_i}{\mu}\right)\right)}_{=:K_i}\Bigg].
\end{aligned} \tag{42}$$

Our goal is to lower bound Q_i for all $i \in [n-1]$. Without loss of generality, we denote the index of the largest entry of $\boldsymbol{w}$ in magnitude as i_0, i.e, $|w_{i_0}| \geq |w_j|$, $\forall j \in [n-1]$. We first claim

$$Q_j \geq Q_{i_0}, \quad \forall j \in [n-1], \tag{43}$$

whose proof is given at the end of this subsection. With this claim, we only need to lower bound Q_{i_0}. We proceed to lower bound $\mathbb{E}_{\boldsymbol{z}}[K_{i_0}]$. Let $X := \boldsymbol{w}_{\setminus\{i_0\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0\}} + w_{i_0} z_{i_0} | \bar{\boldsymbol{\Omega}} \sim \mathcal{N}(0, \left\|\boldsymbol{w}_{\mathcal{J}\setminus\{i_0\}}\right\|_2^2 + w_{i_0}^2) := \mathcal{N}(0, \sigma_X^2)$ and $Y := \boldsymbol{w}_{\setminus\{i_0\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0\}} + h_n z_n | \bar{\boldsymbol{\Omega}} \sim \mathcal{N}(0, \left\|\boldsymbol{w}_{\mathcal{J}\setminus\{i_0\}}\right\|_2^2 + h_n^2) := \mathcal{N}(0, \sigma_Y^2)$. By the fundamental theorem of calculus, we have

$$K_{i_0} = \tanh^2\left(\frac{Y}{\mu}\right) - \tanh^2\left(\frac{X}{\mu}\right)$$

$$
\begin{aligned}
&= \frac{2}{\mu}\int_{|X|}^{|Y|} \tanh\left(\frac{x}{\mu}\right)\cdot\left(1-\tanh^2\left(\frac{x}{\mu}\right)\right)dx \\
&\geq \frac{2}{\mu}\int_{|X|}^{|Y|} \left[2\exp\left(\frac{-2x}{\mu}\right)-\exp\left(\frac{-4x}{\mu}\right)\right]\left[1-2\exp\left(\frac{-2x}{\mu}\right)\right]dx \\
&\geq \frac{2}{\mu}\int_{|X|}^{|Y|} \left[2\exp\left(\frac{-2x}{\mu}\right)-5\exp\left(\frac{-4x}{\mu}\right)\right]dx \\
&= \underbrace{2\left[\exp\left(\frac{-2\,|X|}{\mu}\right)-\exp\left(\frac{-2\,|Y|}{\mu}\right)\right]}_{K_1}-\underbrace{\frac{5}{2}\left[\exp\left(\frac{-4\,|X|}{\mu}\right)-\exp\left(\frac{-4\,|Y|}{\mu}\right)\right]}_{K_2}, \qquad (44)
\end{aligned}
$$

where the third line follows from the bounds $2\exp(-2x/\mu)-\exp(-4x/\mu)\leq 1-\tanh^2(x/\mu)$ and $\tanh(x/\mu)\leq 1-\exp(-2x/\mu)$ in [16, Lemma 29]. To continue, we record the lemma rephrased from [16, Lemma 32, 40] and obtain the following lemma by directly repeating integration by parts.

Lemma 14 ([16, Lemma 32, 40])**.** *Let $X\sim\mathcal{N}(0,\sigma_X^2)$. For any $a>0$, we have*

$$
\frac{1}{\sqrt{2\pi}}\left(\frac{1}{a\sigma_X}-\frac{1}{a^3\sigma_X^3}+\frac{3}{a^5\sigma_X^5}-\frac{15}{a^7\sigma_X^7}\right)\leq \mathbb{E}\left[\exp(-aX)\mathbb{1}\{X>0\}\right]\leq\frac{1}{\sqrt{2\pi}}\left(\frac{1}{a\sigma_X}-\frac{1}{a^3\sigma_X^3}+\frac{3}{a^5\sigma_X^5}\right).
$$

Therefore, K_1 can be bounded as

$$
\begin{aligned}
K_1 &= 2\mathbb{E}\left[\exp\left(\frac{-2\,|X|}{\mu}\right)-\exp\left(\frac{-2\,|Y|}{\mu}\right)\right] \\
&= 4\mathbb{E}\left[\exp\left(\frac{-2X}{\mu}\right)\mathbb{1}\{X>0\}-\exp\left(\frac{-2Y}{\mu}\right)\mathbb{1}\{Y>0\}\right] \\
&\geq \frac{4}{\sqrt{2\pi}}\left(\frac{\mu}{2\sigma_X}-\frac{\mu^3}{8\sigma_X^3}+\frac{3\mu^5}{32\sigma_X^5}-\frac{15\mu^7}{2^7\sigma_X^7}\right)-\frac{4}{\sqrt{2\pi}}\left(\frac{\mu}{2\sigma_Y}-\frac{\mu^3}{8\sigma_Y^3}+\frac{3\mu^5}{32\sigma_Y^5}\right) \\
&= \frac{2}{\sqrt{2\pi}}\left[\left(\frac{\mu}{\sigma_X}-\frac{\mu}{\sigma_Y}\right)-\left(\frac{\mu^3}{4\sigma_X^3}-\frac{\mu^3}{4\sigma_Y^3}\right)+\left(\frac{3\mu^5}{16\sigma_X^5}-\frac{3\mu^5}{16\sigma_Y^5}\right)-\frac{15\mu^7}{2^6\sigma_X^7}\right].
\end{aligned}
$$

Similarly, we have

$$
K_2 \leq \frac{5}{4\sqrt{2\pi}}\left[\left(\frac{\mu}{\sigma_X}-\frac{\mu}{\sigma_Y}\right)-\left(\frac{\mu^3}{16\sigma_X^3}-\frac{\mu^3}{16\sigma_Y^3}\right)+\left(\frac{3\mu^5}{4^4\sigma_X^5}-\frac{3\mu^5}{4^4\sigma_Y^5}\right)-\frac{15\mu^7}{4^6\sigma_Y^7}\right].
$$

Plugging the above bounds back into (44), we have

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}}\left[K_{i_0}\right]\geq\mathbb{E}_{X,Y}\left[K_1-K_2\right] \\
&\geq \frac{2}{\sqrt{2\pi}}\left[\left(\frac{\mu}{\sigma_X}-\frac{\mu}{\sigma_Y}\right)-\left(\frac{\mu^3}{4\sigma_X^3}-\frac{\mu^3}{4\sigma_Y^3}\right)+\left(\frac{3\mu^5}{16\sigma_X^5}-\frac{3\mu^5}{16\sigma_Y^5}\right)-\frac{15\mu^7}{2^6\sigma_X^7}\right] \qquad (45)\\
&-\frac{5}{4\sqrt{2\pi}}\left[\left(\frac{\mu}{\sigma_X}-\frac{\mu}{\sigma_Y}\right)-\left(\frac{\mu^3}{16\sigma_X^3}-\frac{\mu^3}{16\sigma_Y^3}\right)+\left(\frac{3\mu^5}{4^4\sigma_X^5}-\frac{3\mu^5}{4^4\sigma_Y^5}\right)-\frac{15\mu^7}{4^6\sigma_Y^7}\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[\frac{3\mu}{4}\left(\frac{1}{\sigma_X}-\frac{1}{\sigma_Y}\right)-\frac{27\mu^3}{64}\left(\frac{1}{\sigma_X^3}-\frac{1}{\sigma_Y^3}\right)+\frac{113\mu^5}{4^5}\left(\frac{1}{\sigma_X^5}-\frac{1}{\sigma_Y^5}\right)-\frac{15\mu^7}{2^5\sigma_X^7}-\frac{75\mu^7}{4^7\sigma_Y^7}\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[\left(\frac{1}{\sigma_X}-\frac{1}{\sigma_Y}\right)\left(\frac{3\mu}{4}-\frac{27\mu^3}{64}\left(\frac{1}{\sigma_X^2}+\frac{1}{\sigma_Y^2}+\frac{1}{\sigma_X\sigma_Y}\right)\right)+\frac{113\mu^5}{4^5}\left(\frac{1}{\sigma_X^5}-\frac{1}{\sigma_Y^5}\right)-\frac{15\mu^7}{2^5\sigma_X^7}-\frac{75\mu^7}{4^7\sigma_Y^7}\right] \\
&\geq \frac{1}{\sqrt{2\pi}}\left[\left(\frac{1}{\sigma_X}-\frac{1}{\sigma_Y}\right)\left(\frac{3\mu}{4}-\frac{27\mu^3}{64}\left(\frac{1}{\sigma_X^2}+\frac{1}{\sigma_Y^2}+\frac{1}{\sigma_X\sigma_Y}\right)\right)-\frac{\mu^7}{2\sigma_X^7}\right] \qquad (46)
\end{aligned}
$$

where the last line follows from the fact $\sigma_X<\sigma_Y$ and $\frac{113\mu^5}{4^5}\left(\frac{1}{\sigma_X^5}-\frac{1}{\sigma_Y^5}\right)>0$.

To continue, since $\sigma_X = \sqrt{\left\|\boldsymbol{w}_{\mathcal{J}\setminus\{i_0\}}\right\|_2^2 + h_{i_0}^2} < 1$ and $\sigma_Y = \sqrt{\left\|\boldsymbol{w}_{\mathcal{J}\setminus\{i_0\}}\right\|_2^2 + h_n^2} < 1$,

$$\frac{1}{\sigma_X} - \frac{1}{\sigma_Y} = \frac{\sigma_Y^2 - \sigma_X^2}{\sigma_X \sigma_Y (\sigma_X + \sigma_Y)} \geq \frac{\sigma_Y^2 - \sigma_X^2}{2} = \frac{1}{2}\left(h_n^2 - h_{i_0}^2\right) \geq \frac{1}{2}\left[h_n^2 - \frac{1}{1+\xi_0} h_n^2\right] \geq \frac{\xi_0}{4n}, \tag{47}$$

where the second inequality uses the fact $h_n^2/h_{i_0}^2 \geq 1+\xi_0$, $h_n \geq 1/\sqrt{n}$ and $\xi_0 \in (0,1)$. In addition, as $|h_{i_0}| = \max_i w_i$ and $\|\boldsymbol{w}\|_2 > \frac{1}{20\sqrt{5}}$, we have $|h_{i_0}| \geq \frac{1}{20\sqrt{5n}}$. So we have $\frac{1}{\sigma_X} \leq \frac{1}{|h_{i_0}|} \leq 20\sqrt{5n}$, such that

$$\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2} + \frac{1}{\sigma_X \sigma_Y} \leq 2061 n \quad \Longrightarrow \quad \frac{27\mu^3}{64}\left(\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2} + \frac{1}{\sigma_X\sigma_Y}\right) \leq \frac{\mu}{4}, \tag{48}$$

provided $\mu \leq c n^{-1/2}$ for a sufficiently small $c > 0$. Plugging (47) and (48) back into (46), we have

$$\mathbb{E}_{\boldsymbol{z}}\left[K_{i_0}\right] \geq \frac{\mu\xi_0}{8\sqrt{2\pi}n} - \frac{\mu^7}{2\sqrt{2\pi}\sigma_X^7} \geq \frac{\mu\xi_0}{16\sqrt{2\pi}n}, \tag{49}$$

conditioned on the support $\bar{\boldsymbol{\Omega}}_{\setminus\{i_0\}}$, provided that $\frac{1}{\sigma_X^7} \leq (20\sqrt{5})^7 n^{7/2}$ and $\mu \leq c\xi_0^{1/6} n^{-3/4}$ for a sufficiently small $c > 0$.

Plugging (49) back into (42), then into (41) with the help of (43), finally, by the assumption $\|\boldsymbol{w}\|_2 \geq \frac{1}{20\sqrt{5}}$ and (39a), we have

$$\frac{\boldsymbol{w}^\top \mathbb{E}\nabla\phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} = \mathbb{E}\left[n\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}}\psi_\mu\left(\boldsymbol{x}^\top \boldsymbol{h}(\boldsymbol{w})\right)}{\|\boldsymbol{w}\|_2}\right] \geq n\frac{\|\boldsymbol{w}\|_2\,\theta(1-\theta)}{\mu}\frac{\mu\xi_0}{16\sqrt{2\pi}n} \geq \frac{\theta\xi_0}{480\sqrt{10\pi}}, \tag{50}$$

where the final bound follows from the constraint $\theta \in (0, 1/3)$.

Proof of (43): For any $j \in [n-1]$ and $j \neq i_0$, by evaluating $\bar{\boldsymbol{\Omega}} \setminus \{i_0\}$ over $\bar{\boldsymbol{\Omega}} \setminus \{i_0, j\}$ and Ω_j sequentially, we can rewrite Q_{i_0} as

$$\begin{aligned} Q_{i_0} &= (1-\theta)\theta^2 \mathbb{E}_{\bar{\boldsymbol{\Omega}}\setminus\{i_0,j\}}\left[\mathbb{E}_{\boldsymbol{z}}\left(\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + h_n z_n + w_j z_j}{\mu}\right) - \tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + w_{i_0} z_{i_0} + w_j z_j}{\mu}\right)\right)\right] \\ &+ (1-\theta)^2\theta \mathbb{E}_{\bar{\boldsymbol{\Omega}}\setminus\{i_0,j\}}\left[\mathbb{E}_{\boldsymbol{z}}\left(\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + h_n z_n}{\mu}\right) - \tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + w_{i_0} z_{i_0}}{\mu}\right)\right)\right].. \end{aligned} \tag{51}$$

Similarly, we can write Q_j as

$$\begin{aligned} Q_j &= (1-\theta)\theta^2 \mathbb{E}_{\bar{\boldsymbol{\Omega}}\setminus\{i_0,j\}}\left[\mathbb{E}_{\boldsymbol{z}}\left(\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + h_n z_n + w_{i_0} z_{i_0}}{\mu}\right) - \tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + w_{i_0} z_{i_0} + w_j z_j}{\mu}\right)\right)\right] \\ &+ (1-\theta)^2\theta \mathbb{E}_{\bar{\boldsymbol{\Omega}}\setminus\{i_0,j\}}\left[\mathbb{E}_{\boldsymbol{z}}\left(\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + h_n z_n}{\mu}\right) - \tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + w_j z_j}{\mu}\right)\right)\right]. \end{aligned} \tag{52}$$

Combining (51) and (52), we have

$$\begin{aligned} &Q_j - Q_{i_0} \\ &= (1-\theta)\theta^2 \mathbb{E}_{\bar{\boldsymbol{\Omega}}\setminus\{i_0,j\}}\Bigg[\underbrace{\mathbb{E}_{\boldsymbol{z}}\left(\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + h_n z_n + w_{i_0} z_{i_0}}{\mu}\right) - \tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + h_n z_n + w_j z_j}{\mu}\right)\right)}_{I_3}\Bigg] \\ &+ (1-\theta)^2\theta \mathbb{E}_{\bar{\boldsymbol{\Omega}}\setminus\{i_0,j\}}\Bigg[\underbrace{\mathbb{E}_{\boldsymbol{z}}\left(\tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + w_{i_0} z_{i_0}}{\mu}\right) - \tanh^2\left(\frac{\boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + w_j z_j}{\mu}\right)\right)}_{I_4}\Bigg]. \end{aligned} \tag{53}$$

To show that $I_3 \geq 0$, let $X_1 := \boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + h_n z_n + w_{i_0} z_{i_0} | \bar{\boldsymbol{\Omega}} \sim \mathcal{N}(0, \left\|\boldsymbol{w}_{\mathcal{J}\setminus\{i_0,j\}}\right\|_2^2 + h_n^2 + w_{i_0}^2 =: \sigma_{X_1}^2)$ and $Y_1 := \boldsymbol{w}_{\setminus\{i_0,j\}}^\top \bar{\boldsymbol{x}}_{\setminus\{i_0,j\}} + h_n z_n + w_j z_j | \bar{\boldsymbol{\Omega}} \sim \mathcal{N}(0, \left\|\boldsymbol{w}_{\mathcal{J}\setminus\{i_0,j\}}\right\|_2^2 + h_n^2 + w_j^2 := \sigma_{Y_1}^2)$. Plugging X_1, Y_1 into the term I_3, we have

$$I_3 = \mathbb{E}_{\boldsymbol{z}} \left(\tanh^2\left(\frac{X_1}{\mu}\right) - \tanh^2\left(\frac{Y_1}{\mu}\right) \right) \geq 0 \tag{54}$$

conditioned on any support $\bar{\boldsymbol{\Omega}} \setminus \{i_0, j\}$, since $\sigma_{X_1}^2 \geq \sigma_{Y_1}^2$ and the function $\tanh^2(x)$ is monotonically increasing with respect to $|x|$. Similarly, we have $I_4 \geq 0$ as well. In view of $I_3, I_4 \geq 0$ and (53), we have (43).

