Structured Local Optima in Sparse Blind Deconvolution

Yuqian Zhang, Han-Wen Kuo, John Wright
Department of Electrical Engineering and Data Science Institute
Columbia University

June 1, 2018

Abstract

Blind deconvolution is a ubiquitous problem of recovering two unknown signals from their convolution. Unfortunately, this is an ill-posed problem in general. This paper focuses on the *short and sparse* blind deconvolution problem, where the one unknown signal is short and the other one is sparsely and randomly supported. This variant captures the structure of the unknown signals in several important applications. We assume the short signal to have unit ℓ^2 norm and cast the blind deconvolution problem as a nonconvex optimization problem over the sphere. We demonstrate that (i) in a certain region of the sphere, every local optimum is close to some shift truncation of the ground truth, and (ii) for a generic short signal of length k, when the sparsity of activation signal $\theta \lesssim k^{-2/3}$ and number of measurements $m \gtrsim \mathrm{poly}\,(k)$, a simple initialization method together with a descent algorithm which escapes strict saddle points recovers a near shift truncation of the ground truth kernel.

1 Introduction

Blind deconvolution is the problem of recovering two unknown signals $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ from their convolution $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$. This fundamental problem recurs across several fields, including astronomy, microscopy data processing [CLC+17], neural spike sorting [Lew98], computer vision [KH96], etc. However, this problem is ill-posed without further priors on the unknown signals, as there are infinitely many pairs of signals $(\boldsymbol{a}, \boldsymbol{x})$ whose convolution equals a given observation $\boldsymbol{y}$. Fortunately, in practice, the target signals $(\boldsymbol{a}, \boldsymbol{x})$ are often structured. In particular, a number of practical applications exhibit a common *short-and-sparse* structure:

In *Neural spike sorting*: Neurons in the brain fire brief voltage spikes when stimulated. The signatures of the spikes encode critical features of the neuron and the occurrence of such spikes are usually sparse and random in time [Lew98, ETS11].

In *Microscopy data analysis*: The nanoscale materials of interests are contaminated by randomly and sparsely distributed "defects", which can dramatically change the electronic structure of the material [CLC+17].

In *Image deblurring*: Blurred images due to camera shake can be modeled as a convolution of the latent sharp image and a kernel capturing the motion of the camera. Although natural images are not sparse, they typically have (approximately) sparse gradients [CW98, LWDF11].

In the above applications, the observation signal $\boldsymbol{y} \in \mathbb{R}^m$ is generated via the convolution of a *short* kernel $\boldsymbol{a}_0 \in \mathbb{R}^k$ with $k \ll m$ and a *sparse* activation coefficient $\boldsymbol{x}_0 \in \mathbb{R}^m$ with $\|\boldsymbol{x}_0\|_0 \ll m$. Without loss of generality, we let $\boldsymbol{y}$ denote the circular convolution of $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$

$$\boldsymbol{y} = \boldsymbol{a}_0 \circledast \boldsymbol{x}_0 = \widetilde{\boldsymbol{a}_0} \circledast \boldsymbol{x}_0, \tag{1.1}$$

with $\widetilde{\boldsymbol{a}_0} \in \mathbb{R}^m$ denoting the zero padded m-length version of $\boldsymbol{a}_0$, which can be expressed as $\widetilde{\boldsymbol{a}_0} = \boldsymbol{\iota}_k \boldsymbol{a}_0$. Here, $\boldsymbol{\iota}_k : \mathbb{R}^k \to \mathbb{R}^m$ is a zero padding operator. Its adjoint $\boldsymbol{\iota}_k^* : \mathbb{R}^m \to \mathbb{R}^k$ acts as a projection onto the lower dimensional space by keeping the first k components.

The short-and-sparse blind deconvolution problem exhibits a *scaled-shift ambiguity*, which derives from the basic properties of a convolution operator. Namely, for any observation signal $\boldsymbol{y}$, and any nonzero scalar α and integer shift τ, the following equality always holds

$$\boldsymbol{y} = \left(\pm\alpha s_\tau[\widetilde{\boldsymbol{a}_0}]\right) \circledast \left(\pm\alpha^{-1} s_{-\tau}[\boldsymbol{x}_0]\right). \tag{1.2}$$

Here, $s_{-\tau}[\boldsymbol{v}]$ denotes the cyclic shift of the vector $\boldsymbol{v}$ by τ entries:

$$s_\tau[\boldsymbol{v}](i) = \boldsymbol{v}\left([i-\tau-1]_m + 1\right), \quad \forall\, i \in \{1, \cdots, m\}. \tag{1.3}$$

Clearly, both scaling and cyclic shifts preserve the short-and-sparse structure of $(\boldsymbol{a}_0, \boldsymbol{x}_0)$. This *scaled-shift symmetry* raises nontrivial challenges for computation, making straightforward convexification approaches ineffective, and leading to complicated nonconvex optimization landscape. [ZLK+17] considers a natural nonconvex formulation of sparse blind deconvolution, in which the kernel $\boldsymbol{a} \in \mathbb{R}^k$ is constrained to have unit Frobenius norm. [ZLK+17] argues that under certain idealized conditions, this problem has well-structured local optima, in the sense that *every local optimum is close to some shift truncation of the ground truth*. The presence of these local optima can be viewed as a result of the shift symmetry associated to the convolution operator: the shifted and truncated kernel $\boldsymbol{\iota}_k^* s_\tau[\widetilde{\boldsymbol{a}_0}]$ can be convolved with the sparse signal $s_{-\tau}[\boldsymbol{x}_0]$ (shifted in the other direction) to produce a near approximation to $\boldsymbol{y}$ that

$$\left(\boldsymbol{\iota}_k^* s_\tau[\widetilde{\boldsymbol{a}_0}]\right) \circledast s_{-\tau}[\boldsymbol{x}_0] \approx \boldsymbol{y}. \tag{1.4}$$

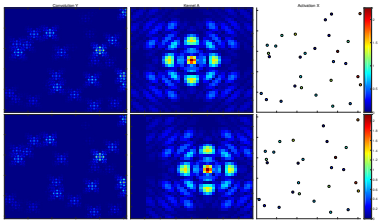

Figure 1: Local Minimum. Top: observation $\boldsymbol{y} = \boldsymbol{a}_0 \circledast \boldsymbol{x}_0$, and ground truth $\boldsymbol{a}_0$, and $\boldsymbol{x}_0$; Bottom: recovered $\boldsymbol{a} \circledast \boldsymbol{x}$, $\boldsymbol{a}$, and $\boldsymbol{x}$ at one local minimum of a natural formulation in [ZLK+17].

In [ZLK+17], the geometric insight about local minima is corroborated with a lot of experiments, but rigorous proof is only available under rather restrictive conditions. In this paper, we adopt the unit Frobenius norm constraint as in [ZLK+17] but consider a different objective function over the kernel sphere $\mathbb{S}^{k-1}$. We formulate the sparse blind deconvolution problem as the following optimization problem over the sphere:

$$\min \; -\left\|\check{\boldsymbol{y}} \circledast \boldsymbol{r}_{\boldsymbol{y}}(\boldsymbol{q})\right\|_4^4 \quad \text{s.t.} \quad \|\boldsymbol{q}\|_F = 1 \tag{1.5}$$

Here, $\check{\boldsymbol{y}}$ denotes the reversal[1] of $\boldsymbol{y}$ and $\boldsymbol{r}_{\boldsymbol{y}}(\boldsymbol{q})$ is a preconditioner which we will discuss in detail later. Convolution $\check{\boldsymbol{y}} \circledast \boldsymbol{r}_{\boldsymbol{y}}(\boldsymbol{q})$ approximates the reversed underlying activation signal $\boldsymbol{x}_0$, and $-\|\cdot\|_4^4$ serves as the sparsity penalty.

This paper studies the function landscape of the short-and-sparse blind deconvolution problem assuming the short k-length convolutional kernel lives on a unit Frobenius norm sphere, denoted as $\mathbb{S}^{k-1}$. We demonstrate that even when $\boldsymbol{x}_0$ is relatively dense, a shift truncation $\boldsymbol{\iota}_k^* s_\tau[\widetilde{\boldsymbol{a}_0}]$ of the ground truth still can be obtained as one local minimum in certain region of the kernel sphere. Such benign region contains the sub-level set of small objective value, and an initial point with small objective value can be easily found. Specifically, for a generic kernel on the sphere [2] $\boldsymbol{a}_0 \in \mathbb{S}^{k-1}$, if the sparsity rate $\theta \lesssim k^{-2/3}$ and the number of measurement $m \gtrsim \mathrm{poly}(k)$, initializing with some k consecutive entries of $\boldsymbol{y}$ and applying any optimization method which (i) is a descent method, and (ii) converges to a local minimizer under a strict saddle hypothesis [JGN+17, XRKM17], produces a near shift-truncation of the ground truth.

[1] Denote $\boldsymbol{y} = [y_1, y_2, \cdots, y_{m-1}, y_m]^T$, then its reversal $\check{\boldsymbol{y}} = [y_1, y_m, y_{m-1}, \cdots, y_2]^T$.

[2] Here, we refer a kernel sampled following a uniform distribution over the sphere as a generic kernel on the sphere.

1.1 Related Works

Even after accounting for the scale ambiguity, the general blind deconvolution problem remains ill-posed. Different types of prior knowledge about the unknown signals have been introduced and to make the blind deconvolution problem well posed. For example, if the signals $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ live on known linear subspaces, the blind deconvolution problem can be cast as a low-rank recovery problem, and solved via semidefinite programming. [ARR12] proves that if one of the subspaces is random and the other satisfies a spectral flatness condition, this approach recovers the pair $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ up to scale. [LLSW16] provides a more efficient nonconvex algorithm for blind deconvolution under this subspace model. [LS15] consider a more complicated model in which one of the signals is sparse in some known dictionary. [LLJB17] considers the case where both convolutional signals are sparse in some known dictionaries. These known dictionaries are assumed to be random (e.g., Gaussian or partial Fourier). Identifiability of these blind deconvolution problems is investigated in [LLB16, LLB17]. [LS17] further addresses a simultaneous demixing and deconvolution problem, where the observation is the superposition of multiple convolutions.

The above results offer efficient and guaranteed algorithms for blind deconvolution problems in which the signals of interest are sparse in a random dictionary. However, in the short-and-sparse blind deconvolution problem in microscopy image analysis or neural spike sorting, the sparse signal is *sparse with respect to the standard basis* rather than a random dictionary. Any cyclic shift of a standard basis is another standard basis, therefore the short-and-sparse blind deconvolution problem is only identifiable up to shifts. This is in contrast to the aforementioned random models, which only exhibit a scale ambiguity. When casting the short-and-sparse blind deconvolution problem as an optimization problem, this shift ambiguity creates a large group of equivalent global solutions (convolutional pairs of opposite shifts $s_\tau[\widetilde{\boldsymbol{a}_0}]$ and $s_{-\tau}[\boldsymbol{x}_0]$) and therefore much more complicated optimization landscape.

For sparsity in the standard basis, [CM14, CM15] show that sparsity alone is not sufficient for unique recovery, by demonstrating the existence of manifolds $(\boldsymbol{a}, \boldsymbol{x})$ of signals that are not identifiable from the convolution $\boldsymbol{y} = \boldsymbol{a} * \boldsymbol{x}$. This construction requires both the support and magnitudes of the two signals to be regular: the support of $\boldsymbol{x}$ needs to have the form $J \cup s_1(J)$ for some set J, and the nonzero entries of $\boldsymbol{x}$ to take on specific values. When $\boldsymbol{x}$ is either Bernoulli or Bernoulli-Gaussian, with probability one, the pair $(\boldsymbol{a}, \boldsymbol{x})$ does not fall in this non-identifiable set. [Chi16] proposes a convex relaxation for a variant of the sparse blind deconvolution problem in which $\boldsymbol{a}$ lies in a random subspace and $\boldsymbol{x}$ is a superposition of spikes with continuous-valued locations. A strong point of this method is that it avoids discretization. Because of the random subspace model on $\boldsymbol{a}$, the results of [Chi16] are not directly comparable to ours. However, if the rates from this work were adapted to the short-and-sparse setting, they would require $\boldsymbol{x}$ to be sparse enough that the observation $\boldsymbol{y}$ contains many isolated (non-overlapping) copies of $\boldsymbol{a}$. This seems to reflect a fundamental limitation of convexification approaches in handling signals with multiple structures [OJF+15]. [WC16] studies another variant where multiple independent observations of circulant convolutions are available, motivated by multi-channel blind deconvolution. Although the convolution kernel is short compared to the total measurements, each independent "short" measurement is self contained. While in the short-and-sparse blind deconvolution problem, only one measurement is available and any "short" measurement heavily depends on adjacent measurements. This nuance leads to much more complicated optimization geometry.

Although the theory of short-and-sparse blind deconvolution remains completely open, many nonconvex algorithms have been developed and practiced in computer vision, where the convolution kernel captures the image blurring process due to camera shake [LWDF11]. Motivated by this physical model, people assume the convolutional kernel to be entry-wise nonnegative and sums up to 1, and then minimize the objective function of following form

$$\min_{\boldsymbol{a}\geq 0, \|\boldsymbol{a}\|_1=1} \ \min_{\boldsymbol{x}} \ \tfrac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{a} \circledast \boldsymbol{x}\right\|_2^2 + \lambda \left\|\boldsymbol{x}\right\|_\star . \tag{1.6}$$

In the image deblurring application, $\boldsymbol{x}$ represents the gradient of a natural image and $\|\cdot\|_\star$ penalizes the sparsity of $\boldsymbol{x}$. However, such formulation always admits one local minimum obtained at the convolutional pair $(\boldsymbol{a}, \boldsymbol{x}) = (\boldsymbol{\delta}, \boldsymbol{y})$ [BVG13, PF14]. In contrast, [WZ13, ZWZ13] carefully compare the difference in MAP and VB approaches, and propose to instead constrain $\boldsymbol{a}$ to have unit Frobenius norm – i.e., to reside on a high-dimensional sphere. [ZLK+17] studies the optimization landscape of the sphere constrained sparse

blind deconvolution and firstly identifies the structure of the local solutions. In particular, [ZLK+17] casts the short-and-sparse blind deconvolution problem as an optimization problem over the sphere:

$$\min_{\boldsymbol{a}\in\mathbb{S}^{k-1}} \min_{\boldsymbol{x}} \tfrac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{a}\circledast\boldsymbol{x}\right\|_2^2+\lambda\left\|\boldsymbol{x}\right\|_1, \tag{1.7}$$

and presents empirical evidence that *local minima $\bar{\boldsymbol{a}}$ are close to certain shift truncations of $\boldsymbol{a}_0$*. [ZLK+17] further proves that a "linearized" version of (1.7), which neglects quadratic interactions in $\boldsymbol{a}$, satisfies this property, in the "dilute limit" in which the sparse signal $\boldsymbol{x}_0$ is a single spike. In this paper, we demonstrate that for a different objective function, this claim holds under much broader conditions than what is proved in [ZLK+17]. In particular, our results allow the sparse signal $\boldsymbol{x}_0$ to be much denser.

1.2 Assumptions and Notations

We assume that $\boldsymbol{x}_0\in\mathbb{R}^m$ follows the Bernoulli-Gaussian (BG) model with sparsity level θ: $\boldsymbol{x}_0(i)=\omega_i g_i$ with $\omega_i\sim\mathrm{Ber}(\theta)$ and $g_i\sim\mathcal{N}(0,1)$, where all the different random variables are jointly independent. For simplicity, we write $\boldsymbol{x}_0\sim_{i.i.d.}\mathrm{BG}(\theta)$.

Throughout this paper, vectors $\boldsymbol{v}\in\mathbb{R}^k$ are indexed as $\boldsymbol{v}=[v_1,v_2,\cdots,v_k]$, and $[\cdot]_m$ denotes the modulo operator of m. We use $\|\cdot\|_2$ to denote the operator norm, $\|\cdot\|_F$ to denote the Frobenius norm, and $\|\cdot\|_p$ to denote the entry wise ℓ^p norm. $(\cdot)_I$ denotes the projection onto subset with index I and $\mathcal{P}_{\mathbb{S}}[\cdot]=\frac{\cdot}{\|\cdot\|_F}$ denotes the projection onto the Frobenius sphere. $(\cdot)^{\circ p}$ is the entry wise p-th order exponent operator. We use C, c to denote positive constants, and their value change across the paper.

2 Problem Formulation and Main Results

In the short-and-sparse blind deconvolution problem, any k consecutive entries in $\boldsymbol{y}$ only depend on $2k-1$ consecutive entries in $\boldsymbol{x}_0$:

$$\boldsymbol{y}_i=\left[y_i,\cdots,y_{1+[i+k-1]_m}\right]^T=\sum_{\tau=-(k-1)}^{k-1}x_{1+[i+\tau-1]_m}\cdot\boldsymbol{\iota}_k^* s_\tau[\widetilde{\boldsymbol{a}_0}] \tag{2.1}$$

$$=\underbrace{\begin{bmatrix} a_k & a_{k-1} & \cdots & a_1 & \cdots & 0 & 0\\ 0 & a_k & \cdots & a_2 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & a_{k-1} & \cdots & a_1 & 0\\ 0 & 0 & \cdots & a_k & \cdots & a_2 & a_1 \end{bmatrix}}_{\boldsymbol{A}_0\in\mathbb{R}^{k\times(2k-1)}}\underbrace{\begin{bmatrix} x_{1+[i-k]_m}\\ \vdots\\ x_i\\ \vdots\\ x_{1+[i+k-2]_m}\end{bmatrix}}_{\boldsymbol{x}_i\in\mathbb{R}^{(2k-1)\times 1}}. \tag{2.2}$$

Write $\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_m]\in\mathbb{R}^{k\times m}$ and $\boldsymbol{X}_0=[\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m]\in\mathbb{R}^{2k-1\times m}$. Using the above expression, we have that

$$\boldsymbol{Y}=\boldsymbol{A}_0\boldsymbol{X}_0. \tag{2.3}$$

Each column $\boldsymbol{x}_i$ of $\boldsymbol{X}_0$ only contains some $2k-1$ entries of $\boldsymbol{x}_0$. The *rows* of $\boldsymbol{X}_0$ are cyclic shifts of the reversal of $\boldsymbol{x}_0$:

$$\boldsymbol{X}_0=\begin{bmatrix} s_0[\check{\boldsymbol{x}}_0]\\ \vdots\\ s_{2k-2}[\check{\boldsymbol{x}}_0]\end{bmatrix}. \tag{2.4}$$

The shifts of $\check{\boldsymbol{x}}_0$ are *sparse vectors* in the linear subspace $\mathrm{row}(\boldsymbol{X}_0)$. Note that if we could recover some shift $s_\tau[\boldsymbol{x}_0]$, we could subsequently determine $s_{-\tau}[\boldsymbol{a}_0]$ by solving a linear system of equations, and hence solve the deconvolution problem, up to the shift ambiguity.

2.1 Finding a Shifted Sparse Signal

In light of the above observations, a natural computational approach to sparse blind deconvolution is to attempt to find $\boldsymbol{x}_0$ by searching for a sparse vector in the linear subspace $\mathrm{row}(\boldsymbol{X}_0)$, e.g., by solving an optimization problem

$$\min \quad \|\boldsymbol{v}\|_\star \quad \text{s.t.} \quad \boldsymbol{v} \in \mathrm{row}\left(\boldsymbol{X}_0\right),\ \|\boldsymbol{v}\|_2 = 1, \tag{2.5}$$

where $\|\cdot\|_\star$ is chosen to encourage sparsity of the target signal [SWW12, SQW15, QSW16, HSSS16].

In sparse blind deconvolution, we do not have access to the row space of $\boldsymbol{X}_0$. Instead, we only observe the subspace $\mathrm{row}(\boldsymbol{Y}) \subset \mathrm{row}(\boldsymbol{X}_0)$. The subspace $\mathrm{row}(\boldsymbol{Y})$ does not necessarily contain the desired sparse vector $\boldsymbol{e}_i^T \boldsymbol{X}_0$, but it *does* contain some approximately sparse vectors. In particular, consider following vector in $\mathrm{row}(\boldsymbol{Y})$,

$$\boldsymbol{v} = \boldsymbol{Y}^T \boldsymbol{a}_0 = \underset{\textbf{sparse}}{\check{\boldsymbol{x}}_0} + \underbrace{\sum_{i \neq 0} \langle \boldsymbol{a}_0, s_i[\boldsymbol{a}_0] \rangle\, s_i[\check{\boldsymbol{x}}_0]}_{\text{“noise” } \boldsymbol{z}}. \tag{2.6}$$

The vector $\boldsymbol{v}$ is a superposition of a sparse signal $\check{\boldsymbol{x}}_0$ and its scaled shifts $\langle \boldsymbol{a}_0, s_i[\boldsymbol{a}_0] \rangle\, s_i[\check{\boldsymbol{x}}_0]$. If the shift-coherence $|\, \langle \boldsymbol{a}_0, s_\tau[\boldsymbol{a}_0] \rangle\, |$ is small[3] and $\boldsymbol{x}_0$ is sparse enough, $\boldsymbol{z}$ can be viewed as small noise.[4] The vector $\boldsymbol{v}$ is not sparse, but it is *spiky*: a few of its entries are much larger than the rest. We deploy a milder sparsity penalty $-\|\cdot\|_4^4$ to recover such a spiky vector, as $\|\cdot\|_4^4$ is very flat around 0 and insensitive to small noise in the signal.[5] This gives

$$\min \quad -\tfrac{1}{4}\|\boldsymbol{v}\|_4^4 \quad \text{s.t.} \quad \boldsymbol{v} \in \mathrm{row}\left(\boldsymbol{Y}\right),\ \|\boldsymbol{v}\|_2 = 1. \tag{2.7}$$

We can express a generic unit vector $\boldsymbol{v} \in \mathrm{row}(\boldsymbol{Y})$ as $\boldsymbol{v} = \boldsymbol{Y}^T \left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} \boldsymbol{q}$, with $\|\boldsymbol{v}\|_2 = \|\boldsymbol{q}\|_2$. This leads to the following equivalent optimization problem over the sphere

$$\min\ \psi\left(\boldsymbol{q}\right) \doteq -\frac{1}{4m} \left\| \boldsymbol{Y}^T \left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} \boldsymbol{q} \right\|_4^4 \quad \text{s.t.} \quad \|\boldsymbol{q}\|_2 = 1. \tag{2.8}$$

Interpretation: preconditioned shifts. This objective $\psi\left(\boldsymbol{q}\right)$ can be rewritten as

$$\psi\left(\boldsymbol{q}\right) = -\frac{1}{4m} \left\| \check{\boldsymbol{y}} \circledast \left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} \boldsymbol{q} \right\|_4^4 = -\frac{1}{4m} \left\| \check{\boldsymbol{x}}_0 \circledast \boldsymbol{A}_0^T \left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} \boldsymbol{q} \right\|_4^4 \sim \|\check{\boldsymbol{x}}_0 \circledast \boldsymbol{\zeta}\|_4^4, \tag{2.9}$$

where $\boldsymbol{\zeta} = \boldsymbol{A}_0^T (\boldsymbol{A}_0 \boldsymbol{A}_0^T)^{-1/2} \boldsymbol{q}$. This approximation becomes accurate as m grows.[6] This objective encourages the convolution of $\check{\boldsymbol{x}}_0$ and $\boldsymbol{\zeta}$ to be as spiky as possible. Reasoning analogous to (2.6) suggests that $\check{\boldsymbol{x}}_0 \circledast \boldsymbol{\zeta}$ will be spiky if

$$\boldsymbol{\zeta} = \boldsymbol{A}_0^T \left(\boldsymbol{A}_0 \boldsymbol{A}_0^T\right)^{-1/2} \boldsymbol{q} \approx \boldsymbol{e}_l, \quad l \in \{1, \cdots, 2k-1\}. \tag{2.10}$$

For simplicity, we define the preconditioned convolution matrix

$$\boldsymbol{A} \doteq \left(\boldsymbol{A}_0 \boldsymbol{A}_0^T\right)^{-1/2} \boldsymbol{A}_0 = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_{2k-1} \end{bmatrix}, \tag{2.11}$$

with column coherence (preconditioned shift coherence) $\mu \doteq \max_{i \neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|$. Then $\boldsymbol{\zeta}$ can also be interpreted as measuring the inner products of $\boldsymbol{q}$ with columns of $\boldsymbol{A}$. Making this intuition rigorous, we will show that minimizing this objective over a certain region of the sphere yields a preconditioned shift truncate $\boldsymbol{a}_l$, from which we can recover a shift truncate of the original signal $\boldsymbol{a}_0$.

2.2 Structured Local Minima

[3] For a generic kernel $\boldsymbol{a}_0$, the shift-coherence is bounded by $|\, \langle \boldsymbol{a}_0, s_\tau[\boldsymbol{a}_0] \rangle\, | \approx 1/\sqrt{k}$ for any shift τ.

[4] In particular, under a Bernoulli-Gaussian model, for each j, $\mathbb{E}[\boldsymbol{z}_j^2] = \theta \sum_{i \neq 0} \langle \boldsymbol{a}_0, s_i[\boldsymbol{a}_0] \rangle^2$.

[5] In comparison, the classical choice $\|\cdot\|_\star = \|\cdot\|_1$ is a strict sparsity penalty that essentially encourages all small entries to be 0.

[6] As $\mathbb{E}_{\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)}[\boldsymbol{Y}\boldsymbol{Y}^T] = \mathbb{E}_{\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)}[\boldsymbol{A}_0 \boldsymbol{X}_0 \boldsymbol{X}_0^T \boldsymbol{A}_0^T] = \theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T$.

We will show that in a certain region $\mathcal{R}_{C_\star} \subset \mathbb{S}^{k-1}$, the preconditioned shift truncations $\boldsymbol{a}_l$ are the *only* local minimizers. Moreover, the other critical points in $\mathcal{R}_{C_\star}$ can be interpreted as resulting from competition between several of these local minima (Figure 2). At any saddle point, there exists strict negative curvature in the direction of a nearby local minimizer which breaks the balance in favor of some particular $\boldsymbol{a}_l$. The region $\mathcal{R}_{C_\star}$ is defined as follows:

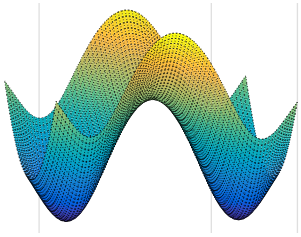

Figure 2: Saddles points are approximately superpositions of local minima.

Definition 2.1. *For fixed $C_\star > 0$, letting κ denote the condition number of $\boldsymbol{A}_0$, and $\mu \doteq \max_{i\neq j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle|$ the column coherence of $\boldsymbol{A}$, we define two regions $\mathcal{R}_{C_\star}$, $\hat{\mathcal{R}}_{C_\star} \subset \mathbb{S}^{k-1}$, as*

$$\mathcal{R}_{C_\star} \doteq \left\{ \boldsymbol{q} \in \mathbb{S}^{k-1} \,|\, \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6 \geq C_\star \mu\kappa^2 \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_3^3 \right\}. \tag{2.12}$$

$$\hat{\mathcal{R}}_{C_\star} \doteq \left\{ \boldsymbol{q} \in \mathbb{S}^{k-1} \,|\, \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6 \geq C_\star \mu\kappa^2 \right\} \subseteq \mathcal{R}_{C_\star}. \tag{2.13}$$

A simpler and smaller region $\hat{\mathcal{R}}_{C_\star}$ is also introduced in Definition (2.1). This region $\hat{\mathcal{R}}_{C_\star}$ can be viewed as a sub-level set for $-\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4$, which is proportional to the objective value $\psi(\boldsymbol{q})$ assuming m is sufficiently large[7]. Therefore, once initialized within $\hat{\mathcal{R}}_{C_\star}$, the iterates produced by a descent algorithm will stay in $\hat{\mathcal{R}}_{C_\star}$.

In particular, at any stationary point $\boldsymbol{q} \in \mathcal{R}_{10}$, the local optimization landscape can be characterized in terms of the number of spikes (entries with nontrivial magnitude[8]) in $\boldsymbol{\zeta}$. If there is only one spike in $\boldsymbol{\zeta}$, then such stationary point $\boldsymbol{q}$ is a local minimum that is close to one local minimizer; if there are more than two spikes in $\boldsymbol{\zeta}$, then such stationary point $\boldsymbol{q}$ is saddle point. Based on the above characterizations of stationary points in $\mathcal{R}_{C_\star}$ with $C_\star \geq 10$, we can deduce that any local minimum is close to some $\boldsymbol{a}_l$, a preconditioned shift truncation of the ground truth $\boldsymbol{a}_0$.

Theorem 2.2 (Main Result)**.** *Assuming observation $\boldsymbol{y} \in \mathbb{R}^m$ is the circulant convolution of $\boldsymbol{a}_0 \in \mathbb{R}^k$ and $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^m$, where the convolutional matrix $\boldsymbol{A}_0$ has minimum singular value $\sigma_{\min} > 0$ and condition number $\kappa \geq 1$, and $\boldsymbol{A}$ has column incoherence $0 \leq \mu < 1$. There exists a positive constant C such that whenever the number of measurements*

$$m \geq C \frac{\min\left\{\mu^{-4/3}, \kappa^2 k^2\right\}}{(1-\theta)^2 \sigma_{\min}^2} \kappa^8 k^4 \log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right) \tag{2.14}$$

and $\theta \geq \log k/k$, then with high probability, any local optima $\bar{\boldsymbol{q}} \in \hat{\mathcal{R}}_{2C_\star}$ satisfies

$$|\langle \bar{\boldsymbol{q}}, \mathcal{P}_{\mathbb{S}}[\boldsymbol{a}_l]\rangle| \geq 1 - c_\star \kappa^{-2} \tag{2.15}$$

for some integer $1 \leq l \leq 2k-1$. Here, $C_\star \geq 10$ and $c_\star = 1/C_\star$.

This theorem says that any local minimum in $\hat{\mathcal{R}}_{2C_\star}$ is close to some normalized column of $\boldsymbol{A}$ given polynomially many observation. The parameters $\sigma_{\min}$, κ and μ effectively measure the spectrum flatness of the ground truth kernel $\boldsymbol{a}_0$ and characterize how broad the results hold. A random like kernel usually has big $\sigma_{\min}$, small κ and μ, which equivalently implies the result holds in a large sub-level set $\hat{\mathcal{R}}_{2C_\star}$ even with fewer observations.

Hence, once assuring the algorithm finds a local minimum in $\hat{\mathcal{R}}_{2C_\star}$, then some shifted truncation of the ground truth kernel $\boldsymbol{a}_0$ can be recovered. In other words, if we can find an initialization point with small objective value, then a descent algorithm minimizing the objective function guarantees that $\boldsymbol{q}$ always stays in $\hat{\mathcal{R}}_{2C_\star}$ in proceeding iterations. Therefore, any descent algorithm that escapes a strict saddle point can be applied to find some $\boldsymbol{a}_l$, or some shift truncation of $\boldsymbol{a}_0$.

[7] Please refer to Section 3 for more arguments.

[8] We call any ζ_l with magnitude no smaller than $2\mu\left\|\boldsymbol{\zeta}\right\|_3^3 / \left\|\boldsymbol{\zeta}\right\|_4^4$ to be nontrivial and defer technical reasonings to later sections.

2.3 Initialization with a Random Sample

Recall that $\boldsymbol{y}_i = \boldsymbol{A}_0 \boldsymbol{x}_i$, which is a sparse superposition of about $2\theta k$ columns of $\boldsymbol{A}_0$. Intuitively speaking, such $\boldsymbol{q}_{\text{init}}$ already encodes certain preferences towards a few preconditioned shift truncations of the ground truth. Therefore, we randomly choose an index i and set the initialization point as

$$\boldsymbol{q}_{\text{init}} = \mathcal{P}_{\mathbb{S}} \left[\left(\boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} \boldsymbol{y}_i \right]. \tag{2.16}$$

Using $\mathbb{E}_{\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta)} [\boldsymbol{Y}\boldsymbol{Y}^T] = \theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T$ again, we have

$$\boldsymbol{\zeta}_{\text{init}} = \boldsymbol{A}^T \boldsymbol{q}_{\text{init}} \approx \mathcal{P}_{\mathbb{S}} \left[\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x}_i \right]. \tag{2.17}$$

For a generic kernel $\boldsymbol{a}_0 \in \mathbb{S}^{k-1}$, $\boldsymbol{A}^T \boldsymbol{A}$ is close to a diagonal matrix, as the magnitudes of off-diagonal entries are bounded by column incoherence μ. Hence, the sparse property of $\boldsymbol{x}_i$ can be approximately preserved, that $\mathcal{P}_{\mathbb{S}} \left[\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{x}_i \right]$ is spiky vector with small $-\|\cdot\|_4^4$. By leveraging the sparsity level θ, one can make sure such initialization point $\boldsymbol{q}_{\text{init}}$ falls in $\hat{\mathcal{R}}_{2C_\star}$. Therefore, we propose Algorithm 1 for solving sparse blind deconvolution with its working conditions stated in Corollary 2.3. For the choice of descent algorithms which escape strict saddle points, there are several such algorithms specially tailored for sphere constrained optimization problems [ABG07, GWY09].

Algorithm 1 Short and Sparse Blind Deconvolution

Input: Observations $\boldsymbol{y} \in \mathbb{R}^m$ and kernel size k.
Output: Recovered Kernel $\bar{\boldsymbol{a}}$.

1: Generate random index $i \in [1, m]$ and set

$$\boldsymbol{q}_{init} = \mathcal{P}_{\mathbb{S}} \left[\left(\boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} \boldsymbol{y}_i \right].$$

2: Solve following nonconvex optimization problem with a descent algorithm that escapes saddle point and find a local minimizer

$$\bar{\boldsymbol{q}} = \arg \min_{\boldsymbol{q} \in \mathbb{S}^{k-1}} \varphi(\boldsymbol{q}).$$

3: Set $\bar{\boldsymbol{a}} = \mathcal{P}_{\mathbb{S}} \left[\left(\boldsymbol{Y}\boldsymbol{Y}^T \right)^{1/2} \bar{\boldsymbol{q}} \right]$.

Corollary 2.3. *Suppose the ground truth $\boldsymbol{a}_0$ kernel has preconditioned shift coherence $0 \le \mu \le \frac{1}{8\times 48} \log^{-3/2}(k)$ and sparse coefficient $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^m$. There exist positive constants $C \ge 2560^4$ and C' such that whenever the sparsity level*

$$\begin{aligned} 64k^{-1} \log k \le \theta \le \min \big\{ & \tfrac{1}{48^2} \mu^{-2} k^{-1} \log^{-2} k, \\ & \left(\tfrac{1}{4} - \tfrac{640}{C^{1/4}} \right) \left(3C_\star \mu \kappa^2 \right)^{-2/3} k^{-1} \left(1 + 36\mu^2 k \log k \right)^{-2} \big\}, \end{aligned}$$

and signal length

$$\begin{aligned} m \ \ge \ & \max \big\{ C\theta^2 \sigma_{\min}^{-2} \kappa^6 k^3 \left(1 + 36\mu^2 k \log k \right)^4 \log(\kappa k), \\ & C' (1-\theta)^{-2} \sigma_{\min}^{-2} \min \left\{ \mu^{-1}, \kappa^2 k^2 \right\} \kappa^8 k^4 \log^3(\kappa k) \big\}, \end{aligned}$$

then with high probability, Algorithm 1 recovers $\bar{\boldsymbol{a}}$ such that

$$\left\| \bar{\boldsymbol{a}} \pm \mathcal{P}_{\mathbb{S}} \left[\boldsymbol{\iota}_k s_\tau [\widetilde{\boldsymbol{a}_0}] \right] \right\|_2 \le 4\sqrt{c_\star} + ck^{-1} \tag{2.18}$$

for some integer shift $-(k-1) \le \tau \le k-1$.

For a generic $\boldsymbol{a}_0 \in \mathbb{S}^{k-1}$, plugging in the numerical estimation of the parameters $\sigma_{\min}$, κ and μ (Figure 3), accurate recovery can be obtained with $m \gtrsim \theta^2 k^6 \operatorname{poly}\log(k)$ measurements and sparsity level $\theta \lesssim k^{-2/3}\operatorname{poly}\log(k)$. For bandpass kernels $\boldsymbol{a}_0$, $\sigma_{\min}$ is smaller and κ, μ are larger, and so our results require $\boldsymbol{x}_0$ to be longer and sparser.

3 Asymptotic Function Landscape

In the next two sections, we discuss some key elements of our analysis. In this section, we first investigate the stationary points of the "population" objective $\mathbb{E}_{\boldsymbol{x}_0}[\psi(\boldsymbol{q})]$. We demonstrate that any local minimizer in $\mathcal{R}_{C_\star}$ is close to a signed column of $\boldsymbol{A}$, a preconditioned shift truncation of $\boldsymbol{a}_0$. In the next section, we then demonstrate that when m is sufficiently large, the "finite sample" objective $\psi(\boldsymbol{q})$ satisfies the same property.

In Section 3.1, we show how to accurately estimate the vector $\boldsymbol{\zeta} = \boldsymbol{A}^T\boldsymbol{q}$ at any stationary point $\boldsymbol{q} \in \mathcal{R}_{C_\star}$. In Section 3.2, we show how the number of spikes in $\boldsymbol{\zeta}$ determines the geometry around a stationary point.

- For any stationary point $\boldsymbol{q} \in \mathcal{R}_{C_\star}$, its preconditioned cross-correlation $\boldsymbol{\zeta}$ has *at least* one large entry (Section 3.2.1). This implies that any stationary point $\boldsymbol{q}$ must be close some local minimizer.
- If $\boldsymbol{\zeta}$ has *only* one large entry, then $\boldsymbol{q}$ is a local minimizer. (Section 3.2.2)
- If $\boldsymbol{\zeta}$ has *more than* one large entry, then $\boldsymbol{q}$ is a strict saddle point. (Section 3.2.3)

With above three characterizations, we can deduce that any local minimizer in $\mathcal{R}_{C_\star}$ is close to some column of $\boldsymbol{A}$, a preconditioned shift truncation of $\boldsymbol{a}_0$.

3.1 Stationary Points

Using $\mathbb{E}_{\boldsymbol{x}_0\sim\text{i.i.d.BG}(\theta)}[\boldsymbol{Y}\boldsymbol{Y}^T] = \theta m \boldsymbol{A}_0\boldsymbol{A}_0^T$ again, the expectation of the objective function $\psi(\boldsymbol{q})$ can be approximated[9] as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_0\sim\text{i.i.d.BG}(\theta)}[\psi(\boldsymbol{q})] &\approx \mathbb{E}_{\boldsymbol{x}_0\sim\text{i.i.d.BG}(\theta)}\left[-\frac{1}{m}\left\|\boldsymbol{Y}^T\left(\theta m\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{q}\right\|_4^4\right] \\ &= -\frac{1}{\theta^2 m^2}\left[3\theta(1-\theta)\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4 + 3\theta^2\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_2^4\right] \\ &= -\frac{3(1-\theta)}{\theta m^2}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4 - \frac{3}{m^2}. \end{aligned} \tag{3.1}$$

In the next section, we will argue that the critical points of the finite sample objective $\psi(\boldsymbol{q})$ are close to those of the asymptotic approximation ϕ. We can therefore study the critical points of ψ by studying the simpler problem

$$\min_{\boldsymbol{q}\in\mathbb{R}^{k-1}} \varphi(\boldsymbol{q}) \doteq -\frac{1}{4}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4 = -\frac{1}{4}\|\boldsymbol{\zeta}\|_4^4. \tag{3.2}$$

The Euclidean gradient and Hessian for $\varphi(\boldsymbol{q})$ can be calculated as

$$\nabla\varphi(\boldsymbol{q}) = -\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3}, \tag{3.3}$$

$$\nabla^2\varphi(\boldsymbol{q}) = -3\boldsymbol{A}\operatorname{diag}\left(\boldsymbol{\zeta}^{\circ 2}\right)\boldsymbol{A}^T. \tag{3.4}$$

We can study the critical points of φ over the sphere using the *Riemannian* gradient and hessian [AMS07]

$$\operatorname{grad}\varphi(\boldsymbol{q}) = \boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\nabla\varphi(\boldsymbol{q})\right] \tag{3.5}$$

$$= -\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} + \boldsymbol{q}\|\boldsymbol{\zeta}\|_4^4, \tag{3.6}$$

$$\operatorname{Hess}\varphi(\boldsymbol{q}) = \boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\nabla^2\varphi(\boldsymbol{q}) - \langle\nabla\varphi(\boldsymbol{q}),\boldsymbol{q}\rangle\boldsymbol{I}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp} \tag{3.7}$$

[9] In Lemma A.1 of the appendix, we give a detailed derivation of this approximation.

$$= -\boldsymbol{P}_{\boldsymbol{q}^\perp}\Big[3\boldsymbol{A}\,\mathrm{diag}(\boldsymbol{\zeta}^{\circ 2})\boldsymbol{A}^T - \|\boldsymbol{\zeta}\|_4^4\,\boldsymbol{I}\Big]\boldsymbol{P}_{\boldsymbol{q}^\perp}. \tag{3.8}$$

Here, $\boldsymbol{P}_{\boldsymbol{q}^\perp} = \boldsymbol{I} - \boldsymbol{q}\boldsymbol{q}^T$ denotes the projection onto the tangent space of the Frobenius sphere at point $\boldsymbol{q} \in \mathbb{S}^{k-1}$.