B.2 Proof of Proposition 1

The directional gradient can be written as a sum of p i.i.d. random variables as following:

$$\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} := \frac{1}{p} \sum_{i=1}^p X_i, \quad \text{where} \quad X_i = \frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i) \boldsymbol{h}(\boldsymbol{w})\right)}{\|\boldsymbol{w}\|_2}.$$

In order to apply the Bernstein's inequality in Lemma 7, we turn to bound the moments of X_i. Plugging in (34), we have

$$\begin{aligned} X_i &= \frac{\boldsymbol{w}^\top \boldsymbol{J_h}(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) \\ &= \left[\frac{\boldsymbol{w}^\top}{\|\boldsymbol{w}\|_2} \quad -\frac{\|\boldsymbol{w}\|_2}{h_n(\boldsymbol{w})} \right] \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) \leq \sqrt{2}n \left\|\mathcal{C}(\boldsymbol{x}_i)\right\|, \end{aligned} \tag{55}$$

where the last inequality follows from $|\tanh(\cdot)| \leq 1$ and $\left\| \left[\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \quad -\frac{\|\boldsymbol{w}\|_2}{h_n(\boldsymbol{w})} \right] \right\|_2 = \sqrt{1 + \frac{\|\boldsymbol{w}\|_2^2}{h_n^2}} \leq \sqrt{1+n} \leq \sqrt{2n}$, since $\|\boldsymbol{w}\|_2^2 \leq \frac{n-1}{n}$ and $h_n = \sqrt{1 - \|\boldsymbol{w}\|_2^2} \geq \frac{1}{\sqrt{n}}$. Invoking Lemma 10, we have for any $m \geq 2$,

$$\mathbb{E} \left|X_i\right|^m \leq (\sqrt{2}n)^m \mathbb{E} \left\|\mathcal{C}(\boldsymbol{x}_i)\right\|^m \leq \frac{m!}{2} \cdot \left(Cn^3 \log n\right)^{m/2} \tag{56}$$

for some constant C. Finally, using (39a), we complete the proof by setting $\sigma^2 = Cn^3 \log n$, $R = \sqrt{Cn^3 \log n}$ and applying the Bernstein's inequality in Lemma 7.

B.3 Proof of Proposition 2

The Hessian of $\phi_o(\boldsymbol{w})$ can be written as a sum of p i.i.d. random matrices as following:

$$\nabla_{\boldsymbol{w}}^2 \phi_o(\boldsymbol{w}) := \frac{1}{p} \sum_{i=1}^p \boldsymbol{Y}_i, \quad \text{where,} \quad \boldsymbol{Y}_i = \nabla_{\boldsymbol{w}}^2 \psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})\right).$$

Plugging in (35), we divide $\boldsymbol{Y}_i$ into two parts as:

$$\begin{aligned} \boldsymbol{Y}_i = &\underbrace{\frac{1}{\mu} \boldsymbol{J_h}(\boldsymbol{w}) \mathcal{C}(\boldsymbol{x}_i)^\top \left[\mathbf{I} - \operatorname{diag}\left(\tanh^2\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right)\right] \mathcal{C}(\boldsymbol{x}_i) \boldsymbol{J_h}(\boldsymbol{w})^\top}_{\boldsymbol{D}_i} \\ &\underbrace{- \frac{1}{h_n} \mathcal{S}_{n-1}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) \boldsymbol{J_h}(\boldsymbol{w}) \boldsymbol{J_h}(\boldsymbol{w})^\top}_{\boldsymbol{E}_i}. \end{aligned}$$

Therefore, we bound the sums of $\boldsymbol{D}_i$ and $\boldsymbol{E}_i$ respectively, using the Bernstein's inequality in Lemma 7.

Bound the concentration of $\boldsymbol{E}_i$: We start by bounding the moments of $\boldsymbol{E}_i$. Recalling the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})$ in (33), we have

$$\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})^{\top} = \mathbf{I} + \frac{\boldsymbol{w}\boldsymbol{w}^{\top}}{h_n^2},$$

and therefore, $\left\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})^{\top}\right\| = 1 + \|\boldsymbol{w}\|_2^2/h_n^2 \le 5$ since for $\|\boldsymbol{w}\|_2 \le 1/2$, we have $h_n(\boldsymbol{w}) \ge 1/2$. Consequently, by the triangle inequality,

$$\|\boldsymbol{E}_i\| \le \frac{1}{h_n}\|\boldsymbol{x}_i\|_2 \left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right\|_2 \left\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})^{\top}\right\| \le 10\sqrt{n}\|\boldsymbol{x}_i\|_2.$$

We can bound the moments of $\boldsymbol{E}_i$ as

$$\mathbb{E}\|\boldsymbol{E}_i\|^m \le 10^m n^{m/2}\mathbb{E}\|\boldsymbol{x}_i\|_2^m \le 10^m n^{m/2} \cdot m! n^{m/2} \le \frac{m!}{2}(20n)^2 \cdot (20n)^{m-2},$$

where the second line follows from Fact 2 and Fact 3 that bound the moments of $\|\boldsymbol{x}_i\|_2$.

Setting $\sigma^2 = 400n^2$, $R = 20n$, we apply the Bernstein's inequality in Lemma 7 and obtain:

$$\mathbb{P}\left(\left\|\frac{1}{p}\sum_{i=1}^{p}\boldsymbol{E}_i - \mathbb{E}\left(\frac{1}{p}\sum_{i=1}^{p}\boldsymbol{E}_i\right)\right\| \ge \frac{t}{2}\right) \le 2n\exp\left(\frac{-pt^2}{c_1 n^2 + c_2 nt}\right) \tag{57}$$

for some large enough constants c_1 and c_2.

Bound the concentration of $\boldsymbol{D}_i$: Using the fact that $1-\tanh^2(\cdot) \le 1$, the spectral norm of $\boldsymbol{D}_i$ can be bounded as

$$\|\boldsymbol{D}_i\| \le \frac{1}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\|^2 \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|^2 \le \frac{5}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\|^2,$$

where we have used again $\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|^2 = \left\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})^{\top}\right\| \le 5$ derived above. Invoking Lemma 10, we obtain

$$\mathbb{E}\|\boldsymbol{D}_i\|^m \le \left(\frac{5}{\mu}\right)^m \mathbb{E}\|\mathcal{C}(\boldsymbol{x}_i)\|^{2m} \le \frac{m!}{2}\left(\frac{Cn\log n}{\mu}\right)^m, \tag{58}$$

for some constant C. Let $\sigma^2 = \frac{C^2n^2\log^2 n}{\mu^2}$, $R = \frac{Cn\log n}{\mu}$, by the Bernstein's inequality in Lemma 7, we have:

$$\mathbb{P}\left(\left\|\frac{1}{p}\sum_{i=1}^{p}\boldsymbol{D}_i - \mathbb{E}\left(\frac{1}{p}\sum_{i=1}^{p}\boldsymbol{D}_i\right)\right\| \ge \frac{t}{2}\right) \le 2n\exp\left(\frac{-p\mu^2t^2}{c_3n^2\log^2 n + c_4\mu nt\log n}\right). \tag{59}$$

for some constants c_3 and c_4.

Recall the Hessian of interest is written as:

$$\nabla^2_{\boldsymbol{w}}\phi_o(\boldsymbol{w}) = \frac{1}{p}\sum_{i=1}^{p}\boldsymbol{Y}_i = \frac{1}{p}\sum_{i=1}^{p}\boldsymbol{D}_i - \frac{1}{p}\sum_{i=1}^{p}\boldsymbol{E}_i. \tag{60}$$

Combining the bounds for $\boldsymbol{D}_i$ (cf. (59)) and $\boldsymbol{E}_i$ (cf. (57)), and observing $\nabla^2_{\boldsymbol{w}}\mathbb{E}\phi_o(\boldsymbol{w}) = \mathbb{E}\nabla^2_{\boldsymbol{w}}\phi_o(\boldsymbol{w})$ from (39b), we obtain the final bound as advertised:

$$\mathbb{P}\left(\left\|\nabla^2_{\boldsymbol{w}}\phi_o(\boldsymbol{w}) - \nabla^2_{\boldsymbol{w}}\mathbb{E}\phi_o(\boldsymbol{w})\right\| \ge t\right) \le 4n\exp\left(\frac{-p\mu^2t^2}{c_5n^2\log^2 n + c_6\mu n\log(n)t}\right).$$

B.4 Proof of Theorem 2

We start by introducing the event

$$\mathcal{A}_0 := \left\{\|\boldsymbol{X}\|_\infty \le 4\sqrt{\log(np)}\right\},$$

which holds with probability at least $1-\theta(np)^{-7}$ by Fact 8.

B.4.1 Proof of (16a)

To show that $\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2}$ is lower bounded uniformly in the region $\mathcal{Q}_1$, we will apply a standard covering argument. Let $\mathcal{N}_1$ be an ϵ-net of $\mathcal{Q}_1$, such that for any $\boldsymbol{w} \in \mathcal{Q}_1$, there exists $\boldsymbol{w}_1 \in \mathcal{N}_1$ with $\|\boldsymbol{w} - \boldsymbol{w}_1\|_2 \le \epsilon$. By standard results [71, Lemma 5.7], the size of $\mathcal{N}_1$ is at most $\lceil 3/\epsilon \rceil^n$, where the value of ϵ will be determined later. We have

$$\begin{aligned}\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} &= \left[\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} - \frac{\boldsymbol{w}_1^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2}\right] + \left[\frac{\boldsymbol{w}_1^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2} - \frac{\boldsymbol{w}_1^\top \mathbb{E}\nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2}\right] + \frac{\boldsymbol{w}_1^\top \mathbb{E}\nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2} \\ &\ge \underbrace{\frac{\boldsymbol{w}_1^\top \mathbb{E}\nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2}}_{\mathrm{I}} - \underbrace{\left|\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} - \frac{\boldsymbol{w}_1^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2}\right|}_{\mathrm{II}} - \underbrace{\left|\frac{\boldsymbol{w}_1^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2} - \frac{\boldsymbol{w}_1^\top \mathbb{E}\nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2}\right|}_{\mathrm{III}}.\end{aligned}$$

In the sequel, we derive bounds for the terms I, II, III respectively.

- For term I, as $\boldsymbol{w}_1 \in \mathcal{N}_1 \subseteq \mathcal{Q}_1$, by Lemma 1, we have

$$\mathrm{I} = \frac{\boldsymbol{w}_1^\top \mathbb{E}\nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2} \ge \frac{\theta}{480\sqrt{10\pi}}\xi_0 := c_1\theta\xi_0.$$

- To bound term II, by the additivity of Lipschitz constants and [16, Proposition 13], we have $\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2}$ is L_1-Lipschitz with

$$L_1 \le \left(\frac{8\sqrt{2}n^{3/2}}{\mu} + 8n^{5/2}\right)\|\boldsymbol{X}\|_\infty + \frac{4n^3}{\mu}\|\boldsymbol{X}\|_\infty^2.$$

  Therefore, under the event $\mathcal{A}_0$, we have $L_1 \le \frac{c_2 n^3}{\mu}\log(np)$ for some constant c_2. Setting $\epsilon = \frac{c_1\theta\xi_0}{3L_1}$, we obtain that

$$\mathrm{II} = \left|\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} - \frac{\boldsymbol{w}_1^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2}\right| \le L_1\|\boldsymbol{w} - \boldsymbol{w}_1\|_2 \le L_1\epsilon \le \frac{c_1\theta\xi_0}{3}.$$

  Along the way, we determine the size of $\mathcal{N}_1$ is upper bounded by

$$|\mathcal{N}_1| \le \lceil 3/\epsilon \rceil^n \le \exp\left\{n\log\left(\frac{c_3 n^3\log(np)}{\mu\theta\xi_0}\right)\right\}.$$

- For term III, by setting $t = \frac{c_1\theta\xi_0}{3}$ in Proposition 1 and the union bound, we have the event

$$\mathcal{A}_1 := \left\{\max_{\boldsymbol{w}_1 \in \mathcal{N}_1}\left|\frac{\boldsymbol{w}_1^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2} - \frac{\boldsymbol{w}_1^\top \mathbb{E}\nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_1)}{\|\boldsymbol{w}_1\|_2}\right| \le \frac{c_1\theta\xi_0}{3}\right\}$$

  holds with probability at least

$$\begin{aligned}1 - |\mathcal{N}_1|\cdot 2\exp\left(\frac{-pt^2}{C_1 n^3\log n + C_2\sqrt{n^3\log(n)}t}\right) &\ge 1 - 2\exp\left(\frac{-c_4 p\theta^2\xi_0^2}{n^3\log n} + n\log\left(\frac{c_3 n^3\log(np)}{\mu\theta\xi_0}\right)\right) \\ &\ge 1 - 2\exp(-c_5 n),\end{aligned}$$

  provided $p \ge \frac{Cn^4}{\theta^2\xi_0^2}\log n\log\left(\frac{n^3\log p}{\mu\theta\xi_0}\right)$ for some sufficiently large C and n is sufficiently large.