As in the Euclidean space, a stationary point on the sphere satisfies $\mathrm{grad}\,[\varphi]\,(\boldsymbol{q}) = \boldsymbol{0}$. Using (3.6), at any stationary point of φ,

$$\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} - \boldsymbol{q}\,\|\boldsymbol{\zeta}\|_4^4 = \boldsymbol{0}. \tag{3.9}$$

Left-multiplying both sides of the equation by $\boldsymbol{A}^T$, we have

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} - \boldsymbol{A}^T\boldsymbol{q}\,\|\boldsymbol{\zeta}\|_4^4 = \boldsymbol{0}. \tag{3.10}$$

For the i-th entry, following equality always holds

$$0 = \|\boldsymbol{a}_i\|_2^2\,\zeta_i^3 + \sum_{j\neq i}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\,\zeta_j^3 - \zeta_i\,\|\boldsymbol{\zeta}\|_4^4 \tag{3.11}$$

$$\Rightarrow \quad 0 = \zeta_i^3 - \zeta_i\underbrace{\frac{\|\boldsymbol{\zeta}\|_4^4}{\|\boldsymbol{a}_i\|_2^2}}_{\alpha_i} + \underbrace{\frac{\sum_{j\neq i}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\,\zeta_j^3}{\|\boldsymbol{a}_i\|_2^2}}_{\beta_i}. \tag{3.12}$$

For simplicity, we deploy the following notations

$$\alpha_i = \frac{\|\boldsymbol{\zeta}\|_4^4}{\|\boldsymbol{a}_i\|_2^2}, \quad \beta_i = \frac{\sum_{j\neq i}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\,\zeta_j^3}{\|\boldsymbol{a}_i\|_2^2}. \tag{3.13}$$

If $\alpha_i \gg \beta_i$, Proposition 3.1 shows that ζ_i is very close to one of three values: 0, or $\pm\sqrt{\alpha_i}$.

Proposition 3.1. *Let $\boldsymbol{q} \in \mathbb{S}^{k-1}$ be a stationary point satisfying $\|\boldsymbol{A}^T\boldsymbol{q}\|_4^6 \geq 4\mu\,\|\boldsymbol{A}^T\boldsymbol{q}\|_3^3$, then the i-th entry of $\boldsymbol{\zeta} = \boldsymbol{A}^T\boldsymbol{q}$ falls in the range*

$$\{0, \pm\sqrt{\alpha_i}\} \pm \frac{2\beta_i}{\alpha_i}, \tag{3.14}$$

with

$$\alpha_i = \frac{\|\boldsymbol{\zeta}\|_4^4}{\|\boldsymbol{a}_i\|_2^2}, \quad \beta_i = \frac{\sum_{j\neq i}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\,\zeta_j^3}{\|\boldsymbol{a}_i\|_2^2}. \tag{3.15}$$

Proof Since $\|\boldsymbol{\zeta}\|_4^6 \geq 4\mu\,\|\boldsymbol{\zeta}\|_3^3$, in this case, $\beta_i \leq \frac{1}{4}\alpha_i^{3/2}$ is satisfied, as

$$\|\boldsymbol{\zeta}\|_4^6 \geq 4\mu\,\|\boldsymbol{\zeta}\|_3^3 \geq 4\,\|\boldsymbol{a}_i\|_2\sum_{i\neq j}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\,\zeta_j^3. \tag{3.16}$$

The roots can be estimated by applying Lemma A.2 with

$$\sqrt{\alpha_i} = \frac{\|\boldsymbol{\zeta}\|_4^2}{\|\boldsymbol{a}_i\|_2}, \tag{3.17}$$

$$\frac{2\beta_i}{\alpha_i} = \frac{2\sum_{j\neq i}\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle\,\zeta_j^3}{\|\boldsymbol{\zeta}\|_4^4} \leq \frac{2\mu\,\|\boldsymbol{\zeta}\|_3^3}{\|\boldsymbol{\zeta}\|_4^4}. \tag{3.18}$$

■

This implies that either $|\langle \boldsymbol{a}_i, \boldsymbol{q}\rangle|$ is large ($\approx \sqrt{\alpha_i}$) or it is very close to zero.

3.2 Function Landscape on $\mathcal{R}_{C_\star}$

In this section, we study the optimization landscape around a stationary point $\boldsymbol{q}$ by bounding the eigenvalues of the Riemannian Hessian $\mathrm{Hess}\left[\varphi\right](\boldsymbol{q})$: if $\mathrm{Hess}\left[\varphi\right](\boldsymbol{q})$ is positive semidefinite, then the φ is convex in a neighborhood of $\boldsymbol{q}$ and hence $\boldsymbol{q}$ is a local minimum; if $\mathrm{Hess}\left[\varphi\right](\boldsymbol{q})$ has a negative eigenvalue, then there exists a direction along which the objective value decreases and hence $\boldsymbol{q}$ is a saddle point.

Note that the Riemannian Hessian $\mathrm{Hess}\left[\varphi\right](\boldsymbol{q})$ at stationary point $\boldsymbol{q}$ is a function of $\boldsymbol{\zeta}$ which can be accurately estimated when constrained in $\mathcal{R}_{C_\star}$ with $C_\star \geq 10$. By plugging the estimation of $\boldsymbol{\zeta}$ in the Riemannian Hessian, we can bound the eigenvalues of $\mathrm{Hess}\left[\varphi\right](\boldsymbol{q})$, and hence we can characterize the optimization landscape around a stationary point $\boldsymbol{q}$.

3.2.1 Nontrivial Preference of a Stationary Point

First, we demonstrate that for any stationary point $\boldsymbol{q} \in \mathcal{R}_{C_\star}$ with $C_\star \geq 10$, $\boldsymbol{\zeta}$ must have at least one large entry.

Lemma 3.2. *For any stationary point $\boldsymbol{q} \in \mathcal{R}_{C_\star}$ with $C_\star \geq 10$,*

$$\|\boldsymbol{\zeta}\|_\infty \geq \frac{2\mu \|\boldsymbol{\zeta}\|_3^3}{\|\boldsymbol{\zeta}\|_4^4}. \tag{3.19}$$

Proof We give a proof by contradiction. Suppose that $\boldsymbol{q} \in \mathcal{R}_{C_\star}$ with $C_\star \geq 10$, and every entry of $\boldsymbol{\zeta}$ has small magnitude such that $\|\boldsymbol{\zeta}\|_\infty < 2\mu \|\boldsymbol{\zeta}\|_3^3 / \|\boldsymbol{\zeta}\|_4^4$, then

$$\|\boldsymbol{\zeta}\|_4^4 \leq \|\boldsymbol{\zeta}\|_\infty^2 \leq \left(\frac{2\beta_i}{\alpha_i}\right)^2 \leq \frac{4\mu^2 \|\boldsymbol{\zeta}\|_3^6}{\|\boldsymbol{\zeta}\|_4^8}, \tag{3.20}$$

which indicates $\|\boldsymbol{\zeta}\|_4^6 \leq 2\mu \|\boldsymbol{\zeta}\|_3^3$ and contradicts the assumption $\|\boldsymbol{\zeta}\|_4^6 > C_\star \mu\kappa^2 \|\boldsymbol{\zeta}\|_3^3$. Therefore, at least one entry of $\boldsymbol{\zeta}$ has large enough magnitude. ∎

Geometrically, the nontrivial entry ζ_i indicates the preference to corresponding column $\boldsymbol{a}_i$, as $\zeta_i = \langle \boldsymbol{a}_i, \boldsymbol{q} \rangle$. Therefore, Lemma 3.2 implies that any stationary point $\boldsymbol{q}$ in $\mathcal{R}_{C_\star}$ should be close to at least one column of $\boldsymbol{A}$.

3.2.2 Local Minima

Suppose $\boldsymbol{q} \in \mathcal{R}_{C_\star}$ ($C_\star \geq 10$) is a stationary point and vector $\boldsymbol{\zeta}$ only has one nontrivial entry ζ_l, then we can demonstrate that the Riemannian Hessian $\mathrm{Hess}\,\varphi(\boldsymbol{q})$ is positive definite, and hence $\boldsymbol{q}$ is a local minimizer near $\boldsymbol{a}_l$.

Lemma 3.3. *Suppose $\boldsymbol{q}$ is a stationary point in $\mathcal{R}_{C_\star}$ with $C_\star \geq 10$, and $\boldsymbol{\zeta} = \boldsymbol{A}^T\boldsymbol{q}$ has only one entry ζ_l of magnitude no smaller than $2\mu \|\boldsymbol{\zeta}\|_3^3 / \|\boldsymbol{\zeta}\|_4^4$. Then $\boldsymbol{q}$ is a local minimum near $\boldsymbol{a}_l$ and $|\langle \boldsymbol{q}, \mathcal{P}_{\mathbb{S}}[\boldsymbol{a}_l] \rangle| > 1 - 2c_\star\kappa^{-2}$ with $c_\star = 1/C_\star$.*

Proof Suppose $\boldsymbol{\zeta}$ has only one big entry ζ_l, and other entries are bounded by $2\beta_l/\alpha_l$

$$\|\boldsymbol{\zeta}\|_4^4 = \zeta_l^4 + \sum_{j\neq l} \zeta_j^4 \tag{3.21}$$

$$\leq \zeta_l^4 + \max_{j\neq l} \zeta_j^2 \cdot \sum_{j\neq l} \zeta_j^2 \tag{3.22}$$

$$\leq \zeta_l^4 + \frac{4\mu^2 \|\boldsymbol{\zeta}\|_3^6}{\|\boldsymbol{\zeta}\|_4^8}, \tag{3.23}$$

with $\|\boldsymbol{\zeta}\|_4^6 \geq C_\star \mu\kappa^2 \|\boldsymbol{\zeta}\|_3^3$, and for simplicity let $c_\star = 1/C_\star$, we have

$$\zeta_l^4 \geq \|\boldsymbol{\zeta}\|_4^4 - \frac{4\mu^2 \|\boldsymbol{\zeta}\|_3^6}{\|\boldsymbol{\zeta}\|_4^8} \geq \left(1 - 4c_\star^2\kappa^{-4}\right) \|\boldsymbol{\zeta}\|_4^4. \tag{3.24}$$

On the other hand, we also have

$$\zeta_l^2 \le \left(\sqrt{\alpha_l} + \frac{2\beta_l}{\alpha_l}\right)^2 \tag{3.25}$$

$$\le \frac{\|\boldsymbol{\zeta}\|_4^4}{\|\boldsymbol{a}_i\|_2^2} + \frac{4\mu\|\boldsymbol{\zeta}\|_3^3}{\|\boldsymbol{a}_i\|_2\|\boldsymbol{\zeta}\|_4^2} + \frac{4\mu^2\|\boldsymbol{\zeta}\|_3^6}{\|\boldsymbol{\zeta}\|_4^8} \tag{3.26}$$

$$\le \frac{\|\boldsymbol{\zeta}\|_4^4}{\|\boldsymbol{a}_i\|_2^2}\left(1 + 4c_\star\kappa^{-2} + 4c_\star^2\kappa^{-4}\right). \tag{3.27}$$

Combining above two inequalities, we have

$$\zeta_l^2 \le \frac{1 + 4c_\star\kappa^{-2} + 4c_\star^2\kappa^{-4}}{1 - 4c_\star^2\kappa^{-4}} \frac{\zeta_l^4}{\|\boldsymbol{a}_i\|_2^2}, \tag{3.28}$$

thus the local minimum $\boldsymbol{q}$ is close to $\boldsymbol{a}_l$:

$$\frac{|\langle \boldsymbol{q}, \boldsymbol{a}_l\rangle|}{\|\boldsymbol{a}_l\|_2} \ge \frac{\sqrt{1 - 4c_\star^2\kappa^{-4}}}{1 + 2c_\star\kappa^{-2}} \ge 1 - 2c_\star\kappa^{-2}. \tag{3.29}$$

Next, we need to verify that the Riemannian Hessian at $\bar{\boldsymbol{q}}$ is definite positive, recall that

$$\operatorname{Hess}\varphi(\boldsymbol{q}) = -\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[3\boldsymbol{A}\operatorname{diag}(\boldsymbol{\zeta}^{\circ 2})\boldsymbol{A}^T - \|\boldsymbol{\zeta}\|_4^4\boldsymbol{I}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}. \tag{3.30}$$

Let $\boldsymbol{v}$ be a unit vector such that $\boldsymbol{v} \perp \boldsymbol{q}$, then

$$\boldsymbol{v}^T \operatorname{Hess}\varphi(\boldsymbol{q})\boldsymbol{v} \tag{3.31}$$

$$= -\boldsymbol{v}^T\left(3\boldsymbol{A}\operatorname{diag}(\boldsymbol{\zeta}^{\circ 2})\boldsymbol{A}^T - \|\boldsymbol{\zeta}\|_4^4\boldsymbol{I}\right)\boldsymbol{v} \tag{3.32}$$

$$= \|\boldsymbol{\zeta}\|_4^4 - 3\boldsymbol{v}^T\boldsymbol{A}\operatorname{diag}(\boldsymbol{\zeta}^{\circ 2})\boldsymbol{A}^T\boldsymbol{v} \tag{3.33}$$

$$= \|\boldsymbol{\zeta}\|_4^4 - 3\langle \boldsymbol{a}_l, \boldsymbol{v}\rangle^2\zeta_l^2 - 3\sum_{i\neq l}\langle \boldsymbol{a}_i, \boldsymbol{v}\rangle^2\zeta_i^2 \tag{3.34}$$

$$\ge \|\boldsymbol{\zeta}\|_4^4 - 3\langle \boldsymbol{a}_l, \boldsymbol{v}\rangle^2\zeta_l^2 - 3\max_{i\neq l}\zeta_i^2. \tag{3.35}$$

The last inequality is due to $\sum_{i\neq l}\langle \boldsymbol{a}_i, \boldsymbol{v}\rangle^2 \le \left\|\boldsymbol{A}^T\boldsymbol{v}\right\|_2^2 = 1$. Since $\boldsymbol{v} \perp \bar{\boldsymbol{q}}$ and ζ_l is the only entry with nontrivial magnitude, then derive from (3.29):

$$\langle \boldsymbol{a}_l, \boldsymbol{v}\rangle^2\zeta_l^2 \le 2c_\star\|\boldsymbol{a}_l\|_2^2\left(\sqrt{\alpha_l} + \frac{2\beta_l}{\alpha_l}\right)^2 \tag{3.36}$$

$$\le 2c_\star\|\boldsymbol{a}_l\|_2^2\cdot(1 + 2c_\star)^2\alpha_l \tag{3.37}$$

$$\le 2c_\star\left(1 + 2c_\star^2\right)^2\|\boldsymbol{\zeta}\|_4^4, \tag{3.38}$$

and

$$\max_{i\neq l}\zeta_i^2 \le \frac{4\beta^2}{\alpha^2} \le \frac{4\mu^2\|\boldsymbol{\zeta}\|_3^6}{\|\boldsymbol{\zeta}\|_4^8} \le \frac{4c_\star^2\|\boldsymbol{\zeta}\|_4^{12}}{\|\boldsymbol{\zeta}\|_4^8} \le 4c_\star^2\|\boldsymbol{\zeta}\|_4^4. \tag{3.39}$$

Hence, the inequality $\boldsymbol{v}^T\operatorname{Hess}\varphi(\boldsymbol{q})\boldsymbol{v} \ge \left(1 - 6c_\star - 36c_\star^2 - 24c_\star^3\right)\|\boldsymbol{\zeta}\|_4^4$ holds for any $\boldsymbol{v}$ satisfying $\boldsymbol{v} \perp \boldsymbol{q}$, thus implies positive curvature along any tangent direction at such stationary point $\boldsymbol{q}$ when $C_\star \ge 10$.

■

The lemma says if $\boldsymbol{q}$ is a stationary point in $\boldsymbol{R}_{C_\star}$ and $\boldsymbol{q}$ is only close to one column $\boldsymbol{a}_l$, then $\boldsymbol{q}$ is a local minimizer and satisfies $|\langle \boldsymbol{q}, \mathcal{P}_{\mathbb{S}}[\boldsymbol{a}_l]\rangle| > 1 - 2c_\star\kappa^{-2}$ with $c_\star = 1/C_\star$.

3.2.3 Saddle Points

At last, if $\boldsymbol{q} \in \mathcal{R}_{C_\star}$ ($C_\star \geq 10$) is a stationary point and vector $\boldsymbol{\zeta}$ has more than one nontrivial entry. Denote any two nontrivial entries of $\boldsymbol{\zeta}$ with ζ_l and $\zeta_{l'}$, then we can prove that the Riemannian Hessian $\operatorname{Hess}\varphi(\boldsymbol{q})$ has negative curvature in the span of $\boldsymbol{a}_l$ and $\boldsymbol{a}_{l'}$, hence $\boldsymbol{q}$ is a saddle point.

Lemma 3.4. *Suppose $\boldsymbol{q}$ is a stationary point in $\mathcal{R}_{C_\star}$ with $C_\star \geq 10$, and $\boldsymbol{\zeta} = \boldsymbol{A}^T\boldsymbol{q}$ has at least two entries ζ_l and $\zeta_{l'}$ with magnitude magnitude $\geq 2\mu \left\|\boldsymbol{\zeta}\right\|_3^3 / \left\|\boldsymbol{\zeta}\right\|_4^4$, then the Riemannian Hessian at $\boldsymbol{q}$ has at least one negative eigenvalue and $\boldsymbol{q}$ is a saddle point.*

Proof Suppose $\boldsymbol{\zeta}$ has at least two big entries ζ_l and $\zeta_{l'}$ satisfying

$$\zeta_l^2 \geq \left(\sqrt{\alpha_l} - \frac{2\beta_l}{\alpha_l}\right)^2 \tag{3.40}$$

$$\geq \frac{\left\|\boldsymbol{\zeta}\right\|_4^4}{\left\|\boldsymbol{a}_l\right\|_2^2} - \frac{4\mu\left\|\boldsymbol{\zeta}\right\|_3^3}{\left\|\boldsymbol{\zeta}\right\|_4^2\left\|\boldsymbol{a}_l\right\|_2} + \frac{4\mu^2\left\|\boldsymbol{\zeta}\right\|_3^6}{\left\|\boldsymbol{\zeta}\right\|_4^8} \tag{3.41}$$

$$> \frac{\left\|\boldsymbol{\zeta}\right\|_4^4}{\left\|\boldsymbol{a}_l\right\|_2^2} - \frac{4\mu\left\|\boldsymbol{\zeta}\right\|_3^3}{\left\|\boldsymbol{\zeta}\right\|_4^2\left\|\boldsymbol{a}_l\right\|_2}, \tag{3.42}$$

and $\zeta_{\ell'}$ likewise. Since the nontrivial entry $\zeta_l = \langle \boldsymbol{a}_l, \boldsymbol{q}\rangle$, and again let $c_\star = 1/C_\star$, it is easy to show that the norm of $\boldsymbol{a}_l$ is sufficiently large:

$$\left\|\boldsymbol{a}_l\right\|_2^2 \geq \zeta_l^2 \geq \left(\sqrt{\alpha_l} - \frac{2\beta_l}{\alpha_l}\right)^2 \tag{3.43}$$

$$\geq (1-2c_\star)^2\,\frac{\left\|\boldsymbol{\zeta}\right\|_4^4}{\left\|\boldsymbol{a}_l\right\|_2^2} \tag{3.44}$$

$$\geq (1-c_\star)^2\, C_\star^{2/3}\frac{\mu^{2/3}\left\|\boldsymbol{\zeta}\right\|_3^2}{\left\|\boldsymbol{a}_l\right\|_2^2}, \tag{3.45}$$

or

$$\left\|\boldsymbol{a}_l\right\|_2 \geq (1-c_\star)^{1/2}\, C_\star^{1/6}\mu^{1/6}\left\|\boldsymbol{\zeta}\right\|_3^{1/2}. \tag{3.46}$$

Similar result holds for $\left\|\boldsymbol{a}_{l'}\right\|_2$, therefore

$$\frac{\mu}{\left\|\boldsymbol{a}_l\right\|_2\left\|\boldsymbol{a}_{l'}\right\|_2} \leq \frac{\mu^{2/3}}{C_\star^{1/3}\left\|\boldsymbol{\zeta}\right\|_3} \leq \frac{C_\star^{-2/3}\left\|\boldsymbol{\zeta}\right\|_4^4}{C_\star^{1/3}\left\|\boldsymbol{\zeta}\right\|_3^3} \leq c_\star. \tag{3.47}$$

Now we are ready to show there exists a unit vector $\boldsymbol{v}$ such that $\boldsymbol{v} \in \operatorname{span}(\boldsymbol{a}_l, \boldsymbol{a}_{l'})$ and $\boldsymbol{v} \perp \boldsymbol{q}$, and the Hessian has negative curvature along such $\boldsymbol{v}$:

$$\begin{aligned}
&\boldsymbol{v}^T \operatorname{Hess}\varphi(\boldsymbol{q})\boldsymbol{v} \\
&= -3\boldsymbol{v}^T\boldsymbol{A}\operatorname{diag}(\boldsymbol{\zeta}^2)\boldsymbol{A}^T\boldsymbol{v} + \left\|\boldsymbol{\zeta}\right\|_4^4 && (3.48)\\
&\leq -3\boldsymbol{v}^T\left(\boldsymbol{a}_l\zeta_l^2\boldsymbol{a}_l^T + \boldsymbol{a}_{l'}\zeta_{l'}^2\boldsymbol{a}_{l'}^T\right)\boldsymbol{v} + \left\|\boldsymbol{\zeta}\right\|_4^4 && (3.49)\\
&< -3\left(\left|\left\langle \frac{\boldsymbol{a}_l}{\left\|\boldsymbol{a}_l\right\|_2}, \boldsymbol{v}\right\rangle\right|^2 + \left|\left\langle \frac{\boldsymbol{a}_{l'}}{\left\|\boldsymbol{a}_{l'}\right\|_2}, \boldsymbol{v}\right\rangle\right|^2\right)\left\|\boldsymbol{\zeta}\right\|_4^4 \\
&\quad + \frac{4\mu\left\|\boldsymbol{\zeta}\right\|_3^3}{\left\|\boldsymbol{\zeta}\right\|_4^2}\left(\left\|\boldsymbol{a}_l\right\|_2 + \left\|\boldsymbol{a}_{l'}\right\|_2\right) + \left\|\boldsymbol{\zeta}\right\|_4^4 && (3.50)\\
&< -3\left(1 - \frac{\mu}{\left\|\boldsymbol{a}_l\right\|_2\left\|\boldsymbol{a}_{l'}\right\|_2}\right)\left\|\boldsymbol{\zeta}\right\|_4^4
\end{aligned}$$

$$+\frac{4\mu\left\|\boldsymbol{\zeta}\right\|_3^3}{\left\|\boldsymbol{\zeta}\right\|_4^2}\left(\left\|\boldsymbol{a}_l\right\|_2+\left\|\boldsymbol{a}_{l'}\right\|_2\right)+\left\|\boldsymbol{\zeta}\right\|_4^4 \tag{3.51}$$

$$\leq\left(-2+11c_\star\right)\left\|\boldsymbol{\zeta}\right\|_4^4. \tag{3.52}$$

The third inequality is implied by Lemma A.3 and is negative when $C_\star \geq 10$.

■

This lemma says if the stationary point $\boldsymbol{q}$ has large inner product with any two columns $\boldsymbol{a}_l$ and $\boldsymbol{a}_{l'}$, then this $\boldsymbol{q}$ is a saddle point and the objective value decreases along the direction that breaks symmetry between $\boldsymbol{a}_l$ and $\boldsymbol{a}_{l'}$. The saddle point $\boldsymbol{q}$ can be seen as resulting from the competition between the two target solutions $\boldsymbol{a}_l$ and $\boldsymbol{a}_{l'}$.

4 Large Sample Concentration

In this section, we argue that the geometric characteristics of $\psi\left(\boldsymbol{q}\right)$ are similar to those of $\varphi\left(\boldsymbol{q}\right)$, by demonstrating that the critical points of the finite sample objective function $\psi(\boldsymbol{q})$ are similar to those of the asymptotic objective function $\varphi(\boldsymbol{q})$:

- **Critical points are close.** The Riemannian gradient (Lemma 4.2) and Hessian (Lemma 4.3) concentrate, such that there is a bijection between critical points $\boldsymbol{q}_\varphi$ of φ and critical points $\boldsymbol{q}_\psi$ of ψ, with $\left\|\boldsymbol{q}_\varphi-\boldsymbol{q}_\psi\right\|_2$ small.
- **Curvature is preserved.** The Riemannian Hessian (Lemma 4.3) concentrates, such that $\mathrm{Hess}[\psi](\boldsymbol{q}_{\mathrm{fs}})$ has a negative eigenvalue if and only if $\mathrm{Hess}[\varphi](\boldsymbol{q}_{\mathrm{pop}})$ has a negative eigenvalue, and $\mathrm{Hess}[\psi](\boldsymbol{q}_{\mathrm{fs}})$ is positive definite if and only if $\mathrm{Hess}[\varphi](\boldsymbol{q}_{\mathrm{pop}})$ is positive definite.

This implies that every local minimizer of the finite sample objective function is close to a preconditioned shift-truncation (Lemma 4.1).

Lemma 4.1. *If the following inequalities hold*

$$\begin{aligned}&\left\|\mathrm{grad}[\psi]\left(\boldsymbol{q}\right)-\frac{3\left(1-\theta\right)}{\theta m^2}\mathrm{grad}[\varphi]\left(\boldsymbol{q}\right)\right\|_2\\ &\quad\leq\quad\frac{3c_\star}{2\kappa^2}\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6,\end{aligned} \tag{4.1}$$

$$\begin{aligned}&\left\|\mathrm{Hess}[\psi]\left(\boldsymbol{q}\right)-\frac{3\left(1-\theta\right)}{\theta m^2}\mathrm{Hess}[\varphi]\left(\boldsymbol{q}\right)\right\|_2\\ &\quad\leq\quad3\left(1-6c_\star-36c_\star^2-24c_\star^3\right)\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4.\end{aligned} \tag{4.2}$$

for all $\boldsymbol{q}\in\mathcal{R}_{2C_\star}$ with $C_\star\geq 10$ and $c_\star=1/C_\star$, then any local minimum $\bar{\boldsymbol{q}}$ of $\psi\left(\boldsymbol{q}\right)$ in $\mathcal{R}_{2C_\star}$ satisfies $\left|\langle\bar{\boldsymbol{q}},\mathcal{P}_{\mathbb{S}}\left[\boldsymbol{a}_l\right]\rangle\right|\geq 1-2c_\star\kappa^{-2}$ for some index l.

Proof Please refer to Appendix B. ■

The Riemannian gradient and Hessian of the finite sample objective function $\psi\left(\boldsymbol{q}\right)$ have similar expressions as those of the asymptotic objective function $\varphi(\boldsymbol{q})$. Let $\boldsymbol{\eta}=\boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}\in\mathbb{S}^{m-1}$. Then

$$\psi\left(\boldsymbol{q}\right)=-\frac{1}{4m}\left\|\boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}\right\|_4^4=-\frac{1}{4m}\left\|\boldsymbol{\eta}\right\|_4^4, \tag{4.3}$$

we calculate the Euclidean gradient and Hessian of the objective function

$$\nabla\psi\left(\boldsymbol{q}\right)=-\frac{1}{m}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3}, \tag{4.4}$$

$$\nabla^2\psi(\boldsymbol{q}) = -\frac{3}{m}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\,\mathrm{diag}(\boldsymbol{\eta}^{\circ 2})\boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}. \tag{4.5}$$

Similarly, the Riemannian gradient and Hessian have the form

$$\begin{aligned}
&\mathrm{grad}[\psi](\boldsymbol{q}) = \boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\nabla\psi(\boldsymbol{q})\right] && (4.6)\\
&= -\frac{1}{m}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3} + \frac{1}{m}\boldsymbol{q}\left\|\boldsymbol{\eta}\right\|_4^4, && (4.7)\\
&\mathrm{Hess}[\psi](\boldsymbol{q}) = \boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\nabla^2\psi(\boldsymbol{q}) - \langle\nabla\psi(\boldsymbol{q}),\boldsymbol{q}\rangle\,\boldsymbol{I}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp} && (4.8)\\
&= \boldsymbol{P}_{\boldsymbol{q}^\perp}\Big[\frac{3}{m}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\,\mathrm{diag}(\boldsymbol{\eta}^{\circ 2})\boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\\
&\quad + \frac{1}{m}\left\|\boldsymbol{\eta}\right\|_4^4\boldsymbol{I}\Big]\boldsymbol{P}_{\boldsymbol{q}^\perp}. && (4.9)
\end{aligned}$$

Since $\boldsymbol{Y} = \boldsymbol{A}_0\boldsymbol{X}_0$, we can see that the Riemannian gradient and Hessian are (complicated) functions of the random circulant matrix $\boldsymbol{X}_0$. Although the entries of the vector $\boldsymbol{x}_0$ are probabilistically independent, the entries of $\boldsymbol{X}_0$ are dependent random variables. To remove the dependence within the random circulant matrix $\boldsymbol{X}_0$, we break $\boldsymbol{X}_0$ into submatrices $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{2k-1}$ that

$$\boldsymbol{X}_i = \left[\boldsymbol{x}_i, \boldsymbol{x}_{i+(2k-1)}, \cdots, \boldsymbol{x}_{i+(m-2k-1)}\right]. \tag{4.10}$$

Each of which is (marginally) distributed as a $(2k-1)\times\frac{m}{2k-1}$ i.i.d. $\mathrm{BG}(\theta)$ random matrix. Indeed, there exists a permutation $\boldsymbol{\Pi}$ such that

$$\boldsymbol{X}_0\boldsymbol{\Pi} = \left[\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_{2k-1}\right]. \tag{4.11}$$

A detailed analysis of (4.7)-(4.9) (see Appendix E and Appendix F in the Appendix) allows us to control the finite sample fluctuations of the gradient and Hessian in terms of analogous quantities for each $\boldsymbol{X}_i$. Because the $\boldsymbol{X}_i$ are i.i.d., they are amenable to standard tools from measure concentration. Taking a union bound over i, we show that the gradient (Lemma Lemma 4.2) and hessian (Lemma Lemma 4.3) concentrate as desired:

Lemma 4.2. *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^m$. There exists positive constant C that whenever*

$$m \geq C\frac{\min\left\{(2C_\star\mu)^{-1}, \kappa^2k^2\right\}}{(1-\theta)^2\sigma_{\min}^2}\kappa^8k^4\log^3(\kappa k), \tag{4.12}$$

and $\theta \geq 1/k$, then with probability no smaller than $1-\exp(-k)-\theta^2(1-\theta)^2k^{-4}-2\exp(-\theta k)-48k^{-7}-48m^{-5}-24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right)$,

$$\left\|\mathrm{grad}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\mathrm{grad}[\varphi](\boldsymbol{q})\right\|_2 \leq c\frac{1-\theta}{\theta m^2}\frac{\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6}{\kappa^2}, \tag{4.13}$$

holds for all $\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}$ with $c \leq 3/(2C_\star) \leq \frac{3}{20}$.

Proof Please refer to section E. ■

Lemma 4.3. *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$. There exists positive constant C that whenever*

$$m \geq C\frac{\min\left\{\left(2C_\star\mu\kappa^2\right)^{-4/3}, k^2\right\}}{(1-\theta)^2\sigma_{\min}^2}\kappa^6k^4\log^3(\kappa k), \tag{4.14}$$

and $\theta \geq 1/k$, then with probability no smaller than $1-\exp(-k)-\theta^2(1-\theta)^2k^{-4}-2\exp(-\theta k)-48k^{-7}-48m^{-5}-24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right)$,

$$\left\|\mathrm{Hess}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\mathrm{Hess}[\varphi](\boldsymbol{q})\right\|_2 \leq c\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4, \tag{4.15}$$

holds for all $\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}$ with positive constant $c \leq 0.048 \leq 3\left(1-6c_\star-36c_\star^2-24c_\star^3\right)$.

Proof Please refer to section F. ■

5 Experiments

5.1 Properties of a Random Kernel

Our results are stated in terms of several parameters, including the condition number κ of $\boldsymbol{A}_0$ and the column coherence of $\boldsymbol{A}$. In Figure 3, we demonstrate the typical values of σ_0, κ, and μ for generic unit-norm kernels of varying dimension $k = 10, 20, \cdots, 1000$.

From this figure, for a generic unit-norm kernel, we have following estimates:

$$\sigma_0 \approx \log^{-1}(k), \tag{5.1}$$

$$\kappa \approx \log^{4/3}(k), \tag{5.2}$$

$$\mu \approx \sqrt{\log(k)/k}. \tag{5.3}$$

On the other hand, if the kernel $\boldsymbol{a}_0$ is bandpass, then both κ and μ are larger. In this situation, our results require more observations m and smaller sparsity rate θ.

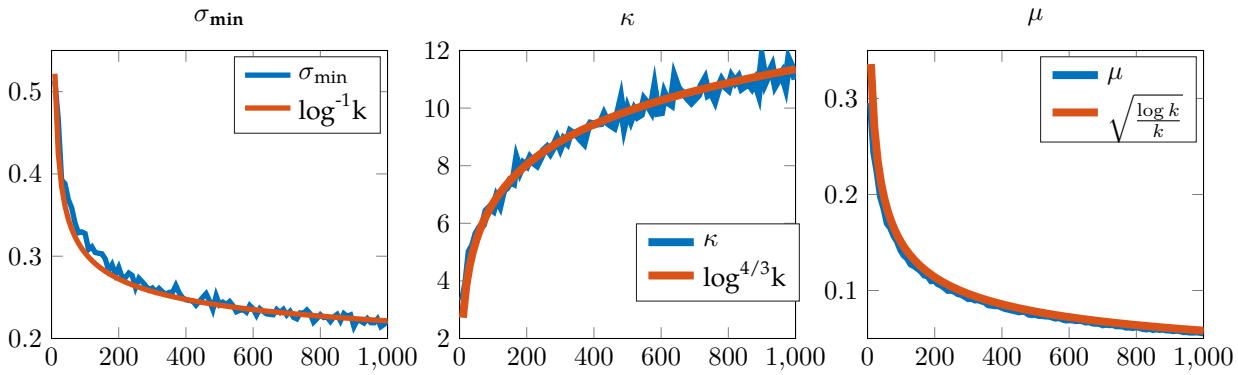

Figure 3: Average of Parameters $\sigma_{\min}$, κ, and μ of a random unit norm kernel $\boldsymbol{a}_0$ over 50 independent trials, as a function of dimension k.

5.2 Recovery Accuracy of Local Minima

We next investigate the performance of Algorithm 1 under varying settings. We define the recover error as $\text{err} = 1 - \max_\tau |\langle \bar{\boldsymbol{a}}, \mathcal{P}_{\mathbb{S}}[\boldsymbol{\iota}_k^* s_\tau[\widetilde{\boldsymbol{a}_0}]]\rangle|$, and calculate the average error from 50 independent experiments. In Figure 4, the left figure plots the average error when we fix the kernel size $k = 50$, and vary the dimension m and the sparsity θ of $\boldsymbol{x}_0$.[10] The right figure plots the average error when we vary the dimensions k, m of both convolution signals, and set the sparsity as $\theta = k^{-2/3}$.

This figure agrees with the theory developed in this paper: when the activation coefficient $\boldsymbol{x}_0$ is long and sparse (large m and small θ), the algorithm obtains a closer estimate of a shift-truncation of the ground truth.

5.3 Recovery Accuracy of the Ground Truth Kernel

In this section, we provide experiment results for the recovery of the ground truth kernel obtained by the annealing algorithm proposed in [ZLK+17]. The annealing algorithm recovers the ground truth kernel by minimizing the Lasso cost in (1.7), initialized at the zero-padded shift truncated kernel rendered from Algorithm

[10] Note that the x-axis is indexed with overlapping ratio $k \cdot \theta$, which indicates how many times the kernel $\boldsymbol{a}_0$ present in a k-length window of $\boldsymbol{y}$ on average.

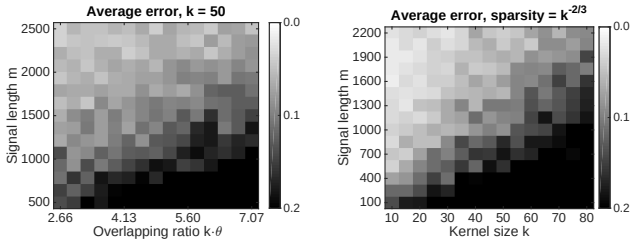

Figure 4: Recovery Error of the Shift Truncated Kernel of Algorithm 1.

Algorithm 1. The recovery accuracy presented in Figure Figure 6 is measured as $\mathrm{err} = \min_\tau \left\| \bar{\boldsymbol{a}}^{(+)} \pm s_\tau[\widetilde{\boldsymbol{a}_0}] \right\|_2$. Here, $\bar{\boldsymbol{a}}^{(+)}$ denote the local minimum in the lifted optimization space.

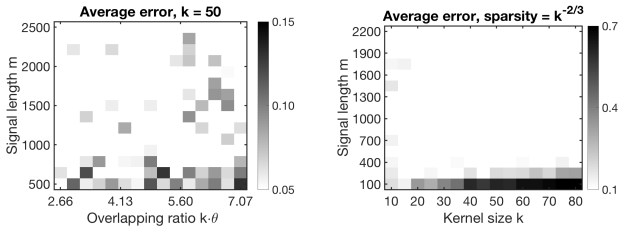

Figure 5: Recovery Error of the Ground Truth Kernel with Algorithm Algorithm 1 finding a shift truncated kernel and the annealing Lasso problem recovering the ground truth kernel.

For comparison, we also present experiment results of the algorithm proposed by [ZLK+17], which is composed of solving two Lasso minimization problems over the original kernel sphere and lifted kernel sphere respectively.

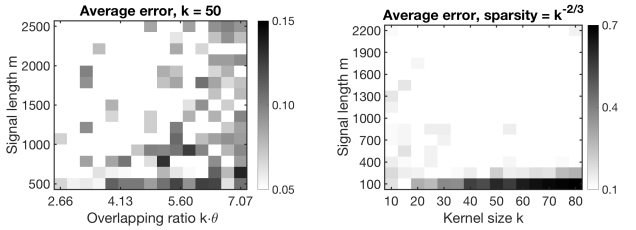

Figure 6: Recovery Error of the Ground Truth Kernel by minimizing the Lasso objective function recovering both the shift truncated kernel as well as the ground truth kernel.

In terms of the recovery accuracy of the ground truth kernel, Algorithm 1 proposed in this paper achieves better recovery for sparser and longer observations, while the [ZLK+17] manifests slight advantages when the observations is limited. As the optimization landscape studied in [ZLK+17] varies with different choice of sparsity parameter λ, it is possible that experiment results for [ZLK+17] could be improved. On the other hand, only empirical knowledge about the choice of λ is available while there is little disciplined understanding. In contrast, Algorithm 1 does not depend on any parameter tuning and guarantees recovery once the working conditions are met.

6 Discussions

Finally, we provide some comments about the results and proof strategy presented in this paper, and discuss directions for future research.

This paper casts the sparse blind deconvolution problem as finding a *spiky* vector in a subspace and studies its optimization landscape. We prove that the geometric property that *any local solution is close to a shift-truncation of the ground truth kernel* holds on a sub-level set of the sphere. This holds even when the observation contains densely overlapping copies of the true kernel. In addition, we propose a simple initialization scheme such that any descent algorithm that escapes strict saddles can recover the local minimum, which is a near shift-truncation of the ground truth kernel.

Sample Complexity. The sample complexity shown in this paper $m \sim k^6$ is suboptimal. Our proofs relies heavily on "worst case" tools such as the triangle inequality, multiplication of operator norm, and union bound. In particular, we believe that the sample complexity can be improved by replacing the sample splitting argument in Section Appendix E and Appendix F in the Appendix with more sophisticated arguments based on decoupling (see also [QZEW17]).

Global Geometry. The theoretical results presented in this paper demonstrate that "all local optima are benign" in the sub-level set $\mathcal{R}_{C_\star}$. Our empirical results suggest that this is a property holds over the whole sphere. Proving this could be challenging, as our characterization of the saddle points only applies when $\|\boldsymbol{\zeta}\|_4^4$ is large. It would be exciting to see if further research investigating other techniques for nonconvex optimization problems could be motivated by our current work.

Convolutional Dictionary Learning. This is a natural and practical extension of blind deconvolution, where the observation is the superposition of several convolutions. The empirical observations and algorithm proposed in [ZLK+17] hold in this more challenging situation. It would be interesting to develop efficient and provable algorithms for convolutional dictionary learning based on the ℓ^4 formulation.

Acknowledgement

The authors gratefully acknowledge support from NSF 1343282, NSF CCF 1527809, and NSF IIS 1546411. It is a great pleasure to acknowledge conversations with Yenson Lau, Sky Cheung, and Abhay Pasupathy.

References

[ABG07] P.-A. Absil, C.G. Baker, and K.A. Gallivan. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, Jul 2007.

[AMS07] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, USA, 2007.

[ARR12] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programing. *arXiv preprint:1211.5608*, 2012.

[Bha97] Rajendra Bhatia. Matrix analysis, 1997.

[BVG13] Alexis Benichoux, Emmanuel Vincent, and Remi Gribonval. A fundamental pitfall in blind deconvolution with sparse and shift-invariant priors. *38th International Conference on Acoustics, Speech, and Signal Processing*, May 2013.

[Chi16] Yuejie Chi. Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):782–794, June 2016.

[CLC+17] Sky Cheung, Yenson Lau, Zhengyu Chen, Ju Sun, Yuqian Zhang, John Wright, and Abhay Pasupathy. Beyond the fourier transform: A nonconvex optimization approach to microscopy analysis. *Submitted*, 2017.

[CM14] Sunav Choudhary and Urbashi Mitra. Fundamental limits of blind deconvolution part I: Ambiguity kernel. *ArXiv e-prints*, abs/1411.3810, November 2014.

[CM15] Sunav Choudhary and Urbashi Mitra. Fundamental limits of blind deconvolution part II: Sparsity-ambiguity trade-offs. *ArXiv e-prints*, abs/1503.03184, March 2015.

[CW98] Tony F. Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE Transactions on Image Processing*, 7(3):370–375, Mar 1998.

[DlPG99] Victor De la Pena and Evarist Giné. *Decoupling: from dependence to independence*. Springer, 1999.

[ETS11] Chaitanya Ekanadham, Daniel Tranchina, and Eero P. Simoncelli. A blind sparse deconvolution method for neural spike identification. In *Advances in Neural Information Processing Systems 24*, pages 1440–1448. 2011.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[GWY09] Donald Goldfarb, Zaiwen Wen, and Wotao Yin. A curvilinear search method for p-harmonic flows on spheres. *SIAM J. Imaging Sciences*, 2(1):84–109, 2009.

[HSSS16] Samuel B. Hopkins, Tselil Schrammand, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: Tensor decomposition and planted sparse vectors. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 178–191, 2016.

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

[KH96] Deepa Kundur and Dimitrios Hatzinakos. Blind image deconvolution. *Signal Processing Magazine, IEEE*, 13(3):43–64, May 1996.