Combining terms, conditioned on $\mathcal{A}_0 \bigcap \mathcal{A}_1$, which holds with probability at least $1 - \theta(np)^{-7} - 2\exp(-c_5 n)$, we have that for all $\boldsymbol{w} \in \mathcal{Q}_1$, (16a) holds since,

$$\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} \ge \mathrm{I} - \mathrm{II} - \mathrm{III} \ge -\frac{c_1\theta\xi_0}{3} - \frac{c_1\theta\xi_0}{3} + c_1\theta\xi_0 = \frac{c_1\theta\xi_0}{3}.$$

B.4.2 Proof of (16b)

The proof is similar to the above proof of (16a) in Appendix B.4.1. Let $\mathcal{N}_2$ be an ϵ-net of $\mathcal{Q}_2$, such that for any $\boldsymbol{w} \in \mathcal{Q}_2$, there exists $\boldsymbol{w}_2 \in \mathcal{N}_2$ with $\|\boldsymbol{w} - \boldsymbol{w}_2\|_2 \le \epsilon$. By standard results [71, Lemma 5.7], the size of $\mathcal{N}_2$ is at most $\lceil 3\mu/(4\sqrt{2}\epsilon)\rceil^n$, where the value of ϵ will be determined later. By the triangle inequality, we have for all $\boldsymbol{w} \in \mathcal{Q}_2$,

$$\nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}) \succeq \underbrace{\inf_{\boldsymbol{w}_2 \in \mathcal{N}_2} \nabla^2_{\boldsymbol{w}} \mathbb{E}\left(\phi_o(\boldsymbol{w}_2)\right)}_{\boldsymbol{H}_1} - \underbrace{\left\|\nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_2) - \nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w})\right\| \mathbf{I}}_{\boldsymbol{H}_2} - \underbrace{\left\|\nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_2) - \nabla^2_{\boldsymbol{w}} \mathbb{E}\left(\phi_o(\boldsymbol{w}_2)\right)\right\| \mathbf{I}}_{\boldsymbol{H}_3}.$$

In the sequel, we derive bounds for the terms $\boldsymbol{H}_1$, $\boldsymbol{H}_2$, $\boldsymbol{H}_3$ respectively.

- For $\boldsymbol{H}_1$, by Theorem 1, we have

$$\boldsymbol{H}_1 = \inf_{\boldsymbol{w}_2 \in \mathcal{N}_2} \nabla^2_{\boldsymbol{w}} \mathbb{E}\left(\phi_o(\boldsymbol{w}_2)\right) \succeq \frac{n\theta}{5\sqrt{2\pi}\mu}\mathbf{I} := \frac{c_5 n\theta}{\mu}\mathbf{I}.$$

- To bound $\boldsymbol{H}_2$, by the additivity of Lipschitz constants and [16, Proposition 14], we have $\nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w})$ is L_2-Lipschitz with

$$L_2 \le \frac{4n^3}{\mu^2} \|\boldsymbol{X}\|_\infty^3 + \left(\frac{4n^2}{\mu} + \frac{8\sqrt{2}n^{3/2}}{\mu}\right) \|\boldsymbol{X}\|_\infty^2 + 8n \|\boldsymbol{X}\|_\infty .$$

Under the event $\mathcal{A}_0$, we have $L_2 \le \frac{c_6 n^3}{\mu^2} \log^{3/2}(np)$ for some constant c_6. Setting $\epsilon = \frac{c_5 n\theta}{3\mu L_2}$, we obtain

$$\left\|\nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_2) - \nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w})\right\| \le \frac{c_5 n\theta}{3\mu}, \quad \text{and} \quad \boldsymbol{H}_2 \preceq \frac{c_5 n\theta}{3\mu}\mathbf{I}.$$

Along the way, we determine the size of $\mathcal{N}_2$ is upper bounded by

$$|\mathcal{N}_2| \le \lceil 3\mu/(4\sqrt{2}\epsilon)\rceil^n \le \exp\left[n \log\left(\frac{c_7 n^2 \log^{3/2}(np)}{\theta}\right)\right].$$

- To bound $\boldsymbol{H}_3$, by setting $t = \frac{c_5 n\theta}{3\mu}$ in Proposition 2 (as $\|\boldsymbol{w}\|_2 \le 1/2$ for $\boldsymbol{w} \in \mathcal{Q}_2$ when $\mu < 1$) and the union bound, we have $\boldsymbol{H}_3 \preceq \frac{c_5 n\theta}{3\mu}\mathbf{I}$ under the event

$$\mathcal{A}_2 := \left\{\max_{\boldsymbol{w}_2 \in \mathcal{N}_2} \left\|\nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}_2) - \nabla^2_{\boldsymbol{w}} \mathbb{E}\left(\phi_o(\boldsymbol{w}_2)\right)\right\| \le \frac{c_5 n\theta}{3\mu}\right\}$$

holds with probability at least

$$\begin{aligned} 1 - |\mathcal{N}_2| \cdot 4n \exp\left(\frac{-p\mu^2 t^2}{C_3 n^2 \log^2 n + C_4 \mu t n \log n}\right) &\ge 1 - 4n \exp\left(-\frac{c_8 p\theta^2}{\log^2 n} + n \log\left(\frac{c_7 n^2 \log^{3/2}(np)}{\theta}\right)\right) \\ &\ge 1 - \exp(-c_9 n), \end{aligned}$$

provided $p \ge \frac{Cn}{\theta^2} \log^2 n \log\left(\frac{n^2 \log^{3/2} p}{\theta}\right)$ for some sufficiently large C and n is sufficiently large.

Combining terms, conditioned on $\mathcal{A}_0 \bigcap \mathcal{A}_2$, which holds with probability at least $1 - \theta(np)^{-7} - \exp(-c_9 n)$, we have (16b) holds since,

$$\nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}) \succeq \boldsymbol{H}_1 - \boldsymbol{H}_2 - \boldsymbol{H}_3 \succeq \frac{c_5 n\theta}{3\mu}\mathbf{I}.$$

B.4.3 Proof of (17)

The characterized geometry of $\phi_o(\boldsymbol{w})$ implies that it has at most one local minimum in $\mathcal{Q}_2$ due to strong convexity, which is denoted as $\boldsymbol{w}_o^\star$. We are going to show that $\boldsymbol{w}_o^\star$ is close to $\mathbf{0}$ in $\mathcal{Q}_2$. By the optimality of $\boldsymbol{w}_o^\star$ and the mean value theorem, we have for some $t \in (0,1)$:

$$\begin{aligned}\phi_o(\mathbf{0}) \geq \phi_o(\boldsymbol{w}_o^\star) &\geq \phi_o(\mathbf{0}) + \langle \nabla_{\boldsymbol{w}}\phi_o(\mathbf{0}), \boldsymbol{w}_o^\star\rangle + \boldsymbol{w}_o^{\star\top}\nabla^2\phi_o(t\boldsymbol{w}_o^\star)\boldsymbol{w}_o^\star \\ &\geq \phi_o(\mathbf{0}) - \|\boldsymbol{w}_o^\star\|_2 \|\nabla_{\boldsymbol{w}}\phi_o(\mathbf{0})\|_2 + \frac{c_5 n\theta}{2\mu}\|\boldsymbol{w}_o^\star\|_2^2,\end{aligned}$$

where the second line follows from (16b) and the Cauchy-Schwartz inequality. Therefore, we have

$$\|\boldsymbol{w}_o^\star\|_2 \leq \frac{2\mu}{c_5 n\theta}\|\nabla_{\boldsymbol{w}}\phi_o(\mathbf{0})\|_2. \tag{61}$$

It remains to bound $\|\nabla_{\boldsymbol{w}}\phi_o(\mathbf{0})\|_2$, which we resort to the Bernstein's inequality in Lemma 7. As $\nabla_{\boldsymbol{w}}\phi_o(\mathbf{0}) = \frac{1}{p}\sum_{i=1}^{p}\nabla_{\boldsymbol{w}}\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\mathbf{0})\right)$, where it is straightforward to check $\mathbb{E}\nabla_{\boldsymbol{w}}\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\mathbf{0})\right) = \mathbf{0}$ due to symmetry. We turn to bound the moments of $\|\nabla_{\boldsymbol{w}}\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\mathbf{0})\right)\|_2$ as follows,

$$\begin{aligned}\|\nabla_{\boldsymbol{w}}\psi_\mu\left(\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\mathbf{0})\right)\|_2 &= \left\|\boldsymbol{J}_{\boldsymbol{h}}(\mathbf{0})\mathcal{C}(\boldsymbol{x}_i)^\top\tanh\left(\frac{\mathcal{C}(\boldsymbol{x})\boldsymbol{h}(\mathbf{0})}{\mu}\right)\right\|_2 \\ &\leq \|\boldsymbol{J}_{\boldsymbol{h}}(\mathbf{0})\|\,\|\mathcal{C}(\boldsymbol{x}_i)\|\left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x})\boldsymbol{h}(\mathbf{0})}{\mu}\right)\right\|_2 \\ &\leq \sqrt{n}\,\|\mathcal{C}(\boldsymbol{x}_i)\|,\end{aligned}$$

where the last inequality follows from $\|\boldsymbol{J}_{\boldsymbol{h}}(\mathbf{0})\| = \left\|\begin{bmatrix}\mathbf{I}_{n-1} & \mathbf{0}\end{bmatrix}\right\| = 1$ and $|\tanh(\cdot)| \leq 1$. Invoking Lemma 10, we have for all $m \geq 2$,

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{w}}\psi_\mu\left(\mathcal{C}(\boldsymbol{x})\boldsymbol{h}(\mathbf{0})\right)\|_2^m\right] \leq \left(\sqrt{n}\right)^m\mathbb{E}\|\mathcal{C}(\boldsymbol{x}_i)\|^m \leq \frac{m!}{2}\cdot\left(Cn^2\log(n)\right)^{m/2}$$

for some constant C. Setting $\sigma^2 = Cn^2\log(n)$, $R = \sqrt{Cn^2\log(n)}$ in the Bernstein's inequality in Lemma 7, we have

$$\mathbb{P}\left(\|\nabla_{\boldsymbol{w}}\phi_o(\mathbf{0})\|_2 \geq t\right) \leq 2(n+1)\exp\left(\frac{-pt^2}{2Cn^2\log(n) + 2\sqrt{Cn^2\log(n)}t}\right).$$

Let $t = c_9\sqrt{\frac{n^2\log(n)\log(p)}{p}}$, we have

$$\|\nabla_{\boldsymbol{w}}\phi_o(\mathbf{0})\|_2 \leq c_9\sqrt{\frac{n^2\log n\log p}{p}} \tag{62}$$

with probability at least $1 - 4np^{-7}$ when $p \geq c_{10}n\log(n)$. Under the sample size requirement on p, we have

$$\|\boldsymbol{w}_o^\star - \mathbf{0}\|_2 \leq \frac{c_6\mu}{\theta}\sqrt{\frac{\log n\log p}{p}} \leq \frac{\mu}{10},$$

for some constant c_6, which ensures $\boldsymbol{w}_o^\star \in \mathcal{Q}_2$.

C Proofs for Section 3.2

C.1 Proof of Lemma 2

Recalling $\boldsymbol{\Delta} = (\boldsymbol{U}' - \boldsymbol{U})\boldsymbol{U}^{-1}$, we have

$$\|\boldsymbol{\Delta}\| = \left\|(\boldsymbol{U}' - \boldsymbol{U})\boldsymbol{U}^{-1}\right\| = \|\boldsymbol{U}' - \boldsymbol{U}\|, \tag{63}$$

since $\boldsymbol{U}$ is an orthonormal matrix, i.e., $\left\|\boldsymbol{U}^{-1}\right\| = 1$. Therefore, it is sufficient to bound $\|\boldsymbol{U}' - \boldsymbol{U}\|$ instead. Plugging in the definition of $\boldsymbol{U}'$ and $\boldsymbol{U}$, we have

$$\begin{aligned}
\|\boldsymbol{U}' - \boldsymbol{U}\| &= \left\|\mathcal{C}(\boldsymbol{g})\boldsymbol{R} - \mathcal{C}(\boldsymbol{g})\left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\right)^{-1/2}\right\| \\
&\leq \|\mathcal{C}(\boldsymbol{g})\| \left\|\boldsymbol{R} - \left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\right)^{-1/2}\right\| \\
&\leq \|\mathcal{C}(\boldsymbol{g})\| \frac{\left\|\boldsymbol{R}^2 - \left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\right)^{-1}\right\|}{\sigma_{\min}\left(\left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\right)^{-1/2}\right)} \\
&\leq \|\mathcal{C}(\boldsymbol{g})\|^2 \left\| \left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\right)^{-1} \right\| \left\|\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\boldsymbol{R}^2 - \mathbf{I}\right\| \\
&= \kappa^2 \left\|\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\boldsymbol{R}^2 - \mathbf{I}\right\|, \qquad (64)
\end{aligned}$$

where the second inequality follows from the fact [72, Theorem 6.2] that for two positive matrices $\boldsymbol{U}, \boldsymbol{V}$, we have $\left\|\boldsymbol{U}^{-1/2} - \boldsymbol{V}^{-1/2}\right\| \leq \frac{\left\|\boldsymbol{U}^{-1} - \boldsymbol{V}^{-1}\right\|}{\sigma_{\min}(\boldsymbol{V}^{-1/2})}$. We continue by plugging in the definition of $\boldsymbol{R}$,

$$\begin{aligned}
\left\|\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\boldsymbol{R}^2 - \mathbf{I}\right\| &= \left\|\left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\right)\cdot\left(\frac{1}{\theta np}\sum_{i=1}^{p}\left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{x}_i)^\top\mathcal{C}(\boldsymbol{x}_i)\mathcal{C}(\boldsymbol{g})\right)\right)^{-1} - \mathbf{I}\right\| \\
&= \left\|\left[\mathbf{I} + \left(\mathcal{C}(\boldsymbol{g})^\top\left[\frac{1}{\theta np}\sum_{i=1}^{p}\mathcal{C}(\boldsymbol{x}_i)^\top\mathcal{C}(\boldsymbol{x}_i) - \mathbf{I}\right]\mathcal{C}(\boldsymbol{g})\right)\cdot\left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\right)^{-1}\right]^{-1} - \mathbf{I}\right\| \\
&:= \left\|(\mathbf{I} + \boldsymbol{A})^{-1} - \mathbf{I}\right\|. \qquad (65)
\end{aligned}$$

where $\boldsymbol{A} = \left(\mathcal{C}(\boldsymbol{g})^\top\left[\frac{1}{\theta np}\sum_{i=1}^{p}\mathcal{C}(\boldsymbol{x}_i)^\top\mathcal{C}(\boldsymbol{x}_i) - \mathbf{I}\right]\mathcal{C}(\boldsymbol{g})\right)\cdot\left(\mathcal{C}(\boldsymbol{g})^\top\mathcal{C}(\boldsymbol{g})\right)^{-1}$.

By Lemma 11, we have when $p \geq Cn\log(n)$,

$$\left\|\frac{1}{\theta np}\sum_{i=1}^{p}\mathcal{C}(\boldsymbol{x}_i)^\top\mathcal{C}(\boldsymbol{x}_i) - \mathbf{I}\right\| \leq C\sqrt{\frac{\log^2 n \log p}{\theta^2 p}}$$

with probability at least $1 - 2np^{-8}$, and $\|\boldsymbol{A}\| \leq C\kappa^2\sqrt{\frac{\log^2(n)\log(p)}{\theta^2 p}}$. Then as long as $\|\boldsymbol{A}\| \leq 1/2$, which holds when $p \geq \frac{C_2\kappa^4\log^2(n)\log p}{\theta^2}$ for some large enough constant C_2, we have

$$\left\|(\mathbf{I} + \boldsymbol{A})^{-1} - \mathbf{I}\right\| \leq \left\|(\mathbf{I} + \boldsymbol{A})^{-1}\right\|\|\boldsymbol{A}\| \leq \frac{\|\boldsymbol{A}\|}{1 - \|\boldsymbol{A}\|} \leq 2\|\boldsymbol{A}\|.$$

Plugging this back into (64), we have

$$\|\boldsymbol{U}' - \boldsymbol{U}\| \leq C_3\kappa^4\sqrt{\frac{\log^2 n\log p}{\theta^2 p}}.$$

C.2 Proof of Lemma 3

We first record some useful facts. For any $\boldsymbol{h} \in \mathcal{S}_0^{(n+)}$, we have the Jacobian matrix $\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}) = \left[\mathbf{I}, -\frac{\boldsymbol{w}}{h_n}\right] \in \mathbb{R}^{(n-1)\times n}$ satisfies

$$\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\| \leq \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|_{\mathrm{F}} \leq \sqrt{n - 1 + \frac{\|\boldsymbol{w}\|_2^2}{h_n^2}} \leq \sqrt{2n}, \qquad (66)$$

since $\|\boldsymbol{w}\|_2 \leq 1$ and $h_n \geq \frac{1}{\sqrt{n}}$. In addition, by the union bound and Lemma 10, we have with probability at least $1 - (np)^{-8}$,

$$\max_{i\in[p]}\|\mathcal{C}(\boldsymbol{x}_i)\| \leq C\sqrt{n\log(np)}, \qquad (67)$$

for some constant C.

C.2.1 Proof of (29a)

Similar to (34), we can write the gradient $\nabla_{\boldsymbol{w}}\phi(\boldsymbol{w})$ as

$$\nabla_{\boldsymbol{w}}\phi(\boldsymbol{w}) = \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\left(\mathbf{I}+\boldsymbol{\Delta}\right)^\top \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right).$$

Recalling the expression of $\nabla_{\boldsymbol{w}}\phi_o(\boldsymbol{w})$ in (34), we write

$$\begin{aligned}
&\nabla_{\boldsymbol{w}}\phi_o(\boldsymbol{w}) - \nabla_{\boldsymbol{w}}\phi(\boldsymbol{w}) \\
&= \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) - \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\left(\mathbf{I}+\boldsymbol{\Delta}\right)^\top \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) \\
&= \underbrace{\frac{1}{p}\sum_{i=1}^{p} \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\mathcal{C}(\boldsymbol{x}_i)^\top \left[\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right]}_{\boldsymbol{g}_1} \\
&\qquad \underbrace{- \frac{1}{p}\sum_{i=1}^{p} \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\boldsymbol{\Delta}^\top \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)}_{\boldsymbol{g}_2}.
\end{aligned}$$

Therefore, we continue to bound $\|\boldsymbol{g}_1\|_2$ and $\|\boldsymbol{g}_2\|_2$.

- To bound $\|\boldsymbol{g}_1\|_2$, we have

$$\begin{aligned}
\|\boldsymbol{g}_1\|_2 &\le \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\| \cdot \max_{i\in[p]} \|\mathcal{C}(\boldsymbol{x}_i)\| \cdot \max_{i\in[p]} \left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right\|_2 \\
&\le \frac{1}{\mu}\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\| \cdot \max_{i\in[p]} \|\mathcal{C}(\boldsymbol{x}_i)\|^2 \cdot \|\boldsymbol{\Delta}\|. \qquad (68)
\end{aligned}$$

  Here, the second line follows from for any $i \in [p]$,

$$\begin{aligned}
&\left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right\|_2 \\
&\le \left\|\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right) - \left(\frac{\mathcal{C}(\boldsymbol{x}_i)(\mathbf{I}+\boldsymbol{\Delta})\boldsymbol{h}}{\mu}\right)\right\|_2 \\
&= \frac{1}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{\Delta}\boldsymbol{h}\|_2 \le \frac{1}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\|\,\|\boldsymbol{\Delta}\|\,\|\boldsymbol{h}\|_2 = \frac{1}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\|\,\|\boldsymbol{\Delta}\|, \qquad (69)
\end{aligned}$$

  where the second line follows from Lemma 9, and the last equality is due to $\|\boldsymbol{h}\|_2 = 1$.

- To bound $\|\boldsymbol{g}_2\|_2$, we have

$$\begin{aligned}
\|\boldsymbol{g}_2\|_2 &\le \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\| \cdot \max_{i\in[p]} \|\mathcal{C}(\boldsymbol{x}_i)\| \cdot \max_{i\in[p]} \left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right\|_2 \cdot \|\boldsymbol{\Delta}\| \\
&\le \sqrt{n}\,\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\| \cdot \max_{i\in[p]} \|\mathcal{C}(\boldsymbol{x}_i)\| \cdot \|\boldsymbol{\Delta}\|, \qquad (70)
\end{aligned}$$

  where the second line uses $|\tanh(\cdot)| \le 1$, and $\left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)(\mathbf{I}+\boldsymbol{\Delta})\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right\|_2 \le \sqrt{n}$.

Combining (68) and (70), we have

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}}\phi_o(\boldsymbol{w}) - \nabla_{\boldsymbol{w}}\phi(\boldsymbol{w})\|_2 &\le \|\boldsymbol{g}_1\|_2 + \|\boldsymbol{g}_2\|_2 \le \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\| \cdot \max_{i\in[p]} \|\mathcal{C}(\boldsymbol{x}_i)\| \cdot \|\boldsymbol{\Delta}\| \left(\sqrt{n} + \frac{1}{\mu}\max_{i\in[p]}\|\mathcal{C}(\boldsymbol{x}_i)\|\right) \\
&\le C\frac{n^{3/2}\log(np)}{\mu}\|\boldsymbol{\Delta}\|,
\end{aligned}$$

for some constant C, where the last line follows from (66) and (67), which holds with probability at least $1-(np)^{-8}$.

C.2.2 Proof of (29b)

First, under the sample size $p \geq \frac{C_2 \kappa^8 n \log^2(n) \log p}{\theta^2}$, from Lemma 2, we can ensure $\|\mathbf{\Delta}\| \leq 1$. Note that

$$\begin{aligned}\left\| \nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}) - \nabla^2_{\boldsymbol{w}} \phi(\boldsymbol{w}) \right\| &= \left\| \frac{1}{p} \sum_{i=1}^p \nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) \boldsymbol{h}(\boldsymbol{w})) - \frac{1}{p} \sum_{i=1}^p \nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})) \right\| \\ &\leq \frac{1}{p} \sum_{i=1}^p \left\| \nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) \boldsymbol{h}(\boldsymbol{w})) - \nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})) \right\|. \end{aligned} \tag{71}$$

Similar to (35), we can write the Hessian $\nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w}))$ as

$$\begin{aligned} &\nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})) \\ &= \frac{1}{\mu} \boldsymbol{J_h}(\boldsymbol{w}) (\mathbf{I} + \mathbf{\Delta}) \mathcal{C}(\boldsymbol{x}_i)^\top \left[\mathbf{I} - \operatorname{diag}\left(\tanh^2 \left(\frac{\mathcal{C}(\boldsymbol{x}) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \right) \right] \mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{J_h}(\boldsymbol{w})^\top \\ &\qquad - \frac{1}{h_n} \mathcal{S}_{n-1}(\boldsymbol{x})^\top (\mathbf{I} + \mathbf{\Delta}) \tanh \left(\frac{\mathcal{C}(\boldsymbol{x}) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \boldsymbol{J_h}(\boldsymbol{w}) \boldsymbol{J_h}(\boldsymbol{w})^\top. \end{aligned} \tag{72}$$

Subtracting $\nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) \boldsymbol{h}(\boldsymbol{w}))$ in (35) from the above equation, we have

$$\begin{aligned} &\nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) \boldsymbol{h}(\boldsymbol{w})) - \nabla^2_{\boldsymbol{w}} \psi_\mu(\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})) \\ &= \underbrace{\frac{1}{\mu} \boldsymbol{J_h}(\boldsymbol{w}) \mathcal{C}(\boldsymbol{x}_i)^\top \left[\operatorname{diag}\left(\tanh^2 \left(\frac{\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) - \tanh^2 \left(\frac{\mathcal{C}(\boldsymbol{x}_i) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \right) \right] \mathcal{C}(\boldsymbol{x}_i) \boldsymbol{J_h}(\boldsymbol{w})^\top}_{\boldsymbol{Q}_1} \\ &\quad \underbrace{- \frac{1}{\mu} \boldsymbol{J_h}(\boldsymbol{w}) \mathbf{\Delta} \mathcal{C}(\boldsymbol{x}_i)^\top \left[\mathbf{I} - \operatorname{diag}\left(\tanh^2 \left(\frac{\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \right) \right] \mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{J_h}(\boldsymbol{w})^\top}_{\boldsymbol{Q}_2} \\ &\quad \underbrace{- \frac{1}{\mu} \boldsymbol{J_h}(\boldsymbol{w}) \mathcal{C}(\boldsymbol{x}_i)^\top \left[\mathbf{I} - \operatorname{diag}\left(\tanh^2 \left(\frac{\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \right) \right] \mathcal{C}(\boldsymbol{x}_i) \mathbf{\Delta} \boldsymbol{J_h}(\boldsymbol{w})^\top}_{\boldsymbol{Q}_3} \\ &\quad \underbrace{+ \frac{1}{h_n} \mathcal{S}_{n-1}(\boldsymbol{x}_i)^\top \left[\tanh \left(\frac{\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) - \tanh \left(\frac{\mathcal{C}(\boldsymbol{x}_i) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \right] \boldsymbol{J_h}(\boldsymbol{w}) \boldsymbol{J_h}(\boldsymbol{w})^\top}_{\boldsymbol{Q}_4} \\ &\quad \underbrace{+ \frac{1}{h_n} \mathcal{S}_{n-1}(\boldsymbol{x}_i)^\top \mathbf{\Delta} \tanh \left(\frac{\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \boldsymbol{J_h}(\boldsymbol{w}) \boldsymbol{J_h}(\boldsymbol{w})^\top,}_{\boldsymbol{Q}_5} \end{aligned}$$

where in the sequel we'll bound these terms respectively.

- $\boldsymbol{Q}_1$ can be bounded as

$$\begin{aligned} \|\boldsymbol{Q}_1\| &\leq \frac{1}{\mu} \|\boldsymbol{J_h}(\boldsymbol{w})\|^2 \|\mathcal{C}(\boldsymbol{x}_i)\|^2 \left\| \tanh^2 \left(\frac{\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \mathbf{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) - \tanh^2 \left(\frac{\mathcal{C}(\boldsymbol{x}_i) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \right\|_\infty \\ &\leq \frac{2}{\mu^2} \|\boldsymbol{J_h}(\boldsymbol{w})\|^2 \|\mathcal{C}(\boldsymbol{x}_i)\|^2 \|\mathcal{C}(\boldsymbol{x}_i) \mathbf{\Delta} \boldsymbol{h}(\boldsymbol{w})\|_\infty \\ &\leq \frac{2}{\mu^2} \|\boldsymbol{J_h}(\boldsymbol{w})\|^2 \|\mathcal{C}(\boldsymbol{x}_i)\|^2 \|\boldsymbol{x}_i\|_2 \|\mathbf{\Delta}\|, \end{aligned}$$

  where the second line follows from Lemma 9, where the last line uses $\|\boldsymbol{h}\|_2 = 1$.

- $\boldsymbol{Q}_2$ can be bounded as

$$\|\boldsymbol{Q}_2\| \leq \frac{1}{\mu} \|\boldsymbol{J_h}(\boldsymbol{w})\|^2 \|\mathcal{C}(\boldsymbol{x}_i)\|^2 \|\mathbf{\Delta}\| (1 + \|\mathbf{\Delta}\|) \leq \frac{2}{\mu} \|\boldsymbol{J_h}(\boldsymbol{w})\|^2 \|\mathcal{C}(\boldsymbol{x}_i)\|^2 \|\mathbf{\Delta}\|,$$

  where we have used $1 - \tanh^2(\cdot) \leq 1$, $\|\mathbf{\Delta}\| \leq 1$ respectively.

- Similar to $\boldsymbol{Q}_2$, $\boldsymbol{Q}_3$ can be bounded as

$$\|\boldsymbol{Q}_3\| \leq \frac{1}{\mu} \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|^2 \|\mathcal{C}(\boldsymbol{x}_i)\|^2 \|\boldsymbol{\Delta}\| .$$

- $\boldsymbol{Q}_4$ can be bounded as

$$\begin{aligned}\|\boldsymbol{Q}_4\| &\leq \frac{1}{h_n} \|\boldsymbol{x}_i\|_2 \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|^2 \left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right\|_2 \\ &\leq \frac{\sqrt{n}}{\mu} \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|^2 \|\mathcal{C}(\boldsymbol{x}_i)\| \|\boldsymbol{x}_i\|_2 \|\boldsymbol{\Delta}\| ,\end{aligned}$$

  where the second line follows from (69) and $h_n \geq 1/\sqrt{n}$.

- $\boldsymbol{Q}_5$ can be bounded as

$$\begin{aligned}\|\boldsymbol{Q}_5\| &\leq \frac{1}{h_n} \|\boldsymbol{x}_i\|_2 \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|^2 \|\boldsymbol{\Delta}\| \left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})}{\mu}\right)\right\|_2 \\ &\leq n \|\boldsymbol{x}_i\|_2 \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|^2 \|\boldsymbol{\Delta}\| ,\end{aligned}$$

  where the second line uses $|\tanh(\cdot)| \leq 1$ and $h_n \geq 1/\sqrt{n}$.