[Lew98] Michael S. Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):53–78, 1998.

[LLB16] Yanjun Li, Kiryung Lee, , and Yoram Bresler. Identifiability in blind deconvolution with subspace or sparsity constraints. *IEEE Transaction of Information Theory*, 62(7):4266 – 4275, July 2016.

[LLB17] Yanjun Li, Kiryung Lee, and Yoram Bresler. Identifiability and stability in blind deconvolution under minimal assumptions. *IEEE Transaction of Information Theory*, 2017.

[LLJB17] Kiryung Lee, Yanjun Li, Marius Junge, and Yoram Bresler. Blind recovery of sparse signals from subsampled convolution. *IEEE Transaction of Information Theory*, 63(2):802–821, February 2017.

[LLSW16] Xiaodong Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. preprint, 2016.

[LS15] Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[LS17] Shuyang Ling and Thomas Strohmer. Blind deconvolution meets blind demixing: Algorithms and performance bounds. *IEEE Transactions on Information Theory*, 63(7):4497–4520, 2017.

[LWDF11] Anat Levin, Yair Weiss, Fredo Durand, and William T. Freeman. Understanding blind deconvolution algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2354–2367, Dec 2011.

[OJF+15] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, May 2015.

[PF14] Daniele Perrone and Paolo Favaro. Total variation blind deconvolution: The devil is in the details. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[QSW16] Qing Qu, Ju Sun, and John Wright. Finding a sparse vector in a subspace: linear sparsity using alternating directions. *IEEE Transactions on Information Theory*, 2016.

[QZEW17] Qing Qu, Yuqian Zhang, Yonina C. Eldar, and John Wright. Convolutional phase retrieval via gradient descent. preprint, 2017.

[SQW15] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere. preprint, 2015.

[SWW12] Daniel Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. preprint, 2012.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[WC16] L. Wang and Y. Chi. Blind deconvolution from multiple sparse inputs. *IEEE Signal Processing Letters*, 23(10):1384–1388, Oct 2016.

[WZ13] David Wipf and Haichao Zhang. Revisiting bayesian blind deconvolution. *arXiv preprint:1305.2362*, 2013.

[XRKM17] Peng Xu, Farbod Roosta-Khorasani, and Michael W. Mahoney. Second-order optimization for non-convex machine learning: An empirical study. *arXiv preprint arXiv:1708.07827*, 2017.

[ZLK+17] Yuqian Zhang, Yenson Lau, Han-Wen Kuo, Sky Cheung, Abhay Pasupathy, and John Wright. On the global geometry of sphere-constrained sparse blind deconvolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[ZWZ13] Haichao Zhang, David Wipf, and Yanning Zhang. Multi-image blind deblurring using a coupled adaptive sparse prior. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), January 2013.

Appendix

Appendix A contains some basic lemmas for quantities used repeatedly; Appendix B presents the proofs of the main theorem and corollary of this paper. Appendix C and Appendix D proves for lemmas around the initialization point $\boldsymbol{q}_{\text{init}}$ and the preconditioning term $\boldsymbol{Y}^T\boldsymbol{Y}$ (or $\boldsymbol{A}_0^T\boldsymbol{A}_0$) respectively. Finite sample concentration for the Riemannian gradient and Hessian are presented in Appendix E and Appendix F respectively.

A Basics

Lemma A.1 (Expectation of the Approximate Objective Function). *Assuming* $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^m$, *then*

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_0}\left[\frac{1}{m}\left\|\boldsymbol{Y}^T\left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{q}\right\|_4^4\right]\\&\quad=3\theta\left(1-\theta\right)\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4+3\theta^2\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_2^4.\end{aligned}\tag{A.1}$$

Proof Let $\boldsymbol{g}\in\mathbb{R}^{2k-1}$ be a standard random Gaussian vector and $\boldsymbol{P}_I$ be the projection operator onto Bernoulli vector $I\sim\text{Ber}(\theta)$ Then any column $\boldsymbol{x}_i\in\mathbb{R}^{2k-1}$ of $\boldsymbol{X}_0$ is equal in distribution to $\boldsymbol{x}_i=\boldsymbol{P}_I\boldsymbol{g}$ with $\boldsymbol{g}\sim_{\text{i.i.d.}}\mathcal{N}(0,1)$.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_0}\left[\frac{1}{m}\left\|\boldsymbol{Y}^T\left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{q}\right\|_4^4\right]\\&=\frac{1}{m}\mathbb{E}_I\mathbb{E}_{\boldsymbol{g}}\left\|\boldsymbol{q}^T\boldsymbol{A}\boldsymbol{X}_0\right\|_4^4 &\text{(A.2)}\\&=\mathbb{E}_I\mathbb{E}_{\boldsymbol{g}}\left\|\boldsymbol{q}^T\boldsymbol{A}\boldsymbol{x}_i\right\|_4^4 &\text{(A.3)}\\&=\mathbb{E}_I\mathbb{E}_{\boldsymbol{g}}\left(\boldsymbol{q}^T\boldsymbol{A}\boldsymbol{P}_I\boldsymbol{g}\right)^4 &\text{(A.4)}\\&=3\mathbb{E}_I\left(\boldsymbol{q}^T\boldsymbol{A}\boldsymbol{P}_I\boldsymbol{A}^T\boldsymbol{q}\right)^2 &\text{(A.5)}\\&=3\mathbb{E}_I\left(\sum_{i\in I}\left\langle\boldsymbol{a}_i,\boldsymbol{q}\right\rangle^4+\sum_{\{i\neq j\}\in I}\left\langle\boldsymbol{a}_i,\boldsymbol{q}\right\rangle^2\left\langle\boldsymbol{a}_j,\boldsymbol{q}\right\rangle^2\right) &\text{(A.6)}\\&=3\theta\left(1-\theta\right)\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4+3\theta^2\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_2^4 &\text{(A.7)}\end{aligned}$$

■

Lemma A.2 (Root Estimation for Cubic Gradient Function). *Consider an equation of the form*

$$f(x)=x\left(\alpha-x^2\right)-\beta=0,\tag{A.8}$$

with $\alpha>0$. *Suppose that* $\beta<\frac{1}{4}\alpha^{3/2}$. *Then* $f(x)=0$ *has three solutions,* x_1,x_2,x_3 *satisfying*

$$\max\left\{\left|x_1-\sqrt{\alpha}\right|,\left|x_2+\sqrt{\alpha}\right|,\left|x_3\right|\right\}\ \leq\ \frac{2\beta}{\alpha}.\tag{A.9}$$

Proof Suppose first that $\beta>0$. Then $f(0)<0$. Moreover,

$$\begin{aligned}f\left(\tfrac{2\beta}{\alpha}\right)&=2\beta-8\beta^3/\alpha^3-\beta &\text{(A.10)}\\&=\beta\left(1-8\beta^2/\alpha^3\right) &\text{(A.11)}\\&>0. &\text{(A.12)}\end{aligned}$$

Hence, f has at least one root in the interval $\left[0, \frac{2\beta}{\alpha}\right]$. Similarly, notice that $f\left(\sqrt{\alpha}\right) < 0$ and that

$$
\begin{aligned}
& f\left(\sqrt{\alpha} - \tfrac{2\beta}{\alpha}\right) \\
&= \alpha^{3/2} - 2\beta - \left(\sqrt{\alpha} - 2\beta/\alpha\right)^3 - \beta && \text{(A.13)} \\
&= \alpha^{3/2} - 3\beta - \alpha^{3/2} + 6\beta - 12\beta^2/\alpha^{3/2} + 8\beta^3/\alpha^3 && \text{(A.14)} \\
&= \beta\left(3 - \frac{12\beta}{\alpha^{3/2}} + \frac{8\beta^2}{\alpha^3}\right) && \text{(A.15)} \\
&> 0. && \text{(A.16)}
\end{aligned}
$$

Thus, there is at least one root in the interval $\left[\sqrt{\alpha} - \frac{2\beta}{\alpha}, \sqrt{\alpha}\right]$. Finally, note that $f\left(-\sqrt{\alpha}\right) < 0$, $\frac{df}{dx}\left(-\sqrt{\alpha}\right) = -2\alpha$, and $\frac{d^2 f}{dx^2}\left(x'\right) = -3x'$ is positive for $x' \leq -\sqrt{\alpha}$. Hence, convexity gives that

$$
\begin{aligned}
& f\left(-\sqrt{\alpha} - \tfrac{2\beta}{\alpha}\right) \\
&\geq f\left(-\sqrt{\alpha}\right) + \frac{df}{dx}\left(-\sqrt{\alpha}\right) \times \left(-2\beta/\alpha\right) && \text{(A.17)} \\
&= -\beta + (-2\alpha) \times (-2\beta/\alpha) && \text{(A.18)} \\
&= 3\beta \quad > \quad 0. && \text{(A.19)}
\end{aligned}
$$

Under this condition, there is at least one root in the interval, $[-\sqrt{\alpha} - 2\beta/\alpha, -\sqrt{\alpha}]$. These three intervals do not overlap, as long as $\frac{4\beta}{\alpha} < \sqrt{\alpha}$, or $\beta < \frac{1}{4}\alpha^{3/2}$.

In the case that $\beta \leq 0$, a symmetric argument applies. Thus there are exactly three solutions to equation (A.8) in the specified intervals. ■

Lemma A.3. *Let $\boldsymbol{a}_l$ and $\boldsymbol{a}_{l'}$ be two nonzero vectors with inner product $\mu_{l,l'} \doteq \langle \boldsymbol{a}_l, \boldsymbol{a}_{l'} \rangle$. Then for any unit vector $\boldsymbol{v} \in \operatorname{span}(\boldsymbol{a}_l, \boldsymbol{a}_{l'})$,*

$$
\left|\left\langle \frac{\boldsymbol{a}_l}{\|\boldsymbol{a}_l\|_2}, \boldsymbol{v} \right\rangle\right|^2 + \left|\left\langle \frac{\boldsymbol{a}_{l'}}{\|\boldsymbol{a}_{l'}\|_2}, \boldsymbol{v} \right\rangle\right|^2 \geq 1 - \frac{|\mu_{l,l'}|}{\|\boldsymbol{a}_l\|_2 \|\boldsymbol{a}_{l'}\|_2}. \tag{A.20}
$$

Proof Let $\boldsymbol{u}$ and $\boldsymbol{u}^\perp$ be two orthogonal unit vectors, such that

$$
\begin{aligned}
\boldsymbol{a}_l &= \|\boldsymbol{a}_l\|_2 \boldsymbol{u}, && \text{(A.21)} \\
\boldsymbol{a}_{l'} &= \frac{\mu_{l,l'}}{\|\boldsymbol{a}_l\|_2}\boldsymbol{u} + \sqrt{\|\boldsymbol{a}_{l'}\|_2^2 - \frac{\mu_{l,l'}^2}{\|\boldsymbol{a}_l\|_2^2}}\boldsymbol{u}^\perp. && \text{(A.22)}
\end{aligned}
$$

Suppose $\boldsymbol{v} = a\boldsymbol{u} + b\boldsymbol{u}^\perp$ with $a^2 + b^2 = 1$. Let $\mu_{\text{rel}} = \frac{\mu_{l,l'}}{\|\boldsymbol{a}_l\|_2 \|\boldsymbol{a}_{l'}\|_2}$, then we can expand the quantity of interests as

$$
\begin{aligned}
& \left|\left\langle \frac{\boldsymbol{a}_l}{\|\boldsymbol{a}_l\|_2}, \boldsymbol{v} \right\rangle\right|^2 + \left|\left\langle \frac{\boldsymbol{a}_{l'}}{\|\boldsymbol{a}_{l'}\|_2}, \boldsymbol{v} \right\rangle\right|^2 \\
&= \left|\left\langle \boldsymbol{u}, a\boldsymbol{u} + b\boldsymbol{u}^\perp \right\rangle\right|^2 \\
&\quad + \left|\left\langle \mu_{\text{rel}}\boldsymbol{u} + \sqrt{1 - \mu_{\text{rel}}^2}\boldsymbol{u}^\perp, a\boldsymbol{u} + b\boldsymbol{u}^\perp \right\rangle\right|^2 && \text{(A.23)} \\
&= a^2 + \left(a\mu_{\text{rel}} + b\sqrt{1 - \mu_{\text{rel}}^2}\right)^2 && \text{(A.24)} \\
&= a^2 + b^2 + \left(a^2 - b^2\right)\mu_{\text{rel}}^2 + 2ab\mu_{\text{rel}}\sqrt{1 - \mu_{\text{rel}}^2} && \text{(A.25)}
\end{aligned}
$$

$$= 1 + \left[a^2 - b^2, 2ab\right] \left[\mu_{\text{rel}}^2, \mu_{\text{rel}}\sqrt{1-\mu_{\text{rel}}^2}\right]^T \tag{A.26}$$

Since $\left[a^2 - b^2, 2ab\right]$ is a unit vector, then above equation is lower bounded by

$$\begin{aligned} & 1 - \left\| \left[\mu_{rel}^2, \mu_{rel}\sqrt{1-\mu_{rel}^2}\right] \right\|_2 \\ & = 1 - |\mu_{rel}| \qquad (A.27) \\ & = 1 - \frac{|\mu_{l,l'}|}{\|\boldsymbol{a}_l\|_2 \|\boldsymbol{a}_{l'}\|_2} \qquad (A.28) \end{aligned}$$

as claimed. ■

Lemma A.4 (Nonzeros in a Bernoulli Vector). *Let $\boldsymbol{v} \sim_{\text{i.i.d.}} \text{Ber}(\theta) \in \mathbb{R}^n$, then*

$$\mathbb{P}\left[\|\boldsymbol{v}\|_0 \geq (1+t)\,\theta n\right] \leq 2\exp\left(-\frac{3t^2}{2t+6}\theta n\right). \tag{A.29}$$

Proof As $\|\boldsymbol{v}\|_0 = v_0 + \cdots + v_{n-1}$, and

$$|v_i - \theta| \leq 1, \quad \mathbb{E}\left[(v_i-\theta)^2\right] = \theta(1-\theta) \leq \theta \tag{A.30}$$

with Bernstein's inequality, we obtain that

$$\begin{aligned} & \mathbb{P}\left[\|\boldsymbol{v}\|_0 \geq (1+t)\,\theta n\right] \\ & \leq 2\exp\left(-\frac{t^2\theta^2 n^2}{2(\theta-\theta^2)n + \frac{2}{3}t\theta n}\right) \qquad (A.31) \\ & \leq 2\exp\left(-\frac{3t^2}{2t+6}\theta n\right), \qquad (A.32) \end{aligned}$$

as claimed. ■

Lemma A.5 (Entry-wise Truncation of a Bernoulli Gaussian Vector). *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^m$, then*

$$\mathbb{P}\left[\|\boldsymbol{x}_0\|_\infty > t\right] \leq 2\theta m e^{-t^2/2}. \tag{A.33}$$

Proof A Bernoulli-Gaussian variable $x = \omega \cdot g$ satisfies

$$\mathbb{P}\left[|x| \geq t\right] = \theta \cdot \mathbb{P}\left[|g| \geq t\right] \leq 2\theta e^{-t^2/2}, \tag{A.34}$$

Taking a union bound over the m entries of $\boldsymbol{x}_0$, we obtain

$$\begin{aligned} \mathbb{P}\left[\|\boldsymbol{x}_0\|_\infty > t\right] & \leq m\mathbb{P}\left[|x| > t\right] \qquad (A.35) \\ & \leq 2\theta m e^{-t^2/2}, \qquad (A.36) \end{aligned}$$

as claimed. ■

Lemma A.6 (Operator Norm of a Bernoulli Gaussian Circulant Matrix). *Let $\boldsymbol{C}_{\boldsymbol{x}_0} \in \mathbb{R}^{m\times m}$ be the circulant matrix generated from $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^m$, then*

$$\mathbb{P}\left[\|\boldsymbol{C}_{\boldsymbol{x}_0}\|_2 \geq t\right] \leq 2m\exp\left(-\frac{t^2}{2\theta m + 2t}\right). \tag{A.37}$$

Proof The operator norm of a circulant matrix is

$$\|\boldsymbol{C}_{\boldsymbol{x}_0}\|_2 = \max_l |\langle \boldsymbol{x}_0, \boldsymbol{w}_l \rangle|, \tag{A.38}$$

where $\boldsymbol{w}_l$ is the l-th (discrete) Fourier basis vector

$$\boldsymbol{w}_l = \left[1,\ e^{l\frac{2\pi j}{m}}, \cdots,\ e^{l(m-1)\frac{2\pi j}{m}}\right]^T, \quad l = 0, \cdots, m-1, \tag{A.39}$$

and j is the imaginary unit. With moment control Bernstein inequality, we obtain

$$\begin{aligned}\mathbb{P}\left[|\langle \boldsymbol{x}_0, \boldsymbol{w}_l \rangle| \geq t\right] &\leq 2\exp\left(-\frac{t^2}{2\theta \|\boldsymbol{w}_l\|_2^2 + 2\|\boldsymbol{w}_l\|_\infty t}\right) \\ &\leq 2\exp\left(-\frac{t^2}{2\theta m + 2t}\right)\end{aligned} \tag{A.40}$$

together with the union bound,

$$\mathbb{P}\left[\|\boldsymbol{C}_{\boldsymbol{x}_0}\|_2 \geq t\right] \leq m\mathbb{P}\left[|\langle \boldsymbol{x}_0, \boldsymbol{w}_l \rangle| \geq t\right] \tag{A.41}$$

$$\leq 2m\exp\left(-\frac{t^2}{2\theta m + 2t}\right), \tag{A.42}$$

as claimed. ■

Lemma A.7 (Norms of $\boldsymbol{\eta}$ and $\bar{\boldsymbol{\eta}}$). *Suppose* $\delta = \left\|\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T - \boldsymbol{I}\right\|_2 \leq 1/\left(2\kappa^2\right)$, *then vectors* $\boldsymbol{\eta} = \boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}$ *and* $\bar{\boldsymbol{\eta}} = \boldsymbol{Y}^T\left(\theta m \boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{q}$ *satisfy*

$$\|\boldsymbol{\eta}\|_\infty \leq \left(1 + \frac{4\kappa^3\delta}{\sigma_{\min}}\right)\left(\frac{2k}{\theta m}\right)^{1/2}\|\boldsymbol{x}_0\|_\infty, \tag{A.43}$$

$$\|\bar{\boldsymbol{\eta}}\|_\infty \leq \left(\frac{2k}{\theta m}\right)^{1/2}\|\boldsymbol{x}_0\|_\infty, \tag{A.44}$$

$$\|\boldsymbol{\eta}\|_6^6 \leq \left(1 + \frac{4\kappa^3\delta}{\sigma_{\min}}\right)^4 \frac{4k^2}{\theta^2 m^2}\|\boldsymbol{x}_0\|_\infty^4, \tag{A.45}$$

$$\|\bar{\boldsymbol{\eta}}\|_2 \leq 1 + \delta/2, \tag{A.46}$$

$$\|\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\|_\infty \leq \frac{4\kappa^3\delta}{\sigma_{\min}}\left(\frac{2k}{\theta m}\right)^{1/2}\|\boldsymbol{x}_0\|_\infty, \tag{A.47}$$

$$\|\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\|_2 \leq (1+\delta/2)\frac{4\kappa^3\delta}{\sigma_{\min}}. \tag{A.48}$$

Proof Since $\delta = \left\|\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T - \boldsymbol{I}\right\|_2$, then

$$\|\boldsymbol{X}_0\|_2 \leq (\theta m)^{1/2}\sqrt{1+\delta} \tag{A.49}$$

$$\leq (\theta m)^{1/2}(1+\delta/2). \tag{A.50}$$

As $\boldsymbol{\eta} = \boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q} = \boldsymbol{X}_0^T\boldsymbol{A}_0^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}$, together with Lemma D.3:

$$\begin{aligned}&\left\|\boldsymbol{A}_0^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}\right\|_\infty \\ &\leq \left\|\boldsymbol{A}_0^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}\right\|_2\end{aligned} \tag{A.51}$$

$$
\begin{aligned}
&\leq \left\| \boldsymbol{A}_0^T \left(\left(\boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} - \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \right) \boldsymbol{q} \right\|_2 \\
&\quad + \left\| \boldsymbol{A}_0^T \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{q} \right\|_2 && \text{(A.52)} \\
&\leq (\theta m)^{-1/2} \frac{4\kappa^3 \delta}{\sigma_{\min}} \left\| \boldsymbol{q} \right\|_2 + (\theta m)^{-1/2} \left\| \boldsymbol{A}^T \boldsymbol{q} \right\|_2 && \text{(A.53)} \\
&\leq (\theta m)^{-1/2} \left(1 + \frac{4\kappa^3 \delta}{\sigma_{\min}} \right) && \text{(A.54)}
\end{aligned}
$$