Combining the above bounds back into (71), we have

$$\begin{aligned}&\left\|\nabla^2_{\boldsymbol{w}}\phi_o(\boldsymbol{w}) - \nabla^2_{\boldsymbol{w}}\phi(\boldsymbol{w})\right\| \\ &\leq \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\|^2 \|\boldsymbol{\Delta}\| \max_{i\in[p]} \left(\frac{2}{\mu^2}\|\mathcal{C}(\boldsymbol{x}_i)\|^2 \|\boldsymbol{x}_i\|_2 + \frac{3}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\|^2 + \frac{\sqrt{n}}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\| \|\boldsymbol{x}_i\|_2 + n\|\boldsymbol{x}_i\|_2\right).\end{aligned}$$

Plugging in (66), (67), and Fact 7, where with probability at least $1-2p^{-8}$,

$$\max_{i\in[p]} \|\mathcal{C}(\boldsymbol{x}_i)\| \leq C\sqrt{n\log(np)}, \quad \max_{i\in[p]} \|\boldsymbol{x}_i\|_2 \leq C\sqrt{n\log p},$$

we have

$$\left\|\nabla^2_{\boldsymbol{w}}\phi_o(\boldsymbol{w}) - \nabla^2_{\boldsymbol{w}}\phi(\boldsymbol{w})\right\| \leq C_9 \frac{n^{5/2}}{\mu^2} \log^{3/2}(np) \|\boldsymbol{\Delta}\| .$$

C.3 Proof of Theorem 3

To begin, by Lemma 3, we have

$$\left|\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}}\phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|_2} - \frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}}\phi(\boldsymbol{w})}{\|\boldsymbol{w}\|_2}\right| \leq \|\nabla_{\boldsymbol{w}}\phi_o(\boldsymbol{w}) - \nabla_{\boldsymbol{w}}\phi(\boldsymbol{w})\|_2 \leq c_g \frac{n^{3/2}\log(np)}{\mu} \|\boldsymbol{\Delta}\| \leq \frac{c_2\xi_0\theta}{2}, \tag{73a}$$

$$\left\|\nabla^2_{\boldsymbol{w}}\phi_o(\boldsymbol{w}) - \nabla^2_{\boldsymbol{w}}\phi(\boldsymbol{w})\right\| \leq c_h \frac{n^{5/2}\log^{3/2}(np)}{\mu^2} \|\boldsymbol{\Delta}\| \leq \frac{c_2 n\theta}{2\mu}, \tag{73b}$$

as long as the sample size satisfies

$$\|\boldsymbol{\Delta}\| \leq c_f\kappa^4 \sqrt{\frac{\log^2 n \log p}{\theta^2 p}} \leq C\frac{\xi_0\theta\mu}{n^{3/2}\log^{3/2}(np)},$$

for some constant C in view of Lemma 2. Translating this into the sample size requirement, it means

$$p \geq C\frac{\kappa^8 n^3 \log^4 p \log^2 n}{\theta^4\mu^2\xi_0^2}.$$

Under the assumption of Theorem 2, and in view of (16a) and (16b), we have

$$\frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w})}{\|\boldsymbol{w}\|} \geq \frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|} - \left| \frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi_o(\boldsymbol{w})}{\|\boldsymbol{w}\|} - \frac{\boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w})}{\|\boldsymbol{w}\|} \right| \geq \frac{c_2 \xi_0 \theta}{2},$$

$$\nabla^2_{\boldsymbol{w}} \phi(\boldsymbol{w}) \succeq \nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}) - \left\| \nabla^2_{\boldsymbol{w}} \phi_o(\boldsymbol{w}) - \nabla^2_{\boldsymbol{w}} \phi(\boldsymbol{w}) \right\| \mathbf{I} \succeq \frac{c_2 n \theta}{2\mu} \mathbf{I}.$$

Now let $\boldsymbol{w}^\star$ be the local minimizer of $\phi(\boldsymbol{w})$ in the region of interest. Similar to the proof in Appendix B.4.3, we have

$$\begin{aligned}
\|\boldsymbol{w}^\star\|_2 &\leq \frac{4\mu}{c_2 n \theta} \|\nabla_{\boldsymbol{w}} \phi(\mathbf{0})\|_2 \\
&\leq \frac{4\mu}{c_2 n \theta} \|\nabla_{\boldsymbol{w}} \phi_o(\mathbf{0})\|_2 + \frac{4\mu}{c_2 n \theta} \|\nabla_{\boldsymbol{w}} \phi(\mathbf{0}) - \nabla_{\boldsymbol{w}} \phi_o(\mathbf{0})\|_2 \\
&\leq \frac{c_9 \mu}{n \theta} \left(\sqrt{\frac{n^2 \log(n) \log(p)}{p}} + \frac{n^{3/2} \log(np)}{\mu} \kappa^4 \sqrt{\frac{\log^2(n) \log(p)}{\theta^2 p}} \right).
\end{aligned}$$

where the first term is bounded by (62) and the second term is bounded by Lemma 3. Under the sample size requirement, the latter term dominates and therefore we have

$$\|\boldsymbol{w}^\star - \mathbf{0}\|_2 \leq \frac{c\kappa^4}{\theta^2} \sqrt{\frac{n \log^3 p \log^2 n}{p}}.$$

D Proofs for Section 3.3

We start by stating a useful observation. Notice that $\left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n} \right)$ is on the tangent space of $\boldsymbol{h}$, i.e.,

$$\left(\mathbf{I} - \boldsymbol{h}\boldsymbol{h}^\top \right) \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n} \right) = \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n} \right),$$

we have the relation

$$\partial f(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n} \right) = \left[\left(\mathbf{I} - \boldsymbol{h}\boldsymbol{h}^\top \right) \nabla f(\boldsymbol{h}) \right]^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n} \right) = \nabla f(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n} \right), \tag{74}$$

holds for both $\partial f(\boldsymbol{h})$ and $\partial f_o(\boldsymbol{h})$.

D.1 Proof of Lemma 4

We first prove the upper bound of $\|\partial f(\boldsymbol{h})\|_2$ in (31), which is simpler. Plugging the bound for $\max_{i \in [p]} \|\mathcal{C}(\boldsymbol{x}_i)\|$ in (67) and $\|\boldsymbol{\Delta}\| \leq 1$ ensured by the sample size requirement and Lemma 2, for any $\boldsymbol{h}$ on the unit sphere, with probability at least $1 - (np)^{-8}$, the manifold gradient satisfies

$$\begin{aligned}
\|\partial f(\boldsymbol{h})\|_2 \leq \|\nabla f(\boldsymbol{h})\|_2 &= \left\| \frac{1}{p} \sum_{i=1}^p (\mathbf{I} + \boldsymbol{\Delta})^\top \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i) (\mathbf{I} + \boldsymbol{\Delta}) \boldsymbol{h}(\boldsymbol{w})}{\mu} \right) \right\|_2 \\
&\leq \sqrt{n} \|\mathbf{I} + \boldsymbol{\Delta}\| \max_{i \in [p]} \|\mathcal{C}(\boldsymbol{x}_i)\| \\
&\leq 2C_1 n \sqrt{\log(np)}.
\end{aligned}$$

We now move to prove the lower bound of the directional gradient in (30). We first consider the directional gradient of $f_o(\boldsymbol{h})$ for the orthogonal case, following the proof procedure of the Theorem 2 to obtain the empirical geometry of $f_o(\boldsymbol{h})$ in the region of interest (shown in Lemma 15), which is proved in Appendix D.4.

Lemma 15 (Uniform concentration for the orthogonal case)**.** *Instate the assumptions of Theorem 2. There exist some constants* c_a, c_b*, such that for* $\boldsymbol{h} \in \mathcal{H}_k = \left\{\boldsymbol{h} : \boldsymbol{h} \in \mathcal{S}_{\xi_0}^{(n+)}, h_k \neq 0, h_n^2/h_k^2 < 4\right\}$*, with probability at least* $1-(np)^{-8} - 2\exp\left(-c_a n\right)$*,*

$$\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) \geq c_b \xi_0 \theta. \tag{75}$$

Based on the result, we derive the bound for the directional gradient of $f(\boldsymbol{h})$ in the general case by bounding the deviation between the directional gradient of $f_o(\boldsymbol{h})$ and $f(\boldsymbol{h})$. Using (74), we can relate the directional gradient of $f(\boldsymbol{h})$ to that of $f_o(\boldsymbol{h})$ as

$$\left(\partial f(\boldsymbol{h}) - \partial f_o(\boldsymbol{h})\right)^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) = \left(\nabla f(\boldsymbol{h}) - \nabla f_o(\boldsymbol{h})\right)^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right). \tag{76}$$

We have

$$\begin{aligned}\left|\left(\partial f(\boldsymbol{h}) - \partial f_o(\boldsymbol{h})\right)^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)\right| &\leq \left\|\nabla f(\boldsymbol{h}) - \nabla f_o(\boldsymbol{h})\right\|_2 \left\|\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right\|_2 \\ &\leq \sqrt{5n}\left\|\nabla f(\boldsymbol{h}) - \nabla f_o(\boldsymbol{h})\right\|_2,\end{aligned}$$

where the last line follows from

$$\left\|\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right\|_2 = \sqrt{\frac{1}{h_k^2} + \frac{1}{h_n^2}} \leq \sqrt{\frac{5}{h_n^2}} \leq \sqrt{5n}, \tag{77}$$

due to the assumption $h_n^2/h_k^2 \leq 4$ and $h_n \geq 1/\sqrt{n}$. Therefore, it is sufficient to bound $\left\|\nabla f(\boldsymbol{h}) - \nabla f_o(\boldsymbol{h})\right\|_2$. By (32), we have

$$\begin{aligned}\left\|\nabla f_o(\boldsymbol{h}) - \nabla f(\boldsymbol{h})\right\|_2 &= \left\|\frac{1}{p}\sum_{i=1}^p \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right) - \frac{1}{p}\sum_{i=1}^p \left(\mathbf{I}+\boldsymbol{\Delta}\right)^\top \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}}{\mu}\right)\right\|_2 \\ &\leq \left\|\frac{1}{p}\sum_{i=1}^p \boldsymbol{\Delta}^\top \mathcal{C}(\boldsymbol{x}_i)^\top \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}}{\mu}\right)\right\|_2 \\ &\quad + \left\|\frac{1}{p}\sum_{i=1}^p \mathcal{C}(\boldsymbol{x}_i)^\top \left[\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}}{\mu}\right)\right]\right\|_2 \\ &\leq \max_{i\in[p]}\left\|\mathcal{C}(\boldsymbol{x}_i)\right\| \cdot \max_{i\in[p]}\left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}}{\mu}\right)\right\|_2 \left\|\boldsymbol{\Delta}\right\| \\ &\quad + \max_{i\in[p]}\left\|\mathcal{C}(\boldsymbol{x}_i)\right\| \cdot \max_{i\in[p]}\left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}}{\mu}\right)\right\|_2 \\ &\leq \max_{i\in[p]}\left\|\mathcal{C}(\boldsymbol{x}_i)\right\| \cdot \left(\sqrt{n}\left\|\boldsymbol{\Delta}\right\| + \frac{1}{\mu}\max_{i\in[p]}\left\|\mathcal{C}(\boldsymbol{x}_i)\right\|\left\|\boldsymbol{\Delta}\right\|\right) \\ &\leq C_1 \frac{n\log(np)}{\mu}\left\|\boldsymbol{\Delta}\right\|\end{aligned}$$

with probability at least $1-(np)^{-8}$, where the penultimate inequality follows from (69), and the last inequality follows from (67). By Lemma 2, there exists some constant C, such that under the sample complexity requirement, we have

$$\left|\left(\partial f(\boldsymbol{h}) - \partial f_o(\boldsymbol{h})\right)^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)\right| \leq C_1 \frac{n^{3/2}\log(np)}{\mu}\kappa^4\sqrt{\frac{\log^2 n \log p}{\theta^2 p}} \leq \frac{c_b\xi_0\theta}{2}. \tag{78}$$

In addition, Lemma 15 guarantees that $\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) \geq c_b\xi_0\theta$. Putting together, we have

$$\partial f(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) \geq \partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) - \left|\left(\partial f(\boldsymbol{h}) - \partial f_o(\boldsymbol{h})\right)^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)\right|$$

$$\geq c_b\xi_0\theta - \frac{c_b\xi_0\theta}{2} = \frac{c_b\xi_0\theta}{2}$$

with probability at least $1 - 2(np)^{-8} - 2\exp(-c_a n)$.

D.2 Proof of Lemma 5

Owing to symmetry, without loss of generality, we will show that if the current iterate $\boldsymbol{h} \in \mathcal{S}_{\xi_0}^{(n+)}$ with $\xi_0 \in (0,1)$, the next iterate

$$\boldsymbol{h}^+ = \frac{\boldsymbol{h} - \eta\partial f(\boldsymbol{h})}{\|\boldsymbol{h} - \eta\partial f(\boldsymbol{h})\|_2}$$

stays in $\mathcal{S}_{\xi_0}^{(n+)}$ for a sufficiently small step size η. For any $i \in [n-1]$, we have

$$\left(\frac{h_n^+}{h_i^+}\right)^2 = \frac{(h_n - \eta[\partial f(\boldsymbol{h})]_n)^2}{(h_i - \eta[\partial f(\boldsymbol{h})]_i)^2} = \frac{(1 - \eta[\partial f(\boldsymbol{h})]_n/h_n)^2}{(h_i/h_n - \eta[\partial f(\boldsymbol{h})]_i/h_n)^2}. \tag{79}$$

By (31) in Lemma 4, which bounds $\|\partial f(\boldsymbol{h})\|_\infty \leq \|\partial f(\boldsymbol{h})\|_2 \leq Cn\sqrt{\log(np)}$ for some constant C, and $h_n \geq 1/\sqrt{n}$, by setting $\eta \leq \frac{1}{10Cn^{3/2}\sqrt{\log(np)}}$, we can lower bound the numerator of (79) as

$$\|\eta\partial f(\boldsymbol{h})\|_\infty / h_n \leq \frac{1}{10} \quad \text{and} \quad (1 - \eta[\partial f(\boldsymbol{h})]_n/h_n)^2 \geq \frac{2}{3}. \tag{80}$$

To continue, we take a similar approach to [20, Lemma D.1], and divide our discussions of the denominator of (79) for different coordinates in three subsets:

$$\mathcal{J}_0 := \{i \in [n-1] : h_i = 0\}, \tag{81a}$$

$$\mathcal{J}_1 := \left\{i \in [n-1] : \frac{h_n^2}{h_i^2} \geq 4, h_i \neq 0\right\}, \tag{81b}$$

$$\mathcal{J}_2 := \left\{i \in [n-1] : \frac{h_n^2}{h_i^2} \leq 4\right\}. \tag{81c}$$

- For any index $i \in \mathcal{J}_0$, we have $h_i = 0$, and then by (79) and (80),

$$\left(\frac{h_n^+}{h_i^+}\right)^2 = \frac{(1 - \eta[\partial f(\boldsymbol{h})]_n/h_n)^2}{(\eta[\partial f(\boldsymbol{h})]_i/h_n)^2} \geq \frac{2/3}{(1/10)^2} \geq 2.$$

- For any index $i \in \mathcal{J}_1$, we have

$$\left(\frac{h_n^+}{h_i^+}\right)^2 = \frac{(1 - \eta[\partial f(\boldsymbol{h})]_n/h_n)^2}{(h_i/h_n - \eta[\partial f(\boldsymbol{h})]_i/h_n)^2} \geq \frac{2/3}{(1/4 + 1/10^2)} \geq 2.$$

- For any index $i \in \mathcal{J}_2$, we have

$$\left(\frac{h_n^+}{h_i^+}\right)^2 = \frac{h_n^2}{h_i^2}\left(1 + \eta\frac{\partial f(\boldsymbol{h})^\top(\boldsymbol{e}_i/h_i - \boldsymbol{e}_n/h_n)}{1 - \eta[\partial f(\boldsymbol{h})]_i/h_i}\right)^2.$$

  Since $\boldsymbol{h} \in \mathcal{H}_i$ as defined in Lemma 4, using (30), we have $\partial f(\boldsymbol{h})^\top(\boldsymbol{e}_i/h_i - \boldsymbol{e}_n/h_n) \geq \frac{c_b}{2}\xi_0\theta > 0$, and consequently,

$$\left(\frac{h_n^+}{h_i^+}\right)^2 \geq \frac{h_n^2}{h_i^2}\left(1 + \eta\frac{\partial f(\boldsymbol{h})^\top(\boldsymbol{e}_i/h_i - \boldsymbol{e}_n/h_n)}{1 - \eta[\partial f(\boldsymbol{h})]_i/h_i}\right)^2 \geq \frac{h_n^2}{h_i^2} \geq 1 + \xi_0, \tag{82}$$

  where the last inequality is due to $\boldsymbol{h} \in \mathcal{S}_{\xi_0}^{(n+)}$.