Norms of $\boldsymbol{\eta}$. Since $\left\| \boldsymbol{X}_0 \boldsymbol{e}_l \right\|_2 \leq \sqrt{2k-1} \left\| \boldsymbol{X}_0 \boldsymbol{e}_l \right\|_\infty$, we have

$$
\begin{aligned}
\left\| \boldsymbol{\eta} \right\|_\infty &= \max_{l \in [1, \cdots, m]} \left\langle \boldsymbol{X}_0 \boldsymbol{e}_l, \boldsymbol{A}_0^T \left(\boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} \boldsymbol{q} \right\rangle && \text{(A.55)} \\
&\leq \max_l \left\| \boldsymbol{X}_0 \boldsymbol{e}_l \right\|_2 \left\| \boldsymbol{A}_0^T \left(\boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} \boldsymbol{q} \right\|_2 && \text{(A.56)} \\
&\leq \sqrt{2k} \left\| \boldsymbol{x}_0 \right\|_\infty \cdot (\theta m)^{-1/2} \left(1 + \frac{4\kappa^3 \delta}{\sigma_{\min}} \right). && \text{(A.57)}
\end{aligned}
$$

At the same time, plugging in $\left\| \boldsymbol{\eta} \right\|_2 = 1$, we have

$$
\left\| \boldsymbol{\eta} \right\|_6^6 \leq \left\| \boldsymbol{\eta} \right\|_2^2 \left\| \boldsymbol{\eta} \right\|_\infty^4 \leq \left(1 + \frac{4\kappa^3 \delta}{\sigma_{\min}} \right)^4 \frac{4k^2}{\theta^2 m^2} \left\| \boldsymbol{x}_0 \right\|_\infty^4. \tag{A.58}
$$

Norms of $\bar{\boldsymbol{\eta}}$. Here, $\bar{\boldsymbol{\eta}} = \boldsymbol{Y}^T \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{q} = \boldsymbol{X}_0^T \boldsymbol{A}_0^T \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{q}$ with

$$
\begin{aligned}
&\left\| \boldsymbol{A}_0^T \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{q} \right\|_\infty \\
&\leq \left\| \boldsymbol{A}_0^T \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{q} \right\|_2 && \text{(A.59)} \\
&= (\theta m)^{-1/2}, && \text{(A.60)}
\end{aligned}
$$

therefore

$$
\begin{aligned}
\left\| \bar{\boldsymbol{\eta}} \right\|_\infty &\leq \max_l \left\| \boldsymbol{X}_0 \boldsymbol{e}_l \right\|_2 \left\| \boldsymbol{A}_0 \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{q} \right\|_2 \\
&\leq \left(\frac{2k}{\theta m} \right)^{1/2} \left\| \boldsymbol{x}_0 \right\|_\infty, && \text{(A.61)} \\
\left\| \bar{\boldsymbol{\eta}} \right\|_2 &\leq \left\| \boldsymbol{X}_0^T \right\|_2 \left\| \boldsymbol{A}_0 \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{q} \right\|_2 \\
&\leq 1 + \delta/2. && \text{(A.62)}
\end{aligned}
$$

Norms of $\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}$. With similar reasoning, we can obtain

$$
\begin{aligned}
&\left\| \boldsymbol{\eta} - \bar{\boldsymbol{\eta}} \right\|_\infty \\
&= \left\| \boldsymbol{Y}^T \left(\boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} \boldsymbol{q} - \boldsymbol{Y}^T \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{q} \right\|_\infty \\
&\leq \max_{l \in [1, \cdots, m]} \left\| \boldsymbol{X}_0 \boldsymbol{e}_l \right\|_2 (\theta m)^{-1/2} \times \\
&\quad \left\| \boldsymbol{A}_0^T \left(\frac{1}{\theta m} \boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} - \boldsymbol{A}_0^T \left(\boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \right\|_2 && \text{(A.63)} \\
&\leq \frac{4\kappa^3 \delta}{\sigma_{\min}} \left(\frac{2k}{\theta m} \right)^{1/2} \left\| \boldsymbol{x}_0 \right\|_\infty, && \text{(A.64)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\|\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}\|_2 \\
&\leq \|\boldsymbol{X}_0\|_2 (\theta m)^{-1/2} \|\boldsymbol{q}\|_2 \times \\
&\quad \left\| \boldsymbol{A}_0^T \left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} - \boldsymbol{A}_0^T \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2} \right\|_2 && \text{(A.65)} \\
&\leq (\theta m)^{-1/2} \frac{4\kappa^3\delta}{\sigma_{\min}} \|\boldsymbol{X}_0\|_2 && \text{(A.66)} \\
&\leq (1+\delta/2) \frac{4\kappa^3\delta}{\sigma_{\min}}, && \text{(A.67)}
\end{aligned}
$$

completing the proof. ■

B Proof of the Main Theorem and Corollary

B.1 Proof of the Main Theorem

Lemma B.1. *If following inequalities hold*

$$
\begin{aligned}
&\left\| \mathrm{grad}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2} \mathrm{grad}[\varphi](\boldsymbol{q}) \right\|_2 \\
&\leq \frac{3c_\star}{2\kappa^2} \frac{1-\theta}{\theta m^2} \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6, && \text{(B.1)} \\
&\left\| \mathrm{Hess}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2} \mathrm{Hess}[\varphi](\boldsymbol{q}) \right\|_2 \\
&\leq 3\left(1-6c_\star-36c_\star^2-24c_\star^3\right) \frac{1-\theta}{\theta m^2} \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4. && \text{(B.2)}
\end{aligned}
$$

for all $\boldsymbol{q} \in \mathcal{R}_{2C_\star}$ with $C_\star \geq 10$ and $c_\star = 1/C_\star$, then any local minimum $\bar{\boldsymbol{q}}$ of $\psi(\boldsymbol{q})$ in $\mathcal{R}_{2C_\star}$ satisfies $|\langle \bar{\boldsymbol{q}}, \mathcal{P}_{\mathbb{S}}[\boldsymbol{a}_l]\rangle| \geq 1-2c_\star\kappa^{-2}$ for some index l.

Proof Let

$$
\boldsymbol{\delta}_{\mathrm{grad}} = \mathrm{grad}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2} \mathrm{grad}[\varphi](\boldsymbol{q}), \tag{B.3}
$$

and let

$$
\bar{\boldsymbol{\delta}}_{\mathrm{grad}} = \frac{\theta m^2}{3(1-\theta)} \boldsymbol{\delta}_{\mathrm{grad}}. \tag{B.4}
$$

Then at any stationary point of $\psi(\boldsymbol{q})$, we have

$$
\begin{aligned}
\boldsymbol{0} &= \boldsymbol{A}^T \mathrm{grad}[\psi](\boldsymbol{q}) && \text{(B.5)} \\
&= \frac{3(1-\theta)}{\theta m^2} \boldsymbol{A}^T \mathrm{grad}[\varphi](\boldsymbol{q}) + \boldsymbol{A}^T\boldsymbol{\delta}_{\mathrm{grad}}. && \text{(B.6)}
\end{aligned}
$$

Hence for any index i, following equality always holds

$$
\begin{aligned}
0 &= \|\boldsymbol{a}_i\|_2^2 \zeta_i^3 + \sum_{j\neq i} \langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle \zeta_j^3 - \zeta_i \|\boldsymbol{\zeta}\|_4^4 + \left\langle \boldsymbol{a}_i, \bar{\boldsymbol{\delta}}_{\mathrm{grad}} \right\rangle \\
&= \zeta_i^3 - \zeta_i \underbrace{\frac{\|\boldsymbol{\zeta}\|_4^4}{\|\boldsymbol{a}_i\|_2^2}}_{\alpha_i} + \underbrace{\frac{\sum_{j\neq i} \langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle \zeta_j^3 + \left\langle \boldsymbol{a}_i, \bar{\boldsymbol{\delta}}_{\mathrm{grad}} \right\rangle}{\|\boldsymbol{a}_i\|_2^2}}_{\beta_i'} && \text{(B.7)}
\end{aligned}
$$

with $\boldsymbol{\zeta} = \boldsymbol{A}^T \boldsymbol{q}$. Under the assumption that

$$\left\| \mathrm{grad}[\psi]\left(\boldsymbol{q}\right) - \frac{3\left(1-\theta\right)}{\theta m^2} \mathrm{grad}[\varphi]\left(\boldsymbol{q}\right) \right\|_2 \leq \frac{3c_\star}{2\kappa^2} \frac{1-\theta}{\theta m^2} \left\|\boldsymbol{\zeta}\right\|_4^6, \tag{B.8}$$

the perturbed part can be bounded via

$$\left| \left\langle \boldsymbol{a}_i, \bar{\boldsymbol{\delta}}_{\mathrm{grad}} \right\rangle \right| \leq \left\|\boldsymbol{a}_i\right\|_2 \left\|\bar{\boldsymbol{\delta}}_{\mathrm{grad}}\right\|_2 \leq \frac{c_\star}{2\kappa^2} \left\|\boldsymbol{a}_i\right\|_2 \left\|\boldsymbol{\zeta}\right\|_4^6, \tag{B.9}$$

and also

$$\frac{\beta_i'}{\alpha_i^{3/2}} \leq \frac{\mu \left\|\boldsymbol{\zeta}\right\|_3^3 + \frac{1}{2} c_\star \kappa^{-2} \left\|\boldsymbol{\zeta}\right\|_4^6}{\left\|\boldsymbol{\zeta}\right\|_4^6} \leq c_\star \kappa^{-2} \leq \frac{1}{4}. \tag{B.10}$$

Then by Lemma A.2, at every stationary point $\bar{\boldsymbol{q}}$, the i-th entry of $\boldsymbol{\zeta}$ resides in the set $\bigcup_{x\in\{0,\pm\sqrt{\alpha_i}\}} [x - \frac{2\beta_i'}{\alpha_i}, x + \frac{2\beta_i'}{\alpha_i}]$ – i.e., $\boldsymbol{\zeta}$ is nearly a trinary vector.

Moreover, we can characterize the curvature of critical points in terms of the number of large entries of $\boldsymbol{\zeta}$. Indeed, whenever $\boldsymbol{\zeta}$ has at least two entries in

$$\bigcup_{x\in\{\pm\sqrt{\alpha_i}\}} \left[x - \frac{2\beta_i'}{\alpha_i}, x + \frac{2\beta_i'}{\alpha_i} \right],$$

using (3.52), there exists a direction of strict negative curvature, provided

$$\begin{aligned} \mathrm{Hess}[\psi]\left(\boldsymbol{q}\right) \prec\ & \frac{3\left(1-\theta\right)}{\theta m^2} \mathrm{Hess}[\varphi]\left(\boldsymbol{q}\right) \\ & +\ 3\left(2-11c_\star\right) \frac{1-\theta}{\theta m^2} \left\|\boldsymbol{\zeta}\right\|_4^4 \boldsymbol{I}. \end{aligned} \tag{B.11}$$

Similarly, whenever $\boldsymbol{\zeta}$ has only one entry in

$$\bigcup_{x\in\{\pm\sqrt{\alpha_i}\}} \left[x - \frac{2\beta_i'}{\alpha_i}, x + \frac{2\beta_i'}{\alpha_i} \right],$$

using (3.35), we have that $\mathrm{Hess}[\psi](\boldsymbol{q}) \succ \boldsymbol{0}$, provided

$$\begin{aligned} \mathrm{Hess}[\psi]\left(\boldsymbol{q}\right) \succ\ & \frac{3\left(1-\theta\right)}{\theta m^2} \mathrm{Hess}[\varphi]\left(\boldsymbol{q}\right) \\ & -3\left(1-6c_\star-36c_\star^2-24c_\star^3\right) \frac{1-\theta}{\theta m^2} \left\|\boldsymbol{\zeta}\right\|_4^4 \boldsymbol{I}. \end{aligned} \tag{B.12}$$

When $C_\star \geq 10$ and $c_\star \leq 0.1$, we have $2 - 11c_\star > 1 - 6c_\star - 36c_\star^2 - 24c_\star^3 \geq 0.016$, and so above characterization obtains. ∎

Theorem B.2 (Main Result). *Assume the observation $\boldsymbol{y} \in \mathbb{R}^m$ is the cyclic convolution of $\boldsymbol{a}_0 \in \mathbb{R}^k$ and $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}\left(\theta\right) \in \mathbb{R}^m$, where the convolution matrix $\boldsymbol{A}_0 \in \mathbb{R}^{k\times(2k-1)}$ has minimum singular value $\sigma_{\min} > 0$ and condition number $\kappa \geq 1$, and $\boldsymbol{A}$ has column incoherence μ. If*

$$m \ \geq\ C \frac{\min\left\{\left(2C_\star\mu\right)^{-1}, \kappa^2 k^2\right\}}{\left(1-\theta\right)^2 \sigma_{\min}^2} \kappa^8 k^4 \log^3\left(\kappa k\right) \tag{B.13}$$

and $\theta \geq \log k/k$, then with probability no smaller than $1 - \exp\left(-k\right) - \theta^2\left(1-\theta\right)^2 k^{-4} - 2\exp\left(-\theta k\right) - 48k^{-7} - 48m^{-5} - 24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right)$, any local minimum $\bar{\boldsymbol{q}}$ of ψ in $\hat{\mathcal{R}}_{2C_\star}$ satisfies $\left|\langle \bar{\boldsymbol{q}}, \mathcal{P}_{\mathbb{S}}\left[\boldsymbol{a}_\tau\right]\rangle\right| \geq 1 - c_\star\kappa^{-2}$ for some integer τ.

Proof From the concentration analysis for the Riemannian gradient (Lemma 4.2) and Hessian (Lemma 4.3), if

$$m \;\geq\; C\frac{\min\left\{(2C_\star\mu)^{-1}, \kappa^2k^2\right\}}{(1-\theta)^2\,\sigma_{\min}^2}\kappa^8k^4\log^3(\kappa k), \tag{B.14}$$

then with probability no smaller than $1-\exp(-k)-\theta^2(1-\theta)^2k^{-4}-2\exp(-\theta k)-24k\exp\left(-\frac{1}{144}\min\left\{k,3\sqrt{\theta m}\right\}\right)-48k^{-7}-48m^{-5}$,

$$\begin{aligned}
&\left\|\mathrm{grad}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\mathrm{grad}[\varphi](\boldsymbol{q})\right\|_2 \\
&\leq \frac{3c_\star}{2\kappa^2}\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6, \qquad\qquad (B.15)\\
&\left\|\mathrm{Hess}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\mathrm{Hess}[\varphi](\boldsymbol{q})\right\|_2 \\
&\leq 3\left(1-6c_\star-36c_\star^2-24c_\star^3\right)\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4. \qquad\qquad (B.16)
\end{aligned}$$

hold for all $\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}$ with $C_\star\geq 10$ and $c_\star = 1/C_\star$. Therefore, by Lemma 4.1 any local minimum $\bar{\boldsymbol{q}}$ of $\psi(\boldsymbol{q})$ in $\mathcal{R}_{2C_\star}$ satisfies $|\langle\bar{\boldsymbol{q}}, \mathcal{P}_{\mathbb{S}}[\boldsymbol{a}_l]\rangle| \geq 1-2c_\star\kappa^{-2}$ for some index l. ■

B.2 Proof of the Main Corollary

Corollary B.3. *Suppose the ground truth kernel $\boldsymbol{a}_0$ has induces coherence $0\leq\mu\leq\frac{1}{8\times 48}\log^{-3/2}(k)$ and sparse coefficient $\boldsymbol{x}_0\sim_{\text{i.i.d.}}\mathrm{BG}(\theta)\in\mathbb{R}^m$. there exist positive constants $C\geq 2560^4$ and C' such that whenever the sparsity level*

$$\begin{aligned}
&64k^{-1}\log k \leq \theta \leq \min\left\{\tfrac{1}{48^2}\mu^{-2}k^{-1}\log^{-2}k,\right. \qquad (B.17)\\
&\left.\left(\tfrac{1}{4}-\tfrac{640}{C^{1/4}}\right)\left(3C_\star\mu\kappa^2\right)^{-2/3}k^{-1}\left(1+36\mu^2k\log k\right)^{-2}\right\},
\end{aligned}$$

and signal length

$$\begin{aligned}
m \;&\geq\; \max\left\{C\theta^2\sigma_{\min}^{-2}\kappa^6k^3\left(1+36\mu^2k\log k\right)^4\log(\kappa k),\right. \qquad (B.18)\\
&\left. C'(1-\theta)^{-2}\sigma_{\min}^{-2}\min\left\{\mu^{-1},\kappa^2k^2\right\}\kappa^8k^4\log^3(\kappa k)\right\},
\end{aligned}$$

then Algorithm 1 recovers $\bar{\boldsymbol{a}}$ such that

$$\left\|\bar{\boldsymbol{a}}\pm\mathcal{P}_{\mathbb{S}}\left[\boldsymbol{\iota}_k s_\tau[\widetilde{\boldsymbol{a}_0}]\right]\right\|_2 \leq 4\sqrt{c_\star}+ck^{-1} \tag{B.19}$$

for some integer shift $\tau\in[-(k-1),k-1]$ with probability no smaller than $1-k^{-1}-8k^{-2}-\exp(-k)-\theta^2(1-\theta)^2k^{-4}-2\exp(-\theta k)-24k\exp\left(-\frac{1}{144}\min\left\{k,3\sqrt{\theta m}\right\}\right)-48k^{-7}-48m^{-5}$.

Proof From the concentration results for the Riemannian gradient, at every point $\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}$, the objective value of $\psi(\boldsymbol{q})$ satisfies

$$\begin{aligned}
&\left|\psi(\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\varphi(\boldsymbol{q}) + \frac{3}{4m^2}\right| \\
&\leq \left|\frac{\left\|\boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}\right\|_4^4}{4m} - \frac{3(1-\theta)\left\|\boldsymbol{\zeta}\right\|_4^4}{4\theta m^2} - \frac{3}{4m^2}\right|
\end{aligned} \tag{B.20}$$

$$\leq \left| \left\langle \boldsymbol{q}, \frac{\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3}}{4m} - \frac{3\left(1-\theta\right)}{4\theta m^2}\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} - \frac{3}{4m^2}\boldsymbol{q} \right\rangle \right| \tag{B.21}$$

$$\leq \left\| \frac{\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3}}{4m} - \frac{3\left(1-\theta\right)}{4\theta m^2}\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} - \frac{3}{4m^2}\boldsymbol{q} \right\|_2 \tag{B.22}$$

$$\begin{aligned} &\leq \frac{1}{4m}\left\| \left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3} - \left(\theta m \boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3} \right\|_2 \\ &\quad + \frac{1}{4\theta^{1/2}m^{3/2}}\left\| \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{Y}\left(\boldsymbol{\eta}^{\circ 3} - \bar{\boldsymbol{\eta}}^{\circ 3}\right) \right\|_2 \\ &\quad + \left\| \frac{1}{4\theta^{1/2}m^{3/2}}\left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{Y}\bar{\boldsymbol{\eta}}^{\circ 3} \right. \\ &\qquad \left. - \frac{3\left(1-\theta\right)}{4\theta m^2}\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} - \frac{3}{4m^2}\boldsymbol{q} \right\|_2 \end{aligned} \tag{B.23}$$

$$\leq \frac{3c_\star}{8\kappa^2}\frac{1-\theta}{\theta m^2}\min_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6 \tag{B.24}$$

with probability no smaller than $1 - 2\exp\left(-\theta k\right) - 24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right) - 48k^{-7} - 48m^{-5}$. The last inequality is derived with similar arguments in Lemma 4.2, for simplicity, we do not present them here. Moreover, with Lemma C.1, we can obtain an initialization point $\boldsymbol{q}_{\text{init}}$ such that

$$\left\|\boldsymbol{A}^T\boldsymbol{q}_{\text{init}}\right\|_4^4 \geq \left(3C_\star\mu\kappa^2\right)^{2/3} \tag{B.25}$$

$$\geq \left(2C_\star\mu\kappa^2\right)^{2/3} + \mu/2. \tag{B.26}$$

Consider any descent method for ψ, which generates a sequence of iterates $\boldsymbol{q}^{(0)} = \boldsymbol{q}_{\text{init}}, \boldsymbol{q}^{(1)}, \ldots, \boldsymbol{q}^{(k)}, \ldots$ such that $\psi(\boldsymbol{q}^{(k)})$ is non-increasing with k. Then

$$\psi\left(\boldsymbol{q}^{(k)}\right) \leq \psi\left(\boldsymbol{q}_{\text{init}}\right) \tag{B.27}$$

$$\begin{aligned} &\leq \frac{3\left(1-\theta\right)}{\theta m^2}\varphi\left(\boldsymbol{q}_{\text{init}}\right) + \frac{3}{4m^2} \\ &\quad + \frac{3c_\star}{8\kappa^2}\frac{1-\theta}{\theta m^2}\min_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6. \end{aligned} \tag{B.28}$$

On the other hand, the finite sample objective function value ψ is close to that of $\frac{3(1-\theta)}{\theta m^2}\varphi\left(\boldsymbol{q}\right) - \frac{3}{4m^2}$,

$$\begin{aligned} &\frac{3\left(1-\theta\right)}{\theta m^2}\varphi\left(\boldsymbol{q}^{(k)}\right) \\ &\leq \psi\left(\boldsymbol{q}^{(k)}\right) + \frac{3}{4m^2} + \frac{3c_\star}{8\kappa^2}\frac{1-\theta}{\theta m^2}\min_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6 \end{aligned} \tag{B.29}$$

$$\leq \frac{3\left(1-\theta\right)}{\theta m^2}\varphi\left(\boldsymbol{q}_{\text{init}}\right) + \frac{3c_\star}{4\kappa^2}\frac{1-\theta}{\theta m^2}\min_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6, \tag{B.30}$$