Combining the above, we have that for all $i \in [n-1]$, $\left(h_n^+/h_i^+\right)^2 \geq 1 + \xi_0$, i.e, $\boldsymbol{h}^+ \in \mathcal{S}_{\xi_0}^{(n+)}$.

D.3 Proof of Theorem 4

First, as the step size requirement satisfies that in Lemma 5, the iterates never jumps out of $\mathcal{S}_{\xi_0}^{(n+)}$, if initialized in it. Denote $\boldsymbol{h}_u^+$ as the unnormalized update of $\boldsymbol{h}$ with step size η on the tangent space of $\boldsymbol{h}$, i.e,

$$\boldsymbol{h}_u^+ = \boldsymbol{h} - \eta \partial f(\boldsymbol{h}) = \boldsymbol{h} - \eta \left(\mathbf{I} - \boldsymbol{h}\boldsymbol{h}^\top\right) \nabla_{\boldsymbol{h}} f(\boldsymbol{h}).$$

and $\boldsymbol{w}_u^+$ the first $(n-1)$ entries of $\boldsymbol{h}_u^+$, whose update can be written with respect to $\phi(\boldsymbol{w})$ as

$$\begin{aligned} \boldsymbol{w}_u^+ &= \boldsymbol{w} - \eta \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \left(\mathbf{I} - \boldsymbol{h}\boldsymbol{h}^\top\right) \nabla_{\boldsymbol{h}} f(\boldsymbol{h}) \\ &= \boldsymbol{w} - \eta \left(\mathbf{I} - \boldsymbol{w}\boldsymbol{w}^\top\right) \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}) \nabla_{\boldsymbol{h}} f(\boldsymbol{h}) \\ &= \boldsymbol{w} - \eta \left(\mathbf{I} - \boldsymbol{w}\boldsymbol{w}^\top\right) \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w}). \end{aligned} \tag{83}$$

The normalized updates are respectively $\boldsymbol{h}^+ = \boldsymbol{h}_u^+ / \|\boldsymbol{h}_u^+\|_2$ and $\boldsymbol{w}^+ = \boldsymbol{w}_u^+ / \|\boldsymbol{h}_u^+\|_2$. By the property that $\boldsymbol{h} \perp \left(\mathbf{I} - \boldsymbol{h}\boldsymbol{h}^\top\right) \nabla_{\boldsymbol{h}} f(\boldsymbol{h})$, we have $\|\boldsymbol{h}_u^+\|_2 \geq \|\boldsymbol{h}\|_2 \geq 1$.

Convergence in the region of $\mathcal{Q}_1 \cap \{\boldsymbol{w} : \boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}\}$. By (83), we have

$$\left\|\boldsymbol{w}_u^+\right\|_2^2 = \|\boldsymbol{w}\|_2^2 - \eta \underbrace{h_n(\boldsymbol{w})^2 \boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w})}_{I_1} + \eta^2 \underbrace{\left\|\left(\mathbf{I} - \boldsymbol{w}\boldsymbol{w}^\top\right) \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w})\right\|_2^2}_{I_2}. \tag{84}$$

- First, I_1 can be bounded as

  $$I_1 = h_n^2 \boldsymbol{w}^\top \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w}) \geq c_1 h_n^2 \|\boldsymbol{w}\|_2 \xi_0 \theta.$$

  for some constant c_1, where the last inequality owes to (26a).

- Second, I_2 can be bounded as

  $$I_2 \leq \left\| \begin{bmatrix} \mathbf{I} & \mathbf{0} \end{bmatrix} \right\| \cdot \left\| \mathbf{I} - \boldsymbol{h}\boldsymbol{h}^\top \right\| \|\nabla_{\boldsymbol{h}} f(\boldsymbol{h})\|_2 \leq 1 \cdot (1 + \|\boldsymbol{h}\|_2^2) \|\nabla_{\boldsymbol{h}} f(\boldsymbol{h})\|_2 \leq c_2 n \sqrt{\log(np)}$$

  for some constant c_2, where the last inequality follows from (31) in Lemma 4, which holds with probability at least $1 - (np)^{-8}$.

Sum up the above results for I_1 and I_2, we have with probability at least $1 - (np)^{-8}$,

$$\left\|\boldsymbol{w}^+\right\|_2 \leq \left\|\boldsymbol{w}_u^+\right\|_2^2 \leq \|\boldsymbol{w}\|_2^2 - \eta c_1 h_n^2(\boldsymbol{w}) \|\boldsymbol{w}\|_2 \eta \xi_0 \theta + c_2 \eta^2 n \sqrt{\log(np)},$$

where the first inequality follows from $\|\boldsymbol{h}_u^+\|_2 \geq 1$. Setting $\eta \leq \frac{c\mu\xi_0\theta}{n^2\sqrt{\log(np)}} \leq \frac{c_1 h_n^2 \|\boldsymbol{w}\|_2 \xi_0 \theta}{2 c_2 n \sqrt{\log(np)}}$ for some sufficiently small c, we have

$$\left\|\boldsymbol{w}^+\right\|_2 \leq \|\boldsymbol{w}\|_2^2 - \eta \frac{c_1}{2} h_n^2(\boldsymbol{w}) \|\boldsymbol{w}\|_2 \eta \xi_0 \theta \leq \|\boldsymbol{w}\|_2^2 - \eta \frac{c_1}{2n} \|\boldsymbol{w}\|_2 \xi_0 \theta. \tag{85}$$

Denote the k-th iteration as $\boldsymbol{h}^{(k)}$, we have that as long as $\boldsymbol{w}^{(k)} \in \mathcal{Q}_1$,

$$\left\|\boldsymbol{w}^{(k+1)}\right\|_2^2 \leq \left\|\boldsymbol{w}^{(k)}\right\|_2^2 - \eta \frac{c_1}{2n} \left\|\boldsymbol{w}^{(k)}\right\|_2 \xi_0 \theta \leq \left\|\boldsymbol{w}^{(k)}\right\|_2^2 - \eta \frac{c_1 \mu}{8\sqrt{2}n} \xi_0 \theta.$$

Telescoping the above inequality for the first T_1 iterations, we have

$$T_1 \cdot \eta \frac{c_1 \mu}{8\sqrt{2}n} \xi_0 \theta \leq \left\|\boldsymbol{w}^{(0)}\right\|_2^2 - \left\|\boldsymbol{w}^{(T_1)}\right\|_2^2 \leq 1, \quad \Longrightarrow \quad T_1 \leq \frac{Cn}{\mu \eta \xi_0 \theta},$$

which means it takes at most T_1 iterations to enter $\mathcal{Q}_2 \cap \{\boldsymbol{w} : \boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}\}$.

Remark 1. Using arguments similar to [21], the iteration complexity in this phase can be improved to $T_1 \lesssim \frac{1}{\eta}\left(\frac{n}{\mu\theta} + \log\left(\frac{1}{\xi_0}\right)\right)$ which only has a logarithmic dependence on ξ_0; we didn't pursue it here as it only leads to a logarithmic improvement to the overall complexity due to our choice of $\xi_0 = O(1/\log n)$.

Convergence in the region of $\mathcal{Q}_2 \cap \{\boldsymbol{w} : \boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}\}$: Denoting the unique local minima in $\mathcal{Q}_2 \cap \{\boldsymbol{w} : \boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}\}$ as $\boldsymbol{w}^\star$, whose norm is bounded in Theorem 3. By setting p sufficiently large, we can ensure that the iterates stay in $\mathcal{Q}_2$ following a similar argument as (84). To begin, we note that $\nabla_{\boldsymbol{w}} \phi(\boldsymbol{w})$ is L-Lipschitz with

$$L \leq \frac{Cn^{3/2} \log(np)}{\mu}, \tag{86}$$

which is proved in Appendix D.3.1, and $c_1 n\theta/\mu$-strongly convex in $\mathcal{Q}_2 \cap \{\boldsymbol{w} : \boldsymbol{h}(\boldsymbol{w}) \in \mathcal{S}_{\xi_0}^{(n+)}\}$. For $\boldsymbol{w} \in \mathcal{Q}_2$, we have

$$\frac{1}{2} \leq 1 - \frac{\mu^2}{32} \leq \left\| \mathbf{I} - \boldsymbol{w}\boldsymbol{w}^\top \right\| \leq 1 + \frac{\mu^2}{32}, \tag{87}$$

with $\mu < c_1 \min\{\theta, \xi_0^{1/6} n^{-3/4}\} \leq 4$ sufficiently small.

We now consider two cases based on the size of $\|\boldsymbol{w}^+\|_2$, which is the next iterate with respect to $\boldsymbol{w}$ given in (83).

1. $\|\boldsymbol{w}^+\|_2 < \|\boldsymbol{w}^\star\|_2$: in this case, we already achieve

$$\left\|\boldsymbol{w}^+ - \mathbf{0}\right\|_2 \leq \|\boldsymbol{w}^\star\|_2 \lesssim \frac{\kappa^4}{\theta^2} \sqrt{\frac{n \log^3 p \log^2 n}{p}}.$$

2. $\|\boldsymbol{w}^+\|_2 \geq \|\boldsymbol{w}^\star\|_2$: by the fundamental theorem of calculus, we have

$$\begin{aligned}
\left\|\boldsymbol{w}_u^+ - \boldsymbol{w}^\star\right\|_2 &= \left\|\boldsymbol{w} - \boldsymbol{w}^\star - \eta \left(\mathbf{I} - \boldsymbol{w}\boldsymbol{w}^\top\right) \nabla_{\boldsymbol{w}} \phi(\boldsymbol{w})\right\|_2 \\
&= \left\| \left[\mathbf{I} - \eta \left(\mathbf{I} - \boldsymbol{w}\boldsymbol{w}^\top\right) \int_0^1 \nabla^2_{\boldsymbol{w}} \phi(\boldsymbol{w}(t)) dt \right] (\boldsymbol{w} - \boldsymbol{w}^\star) \right\|_2 \\
&\leq \left\| \left[\mathbf{I} - \eta \left(\mathbf{I} - \boldsymbol{w}\boldsymbol{w}^\top\right) \int_0^1 \nabla^2_{\boldsymbol{w}} \phi(\boldsymbol{w}(t)) dt \right] \right\| \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 \\
&\leq \left(1 - \frac{c_1 n\theta\eta}{2\mu}\right) \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 .
\end{aligned} \tag{88}$$

where $\boldsymbol{w}(t) := \boldsymbol{w} + t(\boldsymbol{w}^\star - \boldsymbol{w})$, $t \in [0,1]$, and the step size $\eta \leq \frac{c\mu\xi_0\theta}{n^2\sqrt{\log(np)}} \leq \frac{1}{2L}$. Moreover, since $\boldsymbol{w}^+ = \boldsymbol{w}_u^+ / \|\boldsymbol{h}^+\|_2 = \boldsymbol{w}_u^+/(1+K)$ for some $K > 0$, we have

$$\begin{aligned}
\left\|\boldsymbol{w}_u^+ - \boldsymbol{w}^\star\right\|_2^2 &= \left\|(1+K)\boldsymbol{w}^+ - \boldsymbol{w}^\star\right\|_2^2 \\
&= \left\|\boldsymbol{w}^+ - \boldsymbol{w}^\star\right\|_2^2 + (2K + K^2) \left\|\boldsymbol{w}^+\right\|_2^2 - 2K \boldsymbol{w}^{\star\top}\boldsymbol{w}^+ \\
&\geq \left\|\boldsymbol{w}^+ - \boldsymbol{w}^\star\right\|_2^2 + (2K + K^2) \left\|\boldsymbol{w}^+\right\|_2^2 - 2K \left\|\boldsymbol{w}^+\right\|_2 \|\boldsymbol{w}^\star\|_2 \geq \left\|\boldsymbol{w}^+ - \boldsymbol{w}^\star\right\|_2^2
\end{aligned} \tag{89}$$

where the last inequality owes to $\|\boldsymbol{w}^+\|_2 \geq \|\boldsymbol{w}^\star\|_2$. Combining (88) and (89), we have the update $\boldsymbol{w}^+$ satisfies

$$\left\|\boldsymbol{w}^+ - \boldsymbol{w}^\star\right\|_2 \leq \left(1 - \frac{c_1 n\theta\eta}{2\mu}\right) \|\boldsymbol{w} - \boldsymbol{w}^\star\|_2 . \tag{90}$$

Therefore, to ensure $\|\boldsymbol{w}^+ - \boldsymbol{w}^\star\|_2 \leq \epsilon$, it takes no more than

$$T_2 \leq \frac{2c_1\mu}{n\theta\eta} \log\left(\frac{3\mu}{2\sqrt{2}\epsilon}\right) \tag{91}$$

iterations, since for any $\boldsymbol{w} \in \mathcal{Q}_2$, we have $\|\boldsymbol{w} - \boldsymbol{w}^*\|_2 \leq \|\boldsymbol{w}\|_2 + \|\boldsymbol{w}^*\|_2 \leq \frac{3\mu}{2\sqrt{2}}$.