Therefore, we obtain that

$$\varphi\left(\boldsymbol{q}^{(k)}\right) \leq \varphi\left(\boldsymbol{q}_{\text{init}}\right) + \frac{\mu}{2} \tag{B.31}$$

$$\leq \varphi\left(\boldsymbol{q}_{\text{init}}\right) + \frac{c_\star}{4\kappa^2}\min_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6, \tag{B.32}$$

which implies that $\boldsymbol{q}^{(k)} \in \hat{\mathcal{R}}_{2C_\star}$ always holds. At last, Theorem B.2 says that any local minimum $\bar{\boldsymbol{q}}$ is close to $\pm\boldsymbol{a}_i$ for some i, in the sense that

$$|\langle \bar{\boldsymbol{q}}, \mathcal{P}_{\mathbb{S}}\left[\boldsymbol{a}_i\right]\rangle| \geq 1 - c_\star \kappa^{-2}. \tag{B.33}$$

Write $\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{A}_0\left(\boldsymbol{I}+\boldsymbol{\Delta}\right)\boldsymbol{A}_0^T$ with $\|\boldsymbol{\Delta}\|_2 \leq \delta$, and let

$$\bar{\boldsymbol{q}} = \pm\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2} + \sqrt{2\left(1-\left|\left\langle \bar{\boldsymbol{q}}, \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2}\right\rangle\right|\right)}\boldsymbol{\delta}, \tag{B.34}$$

with $\|\boldsymbol{\delta}\|_2 = 1$. Since

$$\boldsymbol{a}_i = \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{\iota}_k^* s_{-(k-i)}[\widetilde{\boldsymbol{a}_0}], \tag{B.35}$$

we have

$$\begin{aligned}
&\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m}\right)^{1/2}\bar{\boldsymbol{q}} \\
&= \pm\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m}\right)^{1/2}\left[\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2}+\sqrt{2\left(1-\left|\left\langle \bar{\boldsymbol{q}}, \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2}\right\rangle\right|\right)}\boldsymbol{\delta}\right] &\text{(B.36)}\\
&= \pm\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m}\right)^{1/2}\left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\frac{\boldsymbol{\iota}_k^* s_{-(k-i)}[\widetilde{\boldsymbol{a}_0}]}{\|\boldsymbol{a}_i\|_2} \\
&\quad + \sqrt{2\left(1-\left|\left\langle \bar{\boldsymbol{q}}, \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2}\right\rangle\right|\right)}\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m}\right)^{1/2}\boldsymbol{\delta} &\text{(B.37)}
\end{aligned}$$

therefore the error can be bounded as

$$\begin{aligned}
&\left\|\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m}\right)^{1/2}\bar{\boldsymbol{q}} \pm \frac{\boldsymbol{\iota}_k^* s_{-(k-i)}[\widetilde{\boldsymbol{a}_0}]}{\|\boldsymbol{a}_i\|_2}\right\|_2 \\
&\leq \left\|\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m}\right)^{1/2}\left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2} - \boldsymbol{I}\right\|_2\left\|\frac{\boldsymbol{\iota}_k^* s_{-(k-i)}[\widetilde{\boldsymbol{a}_0}]}{\|\boldsymbol{a}_i\|_2}\right\|_2 \\
&\quad + \sqrt{2\left(1-\left|\left\langle \bar{\boldsymbol{q}}, \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2}\right\rangle\right|\right)}\left\|\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m}\right)^{1/2}\right\|_2. &\text{(B.38)}
\end{aligned}$$

Finally, using the fact that for any nonzero vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ that $\langle \boldsymbol{u}, \boldsymbol{v}\rangle \geq 0$,

$$\left\|\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} - \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\right\|_2 \leq \frac{\sqrt{2}}{\|\boldsymbol{v}\|_2}\|\boldsymbol{u}-\boldsymbol{v}\|_2 \tag{B.39}$$

always holds. Therefore,

$$\begin{aligned}
&\|\bar{\boldsymbol{a}} \pm \mathcal{P}_{\mathbb{S}}\left[\boldsymbol{\iota}_k s_i[\widetilde{\boldsymbol{a}_0}]\right]\|_2 \\
&= \left\|\mathcal{P}_{\mathbb{S}}\left[\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{1/2}\bar{\boldsymbol{q}}\right] \pm \mathcal{P}_{\mathbb{S}}\left[\boldsymbol{\iota}_k s_i[\widetilde{\boldsymbol{a}_0}]\right]\right\|_2 &\text{(B.40)}\\
&\leq \frac{\sqrt{2}\|\boldsymbol{a}_i\|_2}{\|\boldsymbol{\iota}_k^* s_{i-k}[\widetilde{\boldsymbol{a}_0}]\|_2}\left\|\left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m}\right)^{1/2}\bar{\boldsymbol{q}} \pm \frac{\boldsymbol{\iota}_k^* s_{i-k}[\widetilde{\boldsymbol{a}_0}]}{\|\boldsymbol{a}_i\|_2}\right\|_2 &\text{(B.41)}\\
&\leq \kappa\sqrt{2\left(1+\delta\right)\left(1-\left|\left\langle \bar{\boldsymbol{q}}, \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2}\right\rangle\right|\right)} &\text{(B.42)}
\end{aligned}$$

$$+\sqrt{2}\kappa \left\| \left(\frac{\boldsymbol{Y}\boldsymbol{Y}^T}{\theta m} \right)^{1/2} \left(\boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} - \boldsymbol{I} \right\|_2$$

$$\leq 2\kappa \sqrt{2\left(1 - \left| \left\langle \bar{\boldsymbol{q}}, \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|_2} \right\rangle \right|\right)} + \sqrt{2}\kappa^3 \delta/\sigma_{\min} \tag{B.43}$$

(Lemma D.2)

$$\leq 4\sqrt{c_\star} + 10\sqrt{2}\kappa^3 \sigma_{\min}^{-1} \sqrt{k \log m/m} \tag{B.44}$$

$$\leq 4\sqrt{c_\star} + ck^{-1}, \tag{B.45}$$

completing the proof. ■

C Initialization

Lemma C.1. *Suppose* $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^m$. *There exists a positive constant* $C > 2560^4$ *such that whenever*

$$m \geq C\theta^2 \sigma_{\min}^{-2} \kappa^6 k^3 \left(1 + 36\mu^2 k \log k\right)^4 \log\left(\kappa k/\sigma_{\min}\right) \tag{C.1}$$

and the sparsity rate

$$64k^{-1} \log k \leq \theta \leq \min\left\{ \tfrac{1}{48^2} \mu^{-2} k^{-1} \log^{-2} k, \right. \tag{C.2}$$
$$\left. \left(\tfrac{1}{4} - \tfrac{640}{C^{1/4}}\right) \left(3C_\star \mu\kappa^2\right)^{-2/3} k^{-1} \left(1 + 36\mu^2 k \log k\right)^{-2} \right\},$$

Then the initialization $\boldsymbol{q}_{\text{init}} = \mathcal{P}_{\mathbb{S}}\left[\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} \boldsymbol{y}_i \right]$ *satisfies*

$$\left\| \boldsymbol{A}^T \boldsymbol{q}_{\text{init}} \right\|_4^6 \geq 3C_\star \mu\kappa^2, \tag{C.3}$$

namely $\boldsymbol{q}_{\text{init}} \in \hat{\mathcal{R}}_{3C_\star}$, *with probability no smaller than* $1 - k^{-1} - 8k^{-2} - 2\exp\left(-\theta k\right) - 48k^{-7} - 48m^{-5} - 24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right)$.

Proof Since

$$m \geq C\frac{\theta^2}{\sigma_{\min}^2} \kappa^6 k^3 \left(1 + 36\mu^2 k \log k\right)^4 \log\left(\kappa k/\sigma_{\min}\right) \tag{C.4}$$

with $C \geq 2560^4$, then from Lemma D.1, then with probability no smaller than $1 - 2\exp\left(-\theta k\right) - 24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right) - 48k^{-7} - 48m^{-5}$, we have

$$\delta \doteq \left\| \frac{1}{\theta m} \boldsymbol{X}_0 \boldsymbol{X}_0^T - \boldsymbol{I} \right\|_2 \tag{C.5}$$

$$\leq 10\sqrt{k \log m/m} \tag{C.6}$$

$$\leq \frac{10\sigma_{\min}}{\theta \sigma_{\min}^{-1} \kappa^3 k \left(1 + 36\mu^2 k \log k\right)^2} \times$$
$$\sqrt{\frac{\log\left(\frac{C\kappa^6 k^3 (1+36\mu^2 k \log k)^4}{\sigma_{\min}^2} \log\left(\frac{\kappa k}{\sigma_{\min}}\right)\right)}{C \log\left(\kappa k/\sigma_{\min}\right)}} \tag{C.7}$$

$$\leq \frac{20\sigma_{\min}}{C^{1/4}\theta\kappa^3 k \left(1 + 36\mu^2 k \log k\right)^2} \tag{C.8}$$

obtains, and the last inequality holds when $C \geq 1000$ that

$$\log\left(37^4 C\right) \leq \log 2\sqrt{C}. \tag{C.9}$$

Therefore

$$\begin{aligned}
&C\sigma_{\min}^{-2}\kappa^6 k^3 \left(1+36\mu^2 k\log k\right)^4 \log\left(\kappa k/\sigma_{\min}\right) \\
&\leq 37^4 C\left(\kappa k/\sigma_{\min}\right)^7 \log^5\left(\kappa k/\sigma_{\min}\right) && \text{(C.10)} \\
&\leq 37^4 C\left(\kappa k/\sigma_{\min}\right)^{12} && \text{(C.11)}
\end{aligned}$$

or

$$\begin{aligned}
&\sqrt{\frac{\log\left(\frac{C\kappa^6 k^3(1+36\mu^2 k\log k)^4}{\sigma_{\min}^2}\log\left(\frac{\kappa k}{\sigma_{\min}}\right)\right)}{C\log\left(\kappa k/\sigma_{\min}\right)}} \\
&\leq \sqrt{\frac{\log\left(37^4 C\right)+12\log\left(\kappa k/\sigma_{\min}\right)}{C\log\left(\kappa k/\sigma_{\min}\right)}} && \text{(C.12)} \\
&\leq \sqrt{\frac{\log 2}{\sqrt{C}\log\left(\kappa k/\sigma_{\min}\right)}+\frac{12}{C}} && \text{(C.13)} \\
&\leq \frac{2}{C^{1/4}} \qquad (k\geq 2, C\geq 16) && \text{(C.14)}
\end{aligned}$$

Moreover, $\kappa^2\delta \leq 1/2$ always holds provided

$$C \geq \left(\frac{40}{\theta k\left(1+36\mu^2 k\log k\right)^2}\right)^4. \tag{C.15}$$

Notice that because θ is lower bounded by $c\log k/k$, the right hand side is indeed bounded by an absolute constant.

Set $\boldsymbol{\zeta}_{\text{init}} = \boldsymbol{A}^T\boldsymbol{q}_{\text{init}}$ and $\hat{\boldsymbol{\zeta}}_{\text{init}} = \mathcal{P}_{\mathbb{S}}\left[\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i\right]$. Then using for any nonzero vectors $\boldsymbol{u}$ and $\boldsymbol{v}$,

$$\left\|\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}-\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\right\|_2 \leq \frac{2}{\|\boldsymbol{v}\|_2}\|\boldsymbol{u}-\boldsymbol{v}\|_2, \tag{C.16}$$

we have that

$$\begin{aligned}
&\left\|\boldsymbol{\zeta}_{\text{init}}-\hat{\boldsymbol{\zeta}}_{\text{init}}\right\|_2 \\
&= \left\|\boldsymbol{A}^T\mathcal{P}_{\mathbb{S}}\left[\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{x}_i\right]-\mathcal{P}_{\mathbb{S}}\left[\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i\right]\right\|_2 && \text{(C.17)} \\
&= \left\|\frac{\boldsymbol{A}^T\left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{x}_i}{\left\|\left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{x}_i\right\|_2}-\frac{\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i}{\|\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i\|_2}\right\|_2 && \text{(C.18)} \\
&\leq \frac{2}{\|\boldsymbol{A}\boldsymbol{x}_i\|_2}\left\|\left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{x}_i-\boldsymbol{A}\boldsymbol{x}_i\right\|_2 && \text{(C.19)} \\
&\leq 2\|\boldsymbol{A}_0\|_2\left\|\left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}-\left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\right\|_2 && \text{(C.20)} \\
&\leq \frac{8\kappa^3\delta}{\sigma_{\min}}, && \text{(C.21)}
\end{aligned}$$

where we have used Lemma D.3 in the final bound.

Since $\|\cdot\|_4^4$ is convex, $\|\boldsymbol{\zeta}_{\text{init}}\|_4^4$ can be lower bounded via

$$\|\boldsymbol{\zeta}_{\text{init}}\|_4^4 \geq \left\|\hat{\boldsymbol{\zeta}}_{\text{init}}\right\|_4^4 + 4\left\langle \hat{\boldsymbol{\zeta}}_{\text{init}}^{\circ 3}, \boldsymbol{\zeta}_{\text{init}} - \hat{\boldsymbol{\zeta}}_{\text{init}}\right\rangle \tag{C.22}$$

$$\geq \left\|\hat{\boldsymbol{\zeta}}_{\text{init}}\right\|_4^4 - 4\left\|\boldsymbol{\zeta}_{\text{init}} - \hat{\boldsymbol{\zeta}}_{\text{init}}\right\|_2 \tag{C.23}$$

$$\geq \left\|\hat{\boldsymbol{\zeta}}_{\text{init}}\right\|_4^4 - \frac{32\kappa^3\delta}{\sigma_{\min}}. \tag{C.24}$$

Let $I = \operatorname{supp}(\boldsymbol{x}_i)$, then the vector $\hat{\boldsymbol{\zeta}}_{\text{init}} = \mathcal{P}_{\mathbb{S}}\left[\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i\right]$ is composed of $|I|$ large components and small components on the off-support I^c of $\boldsymbol{x}_i$.

Dense Component of $\hat{\boldsymbol{\zeta}}_{\text{init}}$. Note that $\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I^c,I}\boldsymbol{x}_i\right\|_2 \leq \left\|\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2$ with $\left\|\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\right\|_\infty \leq \mu$. We have

$$\mathbb{E}\left[\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right] = \boldsymbol{0} \tag{C.25}$$

$$\mathbb{E}\left[\left|\boldsymbol{e}_j^T \operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right|^2\right] = \theta\left\|\boldsymbol{e}_j^T \operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\right\|_2^2 \leq \mu^2\theta k \tag{C.26}$$

With Bernstein's Inequality, the summation of moment-bounded independent random variables can be controlled via

$$\mathbb{P}\left[\left|\boldsymbol{e}_j^T \operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right| \geq \mu t\right] \leq 2\exp\left(-\frac{t^2}{2\theta k + 2t}\right) \tag{C.27}$$

and via union bound

$$\mathbb{P}\left[\left\|\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2 \geq 2k\left(\mu t\right)^2\right] \leq 4k\exp\left(-\frac{t^2}{2\theta k + 2t}\right) \tag{C.28}$$

Therefore, setting $t^2 = 9\theta k\log k$, we obtain

$$\left\|\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2 \leq 18\mu^2\theta k^2\log k \tag{C.29}$$

with failure probability bounded by

$$4k\exp\left(-\frac{9\theta k\log k}{2\theta k + 2\sqrt{9\theta k\log k}}\right)$$

$$= 4k\exp\left(-\frac{9\log k}{2 + 6\sqrt{(\theta k)^{-1}\log k}}\right) \tag{C.30}$$

$$\leq 4k^{-2} \tag{C.31}$$

The last inequality is derived under the assumption $(\theta k)^{-1}\log k \leq \frac{1}{64}$.

Spiky Component of $\hat{\boldsymbol{\zeta}}_{\text{init}}$. On the other hand,

$$\mathbb{E}\left[\left\|\operatorname{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2\right] = \theta\left\|\operatorname{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\right\|_F^2 \tag{C.32}$$

$$= \theta k. \tag{C.33}$$

For $\mathrm{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i$, applying the moment control Bernstein Inequality, we have

$$\mathbb{P}\left[\left|\left\|\mathrm{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2-\mathbb{E}\left[\cdot\right]\right|\geq t\right]\leq 2\exp\left(-\frac{t^2}{2\theta k+2t}\right). \tag{C.34}$$

By setting $t=2\sqrt{\theta k\log k}$, we obtain that with probability no smaller than $1-k^{-1}$,

$$\left\|\mathrm{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2\geq \theta k-2\sqrt{\theta k\log k}. \tag{C.35}$$

Denote the following events for the entry-wise magnitude

$$\mathcal{E}_j=\left\{|\boldsymbol{e}_j^T\mathrm{offdiag}(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{x}_i|\leq \mu t\right\}, \tag{C.36}$$

and for the support size

$$\mathcal{E}_{\mathrm{supp}}=\left\{\|\boldsymbol{x}_i\|_0\leq 4\theta k\right\}. \tag{C.37}$$

On their intersection $\mathcal{E}_{\mathrm{supp}}\cap\bigcap_{j=1}^{2k}\mathcal{E}_j$, we have

$$\left\|\mathrm{offdiag}(\boldsymbol{A}^T\boldsymbol{A})_{I,I}\boldsymbol{x}_i\right\|_2^2\leq 4\theta k(\mu t)^2. \tag{C.38}$$

The the failure probability can be bounded from the union bound as

$$\begin{aligned}
&\mathbb{P}\left[\left\|\mathrm{offdiag}(\boldsymbol{A}^T\boldsymbol{A})_{I,I}\boldsymbol{x}_i\right\|_2^2\geq 4\theta k(\mu t)^2\right]\\
&\leq \mathbb{P}\left[\left(\mathcal{E}_{\mathrm{supp}}\cap\bigcap_j\mathcal{E}_j\right)^c\right] &\text{(C.39)}\\
&=\mathbb{P}\left[\mathcal{E}_{\mathrm{supp}}^c\cup\bigcup_j\mathcal{E}_j^c\right] &\text{(C.40)}\\
&\leq \mathbb{P}\left[\mathcal{E}_{\mathrm{supp}}^c\right]+\sum_j\mathbb{P}\left[\mathcal{E}_j^c\right] &\text{(C.41)}\\
&\leq \exp(-\theta k)+4k\exp\left(-\frac{t^2}{2\theta k+2t}\right). &\text{(C.42)}
\end{aligned}$$

Therefore, by setting $t^2=9\theta k\log k$, we obtain

$$\left\|\mathrm{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2^2\leq 36\mu^2\theta^2k^2\log k \tag{C.43}$$

with probability no smaller than $1-\exp\left(-\theta k\right)-8k^{-2}$. Therefore, with probability no smaller than $1-k^{-1}-8k^{-2}-\exp\left(-\theta k\right)$,

$$\begin{aligned}
\left\|\mathrm{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2&\geq \theta k-2\sqrt{\theta k\log k} &\text{(C.44)}\\
\left\|\mathrm{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2^2&\leq 36\mu^2\theta^2k^2\log k &\text{(C.45)}
\end{aligned}$$

and via Cauchy-Schwatz inequality, we obtain

$$\begin{aligned}
&\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2^2 &\text{(C.46)}\\
&=\left\|\mathrm{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i+\mathrm{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2^2 &\text{(C.47)}
\end{aligned}$$

$$
\begin{aligned}
&= \left\|\operatorname{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2 + \left\|\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2^2 \\
&\quad + 2\left\langle \operatorname{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i, \operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\rangle && \text{(C.48)} \\
&\geq \left\|\operatorname{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2 \\
&\quad - 2\left\|\operatorname{diag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2 \left\|\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2 && \text{(C.49)} \\
&\geq \theta k\left(1 - 2\sqrt{(\theta k)^{-1}\log k} - 12\mu\sqrt{\theta k \log k}\right) && \text{(C.50)} \\
&\geq \theta k/2. && \text{(C.51)}
\end{aligned}
$$

The last equation is derived by plugging in

$$
(\theta k)^{-1}\log k \leq \tfrac{1}{64}, \quad \mu^2\theta k\log k \leq \tfrac{1}{48^2} \tag{C.52}
$$

under the assumption

$$
64k^{-1}\log k \leq \theta \leq \tfrac{1}{48^2}\mu^{-2}k^{-1}\log^{-1} k. \tag{C.53}
$$

Lower Bound of $\|\cdot\|_4^4$. Since with probability no smaller than $1-4k^{-2}$, $\left\|\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2 \leq 36\mu^2\theta k^2\log k$ obtains and the relative $\|\cdot\|_2^2$ norm between the flat entries to the spiky entries in $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i$ can be bounded as

$$
\begin{aligned}
\frac{\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I^c,I}\boldsymbol{x}_i\right\|_2^2}{\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2^2} &\leq \frac{\left\|\operatorname{offdiag}\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{x}_i\right\|_2^2}{\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2^2} && \text{(C.54)} \\
&\leq 36\mu^2 k\log k \doteq r. && \text{(C.55)}
\end{aligned}
$$