Summing up, to achieve $\|\boldsymbol{w}^{(T)} - \mathbf{0}\|_2 \lesssim \frac{\kappa^4}{\theta^2}\sqrt{\frac{n\log^3 p \log^2 n}{p}} + \epsilon$, the total number of iterates is bounded by

$$T = T_1 + T_2 \lesssim \frac{n}{\mu\eta\xi_0\theta} + \frac{\mu}{n\theta\eta} \log\left(\frac{\mu}{\epsilon}\right).$$

We next translate this bound into bounds of $\left\|\boldsymbol{h}^{(T)}-\boldsymbol{e}_n\right\|_2$ and $\mathrm{dist}(\widehat{\boldsymbol{g}}_{\mathrm{inv}},\boldsymbol{g}_{\mathrm{inv}})$, where the latter leads to Corollary 1. For notation simplicity, we denote $\beta := \frac{\kappa^4}{\theta^2}\sqrt{\frac{n\log^3 p\log^2 n}{p}}+\epsilon<1$ which holds for sufficiently large sample size.[4] First, observe that $\boldsymbol{h}^{(T)}=\boldsymbol{h}(\boldsymbol{w}^{(T)})$ satisfies

$$
\begin{aligned}
\left\|\boldsymbol{h}^{(T)}-\boldsymbol{e}_n\right\|_2 &= \sqrt{\left\|\boldsymbol{w}^{(T)}\right\|_2^2+\left(1-\sqrt{1-\left\|\boldsymbol{w}^{(T)}\right\|_2^2}\right)^2} \\
&\le \sqrt{\left(1-\sqrt{1-\beta^2}\right)^2+\beta^2} \le \sqrt{2\beta^2} \lesssim \frac{\kappa^4}{\theta^2}\sqrt{\frac{n\log^3 p\log^2 n}{p}}+\epsilon. \qquad (92)
\end{aligned}
$$

Since we consider the loss function $f(\boldsymbol{h})$ in (21) throughout the proof, the estimate of $\boldsymbol{g}_{\mathrm{inv}}$ is given by $\widehat{\boldsymbol{g}}_{\mathrm{inv}}=\boldsymbol{R}\boldsymbol{U}^\top\boldsymbol{h}^{(T)}$. Hence, we have

$$
\begin{aligned}
\mathrm{dist}(\widehat{\boldsymbol{g}}_{\mathrm{inv}},\boldsymbol{g}_{\mathrm{inv}}) &= \mathrm{dist}(\boldsymbol{R}\boldsymbol{U}^\top\boldsymbol{h}^{(T)},\mathcal{C}(\boldsymbol{g}_{\mathrm{inv}})\boldsymbol{e}_n) \le \left\|\mathcal{C}(\boldsymbol{g})^{-1}\right\|\left\|\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{U}^\top\boldsymbol{h}^{(T)}-\boldsymbol{e}_n\right\|_2 \\
&\le \left\|\mathcal{C}(\boldsymbol{g})^{-1}\right\|\left\|\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{U}^\top(\boldsymbol{h}^{(T)}-\boldsymbol{e}_n)\right\|_2+\left\|\mathcal{C}(\boldsymbol{g})^{-1}\right\|\left\|(\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{U}^\top-\mathbf{I})\boldsymbol{e}_n\right\|_2 \\
&\le \left\|\mathcal{C}(\boldsymbol{g})^{-1}\right\|\left\|\boldsymbol{\Delta}+\mathbf{I}\right\|\left\|\boldsymbol{h}^{(T)}-\boldsymbol{e}_n\right\|_2+\left\|\mathcal{C}(\boldsymbol{g})^{-1}\right\|\left\|\boldsymbol{\Delta}\right\|\left\|\boldsymbol{e}_n\right\|_2 \qquad (93)
\end{aligned}
$$

where we used the definition of $\boldsymbol{\Delta}=\mathcal{C}(\boldsymbol{g})\boldsymbol{R}\boldsymbol{U}^\top-\mathbf{I}$ (cf. (27)). Under the sample size requirement, we have $\|\boldsymbol{\Delta}\|\le c_f\kappa^4\sqrt{\frac{\log^2 n\log p}{\theta^2 p}}<1$ with probability at least $1-2np^{-8}$ by Lemma 2. Plugging it into (93), we have

$$
\begin{aligned}
\mathrm{dist}(\widehat{\boldsymbol{g}}_{\mathrm{inv}},\boldsymbol{g}_{\mathrm{inv}}) &\le \frac{2}{\sigma_n(\mathcal{C}(\boldsymbol{g}))}\left\|\boldsymbol{h}^{(T)}-\boldsymbol{e}_n\right\|_2+\frac{1}{\sigma_n(\mathcal{C}(\boldsymbol{g}))}\|\boldsymbol{\Delta}\| \\
&\lesssim \frac{\kappa^4}{\theta^2\sigma_n(\mathcal{C}(\boldsymbol{g}))}\sqrt{\frac{n\log^3 p\log^2 n}{p}}+\frac{\epsilon}{\sigma_n(\mathcal{C}(\boldsymbol{g}))}, \qquad (94)
\end{aligned}
$$

where the last line follows from (92).

D.3.1 Proof of (86)

Recalling $\psi_\mu'(x)=\tanh(x/\mu)$, for any $\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathcal{Q}_2$, using the expression

$$
\nabla_{\boldsymbol{w}}\phi(\boldsymbol{w})=\frac{1}{p}\sum_{i=1}^p \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w})\left(\mathbf{I}+\boldsymbol{\Delta}\right)^\top\mathcal{C}(\boldsymbol{x}_i)^\top\psi_\mu'\left(\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w})\right),
$$

we have

$$
\begin{aligned}
&\nabla_{\boldsymbol{w}}\phi(\boldsymbol{w}_1)-\nabla_{\boldsymbol{w}}\phi(\boldsymbol{w}_2) \\
&\quad=\underbrace{\frac{1}{p}\sum_{i=1}^p\left[\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}_1)-\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}_2)\right]\left(\mathbf{I}+\boldsymbol{\Delta}\right)^\top\mathcal{C}(\boldsymbol{x}_i)^\top\psi_\mu'\left(\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w}_1)\right)}_{\boldsymbol{g}_1} \\
&\qquad+\underbrace{\frac{1}{p}\sum_{i=1}^p\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}_2)\left(\mathbf{I}+\boldsymbol{\Delta}\right)^\top\mathcal{C}(\boldsymbol{x}_i)^\top\left[\psi_\mu'\left(\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w}_1)\right)-\psi_\mu'\left(\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w}_2)\right)\right]}_{\boldsymbol{g}_2}.
\end{aligned}
$$

We bound $\|\boldsymbol{g}_1\|_2$ and $\|\boldsymbol{g}_2\|_2$ respectively. Under the sample size requirement, from Lemma 2, we can ensure $\|\boldsymbol{\Delta}\|\le 1$. To bound $\|\boldsymbol{g}_1\|_2$, we have

$$
\|\boldsymbol{g}_1\|_2\le\left\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}_1)-\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}_2)\right\|\cdot\left\|\mathbf{I}+\boldsymbol{\Delta}\right\|\cdot\max_{i\in[p]}\left\|\mathcal{C}(\boldsymbol{x}_i)\right\|\cdot\left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w}_1)}{\mu}\right)\right\|_2
$$

[4]Note that $\|\boldsymbol{w}^{(T)}\|_2<1$ by the spherical constraint, the bound is vacuous otherwise

$$\leq C\sqrt{n^2\log(np)}\left\|\left[\mathbf{0} \quad \left(\tfrac{\boldsymbol{w}_2}{h_n(\boldsymbol{w}_2)}-\tfrac{\boldsymbol{w}_1}{h_n(\boldsymbol{w}_1)}\right)\right]\right\|,$$

where the second line follows from $|\tanh(\cdot)|\leq 1$ and (67), which holds with probability at least $1-(np)^{-8}$. To continue, we observe that

$$\begin{aligned}
\left\|\left[\mathbf{0}, \quad \left(\tfrac{\boldsymbol{w}_2}{h_n(\boldsymbol{w}_2)}-\tfrac{\boldsymbol{w}_1}{h_n(\boldsymbol{w}_1)}\right)\right]\right\| &= \left\|\frac{\boldsymbol{w}_2}{h_n(\boldsymbol{w}_2)}-\frac{\boldsymbol{w}_1}{h_n(\boldsymbol{w}_1)}\right\|_2 \\
&\leq \left\|\frac{\boldsymbol{w}_2}{h_n(\boldsymbol{w}_2)}-\frac{\boldsymbol{w}_1}{h_n(\boldsymbol{w}_2)}\right\|_2+\left\|\frac{\boldsymbol{w}_1}{h_n(\boldsymbol{w}_2)}-\frac{\boldsymbol{w}_1}{h_n(\boldsymbol{w}_1)}\right\|_2 \\
&\overset{\text{(i)}}{\leq} 2\left\|\boldsymbol{w}_2-\boldsymbol{w}_1\right\|_2+\left\|\boldsymbol{w}_1\right\|_2\left|\frac{1}{h_n(\boldsymbol{w}_2)}-\frac{1}{h_n(\boldsymbol{w}_1)}\right| \\
&\overset{\text{(ii)}}{\leq} 2\left\|\boldsymbol{w}_2-\boldsymbol{w}_1\right\|_2+8\left\|\boldsymbol{w}_2-\boldsymbol{w}_1\right\|_2 \\
&\leq 10\left\|\boldsymbol{w}_2-\boldsymbol{w}_1\right\|_2
\end{aligned}$$

where (i) follows from $h_n(\boldsymbol{w}_2)=\sqrt{1-\|\boldsymbol{w}_2\|_2^2}\geq 1/2$ when we used the fact that $\|\boldsymbol{w}\|_2\leq\mu/(4\sqrt{2})\leq\sqrt{3}/2$ in $\mathcal{Q}_2$, and (ii) follows from the Lipschitz smoothness of $1/h_n(\boldsymbol{w})$ and $h_n(\boldsymbol{w})\geq 1/2$:

$$\left|\frac{1}{h_n(\boldsymbol{w}_2)}-\frac{1}{h_n(\boldsymbol{w}_1)}\right|\leq\left(\max_{\boldsymbol{w}\in\mathcal{Q}_2}\frac{1}{(1-\|\boldsymbol{w}\|_2^2)^{3/2}}\right)\|\boldsymbol{w}_2-\boldsymbol{w}_1\|_2\leq 8\|\boldsymbol{w}_2-\boldsymbol{w}_1\|_2. \tag{95}$$

Therefore, we have

$$\left\|\boldsymbol{g}_1\right\|_2\leq C\sqrt{n^2\log(np)}\left\|\boldsymbol{w}_2-\boldsymbol{w}_1\right\|_2.$$

To bound $\|\boldsymbol{g}_2\|_2$, we have

$$\begin{aligned}
\|\boldsymbol{g}_2\|_2 &\leq \|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}_2)\|\left\|\mathbf{I}+\boldsymbol{\Delta}\right\|\cdot\max_{i\in[p]}\|\mathcal{C}(\boldsymbol{x}_i)\|\cdot\max_{i\in[p]}\left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w}_1)}{\mu}\right)-\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\left(\mathbf{I}+\boldsymbol{\Delta}\right)\boldsymbol{h}(\boldsymbol{w}_2)}{\mu}\right)\right\|_2 \\
&\leq \frac{1}{\mu}\left\|\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{w}_2)\right\|\left\|\mathbf{I}+\boldsymbol{\Delta}\right\|^2\cdot\max_{i\in[p]}\left\|\mathcal{C}(\boldsymbol{x}_i)\right\|^2\cdot\left\|\boldsymbol{h}(\boldsymbol{w}_1)-\boldsymbol{h}(\boldsymbol{w}_2)\right\|_2 \\
&\leq \frac{Cn^{3/2}\log(np)}{\mu}\left\|\boldsymbol{w}_1-\boldsymbol{w}_2\right\|_2,
\end{aligned}$$

where the second line follows from a similar argument in (69), the last line follows from (66) and (67), which holds with probability at least $1-(np)^{-8}$, and

$$\left\|\boldsymbol{h}(\boldsymbol{w}_1)-\boldsymbol{h}(\boldsymbol{w}_2)\right\|_2\leq\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_2+\left|\frac{1}{h_n(\boldsymbol{w}_2)}-\frac{1}{h_n(\boldsymbol{w}_1)}\right|\leq 9\|\boldsymbol{w}_1-\boldsymbol{w}_2\|_2,$$

following (95). Combining the bounds on $\|\boldsymbol{g}_1\|_2$ and $\|\boldsymbol{g}_2\|_2$ achieve the desired result.

D.4 Proof of Lemma 15

The proof follows a standard covering argument similar to the proof of Theorem 2 in Appendix B.4. To begin, we need the following propositions, proved in Appendix D.4.1, D.4.2, and D.4.3, respectively.

Proposition 3. *For any $\xi_0\in(0,1)$, $\theta\in(0,\frac{1}{3})$, $k\in[n-1]$, there exists some constant c_1 such that when $\mu<c_1\min\{\theta,\xi_0^{1/6}n^{-3/4}\}$, for any $\boldsymbol{h}\in\mathcal{H}_k$, we have*

$$\mathbb{E}\partial f_o(\boldsymbol{h})^\top\left(\frac{\boldsymbol{e}_k}{h_k}-\frac{\boldsymbol{e}_n}{h_n}\right)\geq\frac{\theta\xi_0}{24\sqrt{2\pi}}.$$

Proposition 4. *For any $\xi_0\in(0,1)$, $\theta\in(0,\frac{1}{3})$, $k\in[n-1]$, there exists some constant c_1 such that when $\mu<c_1\min\{\theta,\xi_0^{1/6}n^{-3/4}\}$, for any fixed $\boldsymbol{h}\in\mathcal{H}_k$, there exists some constant C such that for any $t>0$*

$$\mathbb{P}\left(\left|\partial f_o(\boldsymbol{h})^\top\left(\frac{\boldsymbol{e}_k}{h_k}-\frac{\boldsymbol{e}_n}{h_n}\right)-\mathbb{E}\partial f_o(\boldsymbol{h})^\top\left(\frac{\boldsymbol{e}_k}{h_k}-\frac{\boldsymbol{e}_n}{h_n}\right)\right|\geq t\right)\leq 2\exp\left(\frac{-pt^2}{2Cn^3\log n+2t\sqrt{Cn^3\log n}}\right).$$

Proposition 5. *For any $\xi_0 \in (0,1)$, $\theta \in (0, \frac{1}{3})$, $k \in [n-1]$, $\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)$ is L_3-Lipschitz in the domain $\mathcal{H}_k$ with*

$$L_3 \le \max_{i\in[p]} \left(\frac{\sqrt{5n}}{\mu} \left\|\mathcal{C}(\boldsymbol{x}_i)\right\|^2 + 4n^{3/2} \left\|\mathcal{C}(\boldsymbol{x}_i)\right\| \right).$$

We now continue to the proof of Lemma 15. In the subset $\mathcal{H}_k$, for any $0 \le \epsilon \le 2\sqrt{\frac{n-1}{n+\xi_0}}$, we have an ϵ-net $\mathcal{N}_3$ of size at most $\lceil 3/\epsilon \rceil^n$, where ϵ will be determined later. Under the event (67) and Proposition 5, we have

$$L_3 \le \frac{c_{10} n^{3/2}}{\mu} \log(np).$$

For all $\boldsymbol{h} \in \mathcal{H}_k$, there exists $\boldsymbol{h}' \in \mathcal{N}_3$ such that $\|\boldsymbol{h}' - \boldsymbol{h}\|_2 \le \epsilon$. By Proposition 5, we have

$$\left| \partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) - \partial f_o(\boldsymbol{h}')^\top \left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right) \right| \le L_3 \|\boldsymbol{h}' - \boldsymbol{h}\|_2 \le \frac{c_{10} n^{3/2}}{\mu} \log(np)\epsilon \le \frac{c_1 \theta \xi_0}{3},$$

which holds when $\epsilon \le \frac{c\mu\theta\xi_0}{n^{3/2}\log(np)}$ for some sufficiently small c. With this choice of ϵ, the covering number of $\mathcal{N}_3$ satisfies

$$|\mathcal{N}_3| \le \exp\left(n \log\left(\frac{c n^{3/2} \log(np)}{\mu\theta\xi_0} \right) \right).$$

Let $\mathcal{A}_3$ denote the event

$$\mathcal{A}_3 := \left\{ \max_{\boldsymbol{h}\in\mathcal{H}_k} \left| \partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) - \mathbb{E}\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) \right| \le \frac{c_1\theta\xi_0}{3} \right\}.$$

Setting $t = \frac{c_1\theta\xi_0}{3}$ in Proposition 4, by the union bound, $\mathcal{A}_3$ holds with probability at least

$$\begin{aligned} 1 - |\mathcal{N}_3| \cdot 2\exp\left(\frac{-pt^2}{2Cn^3\log n + 2t\sqrt{Cn^3\log n}} \right) &\ge 1 - 2\exp\left(\frac{-c_{11} p\theta^2\xi_0^2}{n^3 \log n} + n\log\left(\frac{c_{10} n^{3/2}\log(np)}{\mu\theta\xi_0} \right) \right) \\ &\ge 1 - 2\exp\left(-c_{12} n\right), \end{aligned}$$

provided $p \ge C\frac{n^4\log n}{\theta^2\xi_0^2} \log\left(\frac{n^{3/2}\log(np)}{\mu\theta\xi_0}\right)$. Finally, we have for all $\boldsymbol{h} \in \mathcal{H}_k$,

$$\begin{aligned} \partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) &= \left[\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) - \partial f_o(\boldsymbol{h}')^\top \left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right) \right] \\ &\quad + \left[\partial f_o(\boldsymbol{h}')^\top \left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right) - \mathbb{E}\partial f_o(\boldsymbol{h}')^\top \left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right) \right] + \mathbb{E}\partial f_o(\boldsymbol{h}')^\top \left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right) \\ &\ge -\frac{c_1\theta\xi_0}{3} - \frac{c_1\theta\xi_0}{3} + c_1\theta\xi_0 = \frac{c_1\theta\xi_0}{3}. \end{aligned}$$

D.4.1 Proof of Proposition 3

First, recall a few notation introduced in Appendix B.1. For $\boldsymbol{x} = \boldsymbol{\Omega} \odot \boldsymbol{z} \sim_{iid} \mathrm{BG}(\theta) \in \mathbb{R}^n$, we denote the first $n-1$ dimension of $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{\Omega}$ as $\bar{\boldsymbol{x}}$, $\bar{\boldsymbol{z}}$ and $\bar{\boldsymbol{\Omega}}$, respectively. Denote $\mathcal{I}$ as the support of $\boldsymbol{\Omega}$ and $\mathcal{J}$ as the support of $\bar{\boldsymbol{\Omega}}$. For any $k \in [n-1]$ with $h_k \neq 0$, by (74) and (32), we have

$$\mathbb{E}\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) = \mathbb{E}\nabla f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) = n \cdot \mathbb{E}\nabla\psi_\mu(\boldsymbol{x}^\top\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right), \tag{96}$$

since the rows of $\mathcal{C}(\boldsymbol{x})$ has the same distribution as $\boldsymbol{x} \sim_{iid} \mathrm{BG}(\theta)$. Further plugging in (34), we rewrite it as:

$$\mathbb{E}\nabla\psi_\mu(\boldsymbol{x}^\top\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) = \mathbb{E}\left[\left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)^\top \tanh\left(\frac{\boldsymbol{x}^\top\boldsymbol{h}}{\mu}\right) \boldsymbol{x} \right]$$

$$= \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{x}^\top \boldsymbol{h}}{\mu}\right)\frac{x_k}{h_k}\right] - \mathbb{E}\left[\tanh\left(\frac{\boldsymbol{x}^\top \boldsymbol{h}}{\mu}\right)\frac{x_n}{h_n}\right]$$
$$= \underbrace{\mathbb{E}_{\boldsymbol{\Omega}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\boldsymbol{x}^\top \boldsymbol{h}}{\mu}\right)\frac{x_k}{h_k}\right]}_{I_1} - \underbrace{\mathbb{E}_{\boldsymbol{\Omega}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\boldsymbol{x}^\top \boldsymbol{h}}{\mu}\right)\frac{x_n}{h_n}\right]}_{I_2}. \tag{97}$$

Evaluating $\mathbb{E}_{\boldsymbol{\Omega}}$ over Ω_k, Ω_n, and $\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}$ sequentially, we can express I_1, I_2 respectively as:

$$\begin{aligned} I_1 &= \theta \mathbb{E}_{\boldsymbol{\Omega}_{\setminus\{k\}}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\boldsymbol{x}_{\setminus\{k\}}^\top \boldsymbol{h}_{\setminus\{k\}} + h_k z_k}{\mu}\right)\frac{z_k}{h_k}\right] \\ &= \theta(1-\theta)\underbrace{\mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\bar{\boldsymbol{x}}_{\setminus\{k\}}^\top \boldsymbol{w}_{\setminus\{k\}} + h_k z_k}{\mu}\right)\frac{z_k}{h_k}\right]}_{I_{11}} + \theta^2 \underbrace{\mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\bar{\boldsymbol{x}}_{\setminus\{k\}}^\top \boldsymbol{w}_{\setminus\{k\}} + h_k z_k + h_n z_n}{\mu}\right)\frac{z_k}{h_k}\right]}_{I_{12}}, \\ I_2 &= (1-\theta)\mathbb{E}_{\boldsymbol{\Omega}_{\setminus\{k\}}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\boldsymbol{x}_{\setminus\{k\}}^\top \boldsymbol{h}_{\setminus\{k\}}}{\mu}\right)\frac{x_n}{h_n}\right] + \theta \mathbb{E}_{\boldsymbol{\Omega}_{\setminus\{k\}}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\boldsymbol{x}_{\setminus\{k\}}^\top \boldsymbol{h}_{\setminus\{k\}} + h_k z_k}{\mu}\right)\frac{x_n}{h_n}\right] \\ &= \theta(1-\theta)\underbrace{\mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\bar{\boldsymbol{x}}_{\setminus\{k\}}^\top \boldsymbol{w}_{\setminus\{k\}} + h_n z_n}{\mu}\right)\frac{z_n}{h_n}\right]}_{I_{21}} + \theta^2 \underbrace{\mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\mathbb{E}_{\boldsymbol{z}}\left[\tanh\left(\frac{\bar{\boldsymbol{x}}_{\setminus\{k\}}^\top \boldsymbol{w}_{\setminus\{k\}} + h_k z_k + h_n z_n}{\mu}\right)\frac{z_n}{h_n}\right]}_{I_{22}}. \end{aligned}$$

Introduce the short-hand notation $X_1 = h_k z_k \sim \mathcal{N}(0, h_k^2)$, $Y_1 = \bar{\boldsymbol{x}}_{\setminus\{k\}}^\top \boldsymbol{w}_{\setminus\{k\}} + h_n z_n$, $X_2 = h_n z_n \sim \mathcal{N}(0, h_n^2)$, $Y_2 = \bar{\boldsymbol{x}}_{\setminus\{k\}}^\top \boldsymbol{w}_{\setminus\{k\}} + h_k z_k$. Invoking Lemma 8, the difference of the second terms of I_1 and I_2 is

$$\begin{aligned} I_{12} - I_{22} &= \mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\left[\mathbb{E}_{X_1,Y_1}\left(\tanh\left(\frac{X_1+Y_1}{\mu}\right)\frac{X_1}{h_k^2}\right)\right] - \mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\left[\mathbb{E}_{X_2,Y_2}\left(\tanh\left(\frac{X_2+Y_2}{\mu}\right)\frac{X_2}{h_n^2}\right)\right] \\ &= \frac{1}{\mu}\mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\left[\mathbb{E}_{X_1,Y_1}\left(1-\tanh^2\left(\frac{X_1+Y_1}{\mu}\right)\right)\right] - \frac{1}{\mu}\mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\left[\mathbb{E}_{X_2,Y_2}\left(1-\tanh^2\left(\frac{X_2+Y_2}{\mu}\right)\right)\right] = 0. \end{aligned}$$

Consequently, we have

$$\begin{aligned} &\mathbb{E}\nabla\psi_\mu(\boldsymbol{x}^\top \boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) = \theta(1-\theta)(I_{11} - I_{21}) \\ &= \frac{\theta(1-\theta)}{\mu}\mathbb{E}_{\bar{\boldsymbol{\Omega}}_{\setminus\{k\}}}\mathbb{E}_{\boldsymbol{z}}\left[\left(1-\tanh^2\left(\frac{\bar{\boldsymbol{x}}_{\setminus\{k\}}^\top \boldsymbol{w}_{\setminus\{k\}} + h_k z_k}{\mu}\right)\right) - \left(1-\tanh^2\left(\frac{\bar{\boldsymbol{x}}_{\setminus\{k\}}^\top \boldsymbol{w}_{\setminus\{k\}} + h_n z_n}{\mu}\right)\right)\right] \\ &\geq \frac{\theta(1-\theta)}{\mu}\frac{\xi_0}{16\sqrt{2\pi}n} = \frac{\theta\xi_0}{24n\sqrt{2\pi}}, \end{aligned}$$

where the second line follows from Lemma 8, and the last line follows from (49) and $\theta \in (0, 1/3)$. Finally, we have

$$\mathbb{E}\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) = n\mathbb{E}\nabla\psi_\mu(\boldsymbol{x}^\top \boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) \geq \frac{\theta\xi_0}{24\sqrt{2\pi}}.$$

D.4.2 Proof of Proposition 4

We start by writing the directional gradient as a sum of p i.i.d. random variables:

$$\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) = \nabla f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) = \frac{1}{p}\sum_{i=1}^{p} \underbrace{\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right)^\top \mathcal{C}(\boldsymbol{x}_i)\left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)}_{Z_i}, \tag{98}$$

where the first equality is due to (74) and the second equality is due to (32). Moreover,

$$|Z_i| \leq \left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right)\right\|_2 \left\|\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right\|_2 \|\mathcal{C}(\boldsymbol{x}_i)\| \leq \sqrt{n}\left\|\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right\|_2 \|\mathcal{C}(\boldsymbol{x}_i)\| \leq \sqrt{5}n\|\mathcal{C}(\boldsymbol{x}_i)\|,$$

where the second inequality follows from $|\tanh(\cdot)| \leq 1$ and the third inequality follows from (77). Therefore, for any $m \geq 2$, the moments of $|Z_i|$ can be controlled by Lemma 10 as

$$\mathbb{E}\,|Z_i|^m \leq \left(\sqrt{5}n\right)^m \mathbb{E}\,\|\mathcal{C}(\boldsymbol{x})\|^m \leq \frac{m!}{2}\left(Cn^3 \log n\right)^{m/2}.$$

The proof is then completed by setting $\sigma^2 = Cn^3 \log n$, $R = \sqrt{Cn^3 \log n}$ and applying the Bernstein's inequality in Lemma 7.

D.4.3 Proof of Proposition 5

Using (98), we have for any $\boldsymbol{h}, \boldsymbol{h}'$,

$$\begin{aligned}
&\left|\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) - \partial f_o(\boldsymbol{h}')^\top \left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right)\right| \\
&= \frac{1}{p}\sum_{i=1}^{p}\left|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right)^\top \mathcal{C}(\boldsymbol{x}_i)\left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}'}{\mu}\right)^\top \mathcal{C}(\boldsymbol{x}_i)\left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right)\right| \\
&\leq \frac{1}{p}\sum_{i=1}^{p}\underbrace{\left|\left[\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}'}{\mu}\right)\right]^\top \mathcal{C}(\boldsymbol{x}_i)\left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)\right|}_{A_i} \\
&\qquad + \frac{1}{p}\sum_{i=1}^{p}\underbrace{\left|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}'}{\mu}\right)^\top \mathcal{C}(\boldsymbol{x}_i)\left[\left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right) - \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)\right]\right|}_{B_i},
\end{aligned}$$

where the second line follows by the triangle inequality. In the sequel, we'll bound A_i and B_i respectively.

- To bound A_i, we have

$$\begin{aligned}
A_i &\leq \left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}}{\mu}\right) - \tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}'}{\mu}\right)\right\|_2 \|\mathcal{C}(\boldsymbol{x}_i)\| \left\|\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right\|_2 \\
&\leq \frac{\sqrt{5n}}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\|^2 \|\boldsymbol{h} - \boldsymbol{h}'\|_2,
\end{aligned}$$

  where the second line follows from (69) and (77).

- To bound B_i, we have

$$\begin{aligned}
B_i &\leq \left\|\tanh\left(\frac{\mathcal{C}(\boldsymbol{x}_i)\boldsymbol{h}'}{\mu}\right)\right\|_2 \|\mathcal{C}(\boldsymbol{x}_i)\| \left\|\left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right) - \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right)\right\|_2 \\
&\leq \sqrt{n}\,\|\mathcal{C}(\boldsymbol{x}_i)\| \sqrt{\left(\frac{1}{h'_k} - \frac{1}{h_k}\right)^2 + \left(\frac{1}{h'_n} - \frac{1}{h_n}\right)^2} \\
&\leq 4n^{3/2}\,\|\mathcal{C}(\boldsymbol{x}_i)\| \sqrt{(h_k - h'_k)^2 + (h_n - h'_n)^2} \leq 4n^{3/2}\,\|\mathcal{C}(\boldsymbol{x}_i)\|\,\|\boldsymbol{h} - \boldsymbol{h}'\|_2
\end{aligned}$$

  where the first line follows from $|\tanh(\cdot)| \leq 1$, and the second line follows from $h_n, h'_n \geq 1/\sqrt{n}$ and $h_n^2/h_k^2 < 4, (h'_n)^2/(h'_k)^2 < 4$.

Combining terms, we have

$$\left|\partial f_o(\boldsymbol{h})^\top \left(\frac{\boldsymbol{e}_k}{h_k} - \frac{\boldsymbol{e}_n}{h_n}\right) - \partial f_o(\boldsymbol{h}')^\top \left(\frac{\boldsymbol{e}_k}{h'_k} - \frac{\boldsymbol{e}_n}{h'_n}\right)\right| \leq \max_{i\in[p]}\left(\frac{\sqrt{5n}}{\mu}\|\mathcal{C}(\boldsymbol{x}_i)\|^2 + 4n^{3/2}\|\mathcal{C}(\boldsymbol{x}_i)\|\right)\|\boldsymbol{h} - \boldsymbol{h}'\|_2.$$