Since

$$
\begin{aligned}
\left\|\hat{\boldsymbol{\zeta}}_{\text{init}}\right\|_4^4 &= \left\|\mathcal{P}_{\mathbb{S}}\left[\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i\right]\right\|_4^4 && \text{(C.56)} \\
&= \frac{1}{\|\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i\|_2^4}\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I^c,I}\boldsymbol{x}_i\right\|_4^4 \\
&\quad + \frac{1}{\|\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i\|_2^4}\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_4^4 && \text{(C.57)} \\
&\geq \frac{1}{\|\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}_i\|_2^4}\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_4^4 && \text{(C.58)} \\
&= \frac{\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right\|_2^4\left\|\mathcal{P}_{\mathbb{S}}\left[\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right]\right\|_4^4}{\left\|\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i + \left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I^c,I}\boldsymbol{x}_i\right\|_2^4} && \text{(C.59)} \\
&\geq \frac{1}{(1+r)^2}\left\|\mathcal{P}_{\mathbb{S}}\left[\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right]\right\|_4^4 && \text{(C.60)}
\end{aligned}
$$

and with high probability $1-\exp(-\theta k)$ according to Lemma A.4, $\mathcal{P}_{\mathbb{S}}\left[\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right]$ satisfies

$$
\left\|\mathcal{P}_{\mathbb{S}}\left[\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right]\right\|_4^4 \geq \frac{1}{\|\boldsymbol{x}_i\|_0} \geq \frac{1}{2\theta(2k-1)}, \tag{C.61}
$$

Together, we have

$$
\|\boldsymbol{\zeta}_{\text{init}}\|_4^4 \geq \left\|\hat{\boldsymbol{\zeta}}_{\text{init}}\right\|_4^4 - \frac{32\kappa^3\delta}{\sigma_{\min}} \tag{C.62}
$$

$$\geq \frac{1}{(1+r)^2}\left\|\mathcal{P}_{\mathbb{S}}\left[\left(\boldsymbol{A}^T\boldsymbol{A}\right)_{I,I}\boldsymbol{x}_i\right]\right\|_4^4 - \frac{640C^{-1/4}}{\theta k\left(1+36\mu^2 k\log k\right)^2} \tag{C.63}$$

$$\geq \left(\frac{1}{4}-\frac{640}{C^{1/4}}\right)\frac{1}{\theta k\left(1+36\mu^2 k\log k\right)^2} \tag{C.64}$$

holds with probability no smaller than $1-k^{-1}-8k^{-2}-2\exp\left(-\theta k\right)-24k\exp\left(-\frac{1}{144}\min\left\{k,3\sqrt{\theta m}\right\}\right)-48k^{-7}-48m^{-5}$. To make sure $\|\boldsymbol{\zeta}_{\text{init}}\|_4^6 \geq 3C_\star\mu\kappa^2$ as desired, we require the sparsity to satisfy

$$\theta \leq \left(\tfrac{1}{4}-\tfrac{640}{C^{1/4}}\right)\left(3C_\star\mu\kappa^2\right)^{-2/3}k^{-1}\left(1+36\mu^2 k\log k\right)^{-2}, \tag{C.65}$$

then the initialization $\boldsymbol{q}_{\text{init}} \in \hat{\mathcal{R}}_{3C_\star}$ follows by Definition 2.1. ■

D Preconditioning

Lemma D.1. *Suppose* $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}\left(\theta\right) \in \mathbb{R}^m$*, then following inequality holds*

$$\left\|\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T-\boldsymbol{I}\right\|_2 \leq 10\sqrt{k\log m/m}, \tag{D.1}$$

with probability no smaller than $1-2\exp\left(-\theta k\right)-24k\exp\left(-\frac{1}{144}\min\left\{k,3\sqrt{\theta m}\right\}\right)-48k^{-7}-48m^{-5}$.

Proof Since

$$\begin{aligned}&\left\|\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T-\boldsymbol{I}\right\|_2\\ &\leq \left\|\text{diag}\left(\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T\right)-\boldsymbol{I}\right\|_2\\ &\quad+\left\|\text{offdiag}\left(\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T\right)\right\|_2,\end{aligned} \tag{D.2}$$

which is bounded by δ with probability no smaller than $1-\varepsilon_d-\varepsilon_o$ whenever the probability that each of the terms is upper bounded by $\delta/2$ satisfies

$$\mathbb{P}\left[\left\|\text{diag}\left(\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T\right)-\boldsymbol{I}\right\|_2 \geq \delta/2\right] \leq \varepsilon_d, \tag{D.3}$$

$$\mathbb{P}\left[\left\|\text{offdiag}\left(\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T\right)-\boldsymbol{I}\right\|_2 \geq \delta/2\right] \leq \varepsilon_o. \tag{D.4}$$

Diagonal of $\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T$**.** Note that $\text{diag}\left(\boldsymbol{X}_0\boldsymbol{X}_0^T\right)=\|\boldsymbol{x}_0\|_2^2\,\boldsymbol{I}$, so

$$\left\|\text{diag}\left(\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T\right)-\boldsymbol{I}\right\|_2 = \left|\frac{1}{\theta m}\|\boldsymbol{x}_0\|_2^2-1\right|. \tag{D.5}$$

We calculate the moment for each summand of $\|\boldsymbol{x}_0\|_2^2$. The summands can be seen as a χ_1^2 random variable but populated with probability θ, whence

$$\mathbb{E}_{x_i\sim\text{BG}(\theta)}\left[\left(x_i^2\right)^p\right]=\theta\,\mathbb{E}_{X_i\sim\chi_1^2}\left[X_i^p\right] \tag{D.6}$$

$$= \theta \frac{\Gamma\left(p+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)} \tag{D.7}$$

$$\leq \frac{\theta p!\,(2)^p}{2} \tag{D.8}$$

$$= \frac{p!}{2}\sigma^2 R^{p-2}. \tag{D.9}$$

Apply Bernstein's inequality for moment bounded random variables (G.4) with $R = 2, \sigma^2 = 4\theta$, then

$$\mathbb{P}\left[\left|\frac{1}{m}\left\|\boldsymbol{x}_0\right\|_2^2 - \theta\right| \geq t\right] \leq 2\exp\left(-\frac{mt^2}{8\theta + 4t}\right). \tag{D.10}$$

By taking $t = \frac{1}{2}\theta\delta$, we obtain

$$\mathbb{P}\left[\left\|\operatorname{diag}\left(\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T\right) - \boldsymbol{I}\right\|_2 \geq \delta/2\right]$$

$$\leq 2\exp\left(-\frac{\theta m\delta^2}{32 + 8\delta}\right) \tag{D.11}$$

$$\leq 2\exp\left(-\frac{100\theta k\log m}{32 + 80\sqrt{k\log m/m}}\right) \tag{D.12}$$

$$\leq 2\exp\left(-\theta k\right). \tag{D.13}$$

Off-diagonal of $\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T$. Note that $\operatorname{offdiag}\left(\boldsymbol{X}_0\boldsymbol{X}_0^T\right)$ is a sub-circulant matrix generated by

$$\boldsymbol{r}_{\boldsymbol{x}_0} = \left[r_{\boldsymbol{x}_0}\left(2k-2\right), \cdots, 0, \cdots, r_{\boldsymbol{x}_0}\left(2k-2\right)\right]^T \tag{D.14}$$

with $r_{\boldsymbol{x}_0}\left(\tau\right) = \left\langle \boldsymbol{x}_0, s_\tau[\boldsymbol{x}_0]\right\rangle$ for $\tau = 1, \cdots, 2k-2$. Equivalently, we can write

$$\boldsymbol{r}_{\boldsymbol{x}_0} = \boldsymbol{R}_{\boldsymbol{x}_0}^T\boldsymbol{x}_0, \tag{D.15}$$

with

$$\boldsymbol{R}_{\boldsymbol{x}_0} = \left[s_{2k-2}[\boldsymbol{x}_0], \cdots, \boldsymbol{0}, \cdots, s_{2k-2}[\boldsymbol{x}_0]\right] \in \mathbb{R}^{m\times(4k-3)}. \tag{D.16}$$

Operator norm of a circulant matrix is defined as the following

$$\left\|\operatorname{offdiag}\left(\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T\right)\right\|_2 = \max_{l=0,\ldots,4k-4}\left|\left\langle \boldsymbol{v}_l, \frac{1}{\theta m}\boldsymbol{r}_{\boldsymbol{x}_0}\right\rangle\right|, \tag{D.17}$$

where $\boldsymbol{v}_l$ is the l-th (discrete) Fourier basis vector

$$\boldsymbol{v}_l = \left[1,\ e^{l\frac{2\pi j}{4k-3}}, \cdots,\ e^{l(4k-4)\frac{2\pi j}{4k-3}}\right]^T, \tag{D.18}$$

and j is the imaginary unit. Let $v_{l,\tau} = \boldsymbol{v}_l\left(2k-2-\tau\right) + \boldsymbol{v}_l\left(2k-2+\tau\right)$, then

$$\left\langle \boldsymbol{v}_l, \boldsymbol{r}_{\boldsymbol{x}_0}\right\rangle = \sum_{\tau=1}^{2k-2} v_{l,\tau}\left\langle \boldsymbol{x}_0, s_\tau[\boldsymbol{x}_0]\right\rangle \tag{D.19}$$

$$= \sum_{\tau=1}^{2k-2} v_{l,\tau}\sum_{i=0}^{m-1}\boldsymbol{x}_0\left(i\right)\boldsymbol{x}_0\left(\left[i+\tau\right]_m\right). \tag{D.20}$$

By decoupling (Theorem 3.4.1 of [DlPG99]), the tail probability of the weighted autocorrelation $\left\langle \boldsymbol{v}_l, \boldsymbol{r}_{\boldsymbol{x}_0}\right\rangle$ can be upper bounded via

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{v}_l, \boldsymbol{r}_{\boldsymbol{x}_0}\right\rangle\right| > t\right]$$

$$= \mathbb{P}\left[\left|\sum_{\tau=1}^{2k-2} v_{l,\tau}\left\langle \boldsymbol{x}_0, s_\tau[\boldsymbol{x}_0]\right\rangle\right| > t\right] \tag{D.21}$$

$$\leq 6\,\mathbb{P}\left[\left|\sum_{\tau=1}^{2k-2} v_{l,\tau}\left\langle \boldsymbol{x}_0, s_\tau[\boldsymbol{x}_0']\right\rangle\right| > \frac{t}{6}\right], \tag{D.22}$$

where $\boldsymbol{x}_0' \sim_{\text{i.i.d.}} \text{BG}(\theta)$ is an independent copy of the random vector $\boldsymbol{x}_0$, we have Plugging in $\langle \boldsymbol{v}_l, \boldsymbol{r}_{\boldsymbol{x}_0}\rangle = \left\langle \boldsymbol{v}_l, \boldsymbol{R}_{\boldsymbol{x}_0}^T \boldsymbol{x}_0\right\rangle = \langle \boldsymbol{R}_{\boldsymbol{x}_0}\boldsymbol{v}_l, \boldsymbol{x}_0\rangle$.

$$\mathbb{P}\left[\left|\left\langle \boldsymbol{v}_l, \frac{1}{\theta m}\boldsymbol{r}_{\boldsymbol{x}_0}\right\rangle\right| > t\right] \leq 6\,\mathbb{P}\left[\left|\frac{1}{\theta m}\left\langle \boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l, \boldsymbol{x}_0\right\rangle\right| > \frac{t}{6}\right]. \tag{D.23}$$

Again with Bernstein's inequality for moment bounded random variable, we have

$$\begin{aligned}&\mathbb{P}\left[\left|\frac{1}{\theta m}\left\langle \boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l, \boldsymbol{x}_0\right\rangle\right| \geq t\right] \\ &\leq 2\exp\left(-\frac{\theta m^2 t^2}{2\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_2^2 + 2\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_\infty mt}\right)\end{aligned} \tag{D.24}$$

Control $\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_2$.

$$\left\|\boldsymbol{R}_{\boldsymbol{x}_0}\boldsymbol{v}_l\right\|_2^2 \leq \left\|\boldsymbol{R}_{\boldsymbol{x}_0}\right\|_2^2 \left\|\boldsymbol{v}_l\right\|_2^2 = k\left\|\boldsymbol{R}_{\boldsymbol{x}_0}\right\|_2^2 \tag{D.25}$$

With tail bound of the operator norm of a circulant matrix in Lemma A.6, we have

$$\mathbb{P}\left[\left\|\boldsymbol{R}_{\boldsymbol{x}_0}\right\|_2 \geq t\right] \leq 4m\exp\left(-\frac{t^2}{2\theta m + 2t}\right) \tag{D.26}$$

Control $\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_\infty$. For a discrete Fourier basis $\boldsymbol{v}_l$ as defined, we have

$$\left\|\boldsymbol{v}_l\right\|_2^2 = \left\|\boldsymbol{v}_l\right\|_0 = 4k-3, \quad \left\|\boldsymbol{v}_l\right\|_\infty = 1 \tag{D.27}$$

Note that

$$\left\|\boldsymbol{R}_{\boldsymbol{x}_0}\boldsymbol{v}_l\right\|_\infty = \max_{\tau=1,\ldots,2k-2}\left|\langle s_\tau[\boldsymbol{x}_0], \boldsymbol{v}_l\rangle\right| \tag{D.28}$$

and moment control Bernstein inequality implies that

$$\mathbb{P}\left[\left|\langle s_\tau[\boldsymbol{x}_0], \boldsymbol{v}_l\rangle\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\theta\left\|\boldsymbol{v}_l\right\|_2^2 + 2\left\|\boldsymbol{v}_l\right\|_\infty t}\right). \tag{D.29}$$

with union bound, we obtain

$$\mathbb{P}\left[\left\|\boldsymbol{R}_{\boldsymbol{x}_0}\boldsymbol{v}_l\right\|_\infty \geq t\right] \leq \sum_{\tau=1}^{2k-2}\mathbb{P}\left[\left|\langle s_\tau[\boldsymbol{x}_0], \boldsymbol{v}_l\rangle\right| \geq t\right] \tag{D.30}$$

$$\leq 4k\exp\left(-\frac{t^2}{8\theta k + 2t}\right) \tag{D.31}$$

Therefore, by plugging in

$$\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_\infty \leq t_1 = 10\sqrt{\theta k\log k}, \tag{D.32}$$

$$\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_2 \leq t_2 = 5\sqrt{\theta m\log m}, \tag{D.33}$$

we obtain the following probabilities

$$\begin{aligned}\mathbb{P}\left[\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_\infty \geq t_1\right] &\leq 4k\exp\left(-\frac{t_1^2}{8\theta k+2t_1}\right)\\ &\leq 4k^{-8}, \qquad \text{(D.34)}\\ \mathbb{P}\left[\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\right\|_2 \geq t_2\right] &\leq 4m\exp\left(-\frac{t_2^2}{2\theta m+2t_2}\right)\\ &\leq 4m^{-6}. \qquad \text{(D.35)}\end{aligned}$$

Denoting event

$$\boldsymbol{E}=\left\{\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_\infty \leq t_1, \left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\right\|_2 \leq t_2\right\}, \qquad \text{(D.36)}$$

and combining these bounds with (D.23), we obtain

$$\begin{aligned}&\mathbb{P}\left[\left\|\operatorname{offdiag}\left(\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T\right)\right\|_2 \geq \delta/2\right]\\ &\leq 6\,\mathbb{P}\left[\max_l\left|\frac{1}{\theta m}\left\langle\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l,\boldsymbol{x}_0\right\rangle\right| \geq \frac{\delta}{12}\right] \qquad \text{(D.37)}\\ &\leq 12k\,\mathbb{P}\left[\left|\frac{1}{\theta m}\left\langle\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l,\boldsymbol{x}_0\right\rangle\right| \geq \frac{\delta}{12}\right] \qquad \text{(D.38)}\\ &\leq 12k\mathbb{P}\left[\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l\right\|_\infty > t_1\right] + 12k\mathbb{P}\left[\left\|\boldsymbol{R}_{\boldsymbol{x}_0'}\right\|_2 > t_2\right]\\ &\quad + 12k\mathbb{P}\left[\left|\frac{1}{\theta m}\left\langle\boldsymbol{R}_{\boldsymbol{x}_0'}\boldsymbol{v}_l,\boldsymbol{x}_0\right\rangle\right| \geq \frac{\delta}{12}\mid\boldsymbol{E}\right] \qquad \text{(D.39)}\\ &\leq 24k\exp\left(-\frac{100\theta km\log m/144}{50\theta m\log m+\frac{200}{12}k\sqrt{\theta m\log k\log m}}\right)\\ &\quad + 12k\left(4k^{-8}+4m^{-6}\right) \qquad \text{(D.40)}\\ &\qquad\qquad \left(t_1=10\sqrt{\theta k\log k},\ t_2=5\sqrt{\theta m\log m}\right)\\ &\leq 24k\exp\left(-\tfrac{1}{144}\min\left\{k,3\sqrt{\theta m}\right\}\right)+48k^{-7}+48m^{-5} \qquad \text{(D.41)}\end{aligned}$$

At last, by combining the control for both the diagonal and off-diagonal term, we obtain that with probability no smaller than $1-2\exp\left(-\theta k\right)-24k\exp\left(-\frac{1}{144}\min\left\{k,3\sqrt{\theta m}\right\}\right)-48k^{-7}-48m^{-5}$,

$$\left\|\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T-\boldsymbol{I}\right\|_2 \leq 10\sqrt{k\log m/m}, \qquad \text{(D.42)}$$

holds and completes the proof. ■

Lemma D.2. *Suppose* $\delta=\left\|\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T-\boldsymbol{I}\right\|_2 \leq 1/\left(2\kappa^2\right)$, *then*

$$\left\|\left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{1/2}\left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}-\boldsymbol{I}\right\|_2 \leq \kappa^2\delta/\sigma_{\min}. \qquad \text{(D.43)}$$

Proof As in by [Bha97], we denote the directional derivative of f at direction $\boldsymbol{\Delta}$ with

$$Df(\boldsymbol{M})\left(\boldsymbol{\Delta}\right)=\left.\frac{d}{dt}\right|_{t=0} f(\boldsymbol{M}+t\boldsymbol{\Delta}), \qquad \text{(D.44)}$$

Denote symmetric matrix $\boldsymbol{M} = \boldsymbol{A}_0\boldsymbol{A}_0^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, with $\lambda_{\max}$ and $\lambda_{\min}$ being its maximum and minimum eigenvalue. Then we have

$$\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{M} + \boldsymbol{\Delta}, \quad \|\boldsymbol{\Delta}\|_2 \leq \lambda_{\max}\delta. \tag{D.45}$$

Then derivative of f with $Df(\boldsymbol{M})$. By differential calculus, we can obtain that

$$\begin{aligned}
&\left\| \left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{1/2} \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2} - \boldsymbol{I} \right\|_2 \\
&= \left\| \left(\boldsymbol{A}_0\boldsymbol{A}_0^T + \boldsymbol{\Delta}\right)^{1/2} \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2} - \boldsymbol{I} \right\|_2 && \text{(D.46)} \\
&= \left\| \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2} \int_{t=0}^{1} Df\left(\boldsymbol{A}_0\boldsymbol{A}_0^T + t\boldsymbol{\Delta}\right)(\boldsymbol{\Delta})\, dt \right\|_2 && \text{(D.47)} \\
&\leq \sup_{t\in[0,1]} \left\| Df\left(\boldsymbol{A}_0\boldsymbol{A}_0^T + t\boldsymbol{\Delta}\right) \right\|_2 \|\boldsymbol{\Delta}\|_2 \left\| \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2} \right\|_2 && \text{(D.48)} \\
&\leq \sup_{t\in[0,1]} \left\| Df\left(\boldsymbol{A}_0\boldsymbol{A}_0^T + t\boldsymbol{\Delta}\right) \right\|_2 \lambda_{\max}\delta/\sigma_{\min} && \text{(D.49)}
\end{aligned}$$

Moreover, we denote $f(t) = t^{1/2}$ and $g(t) = t^2$, then $f = g^{-1}$. The directional derivative of g has following form

$$Dg(\boldsymbol{M})(\boldsymbol{X}) = \boldsymbol{M}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{M}, \tag{D.50}$$

and directional derivative $\boldsymbol{Z} = Df(\boldsymbol{M})(\boldsymbol{X})$ satisfies

$$\boldsymbol{M}\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{M} = \boldsymbol{X}. \tag{D.51}$$

Denote $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ with $\boldsymbol{U}$ orthogonal, without loss of generality,

$$\boldsymbol{\Lambda}\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{\Lambda} = \boldsymbol{X}. \tag{D.52}$$

Applying Theorem VII.2.3 of [Bha97], we have

$$\begin{aligned}
\|Df(\boldsymbol{M})(\boldsymbol{X})\|_2 &= \sup_{\|\boldsymbol{X}\|_2\leq 1} \|\boldsymbol{Z}\|_2 && \text{(D.53)} \\
&\leq \int_{t=0}^{\infty} \left\| e^{-\boldsymbol{\Lambda}t}\boldsymbol{X}e^{-\boldsymbol{\Lambda}t} \right\|_2 dt && \text{(D.54)} \\
&\leq \int_{t=0}^{\infty} e^{-2\lambda_{\min}t} \|\boldsymbol{X}\|_2\, dt && \text{(D.55)}
\end{aligned}$$

and

$$\begin{aligned}
&\sup_{t\in[0,1]} \left\| Df\left(\boldsymbol{A}_0\boldsymbol{A}_0^T + t\boldsymbol{\Delta}\right) \right\|_2 \\
&\leq \frac{\|\boldsymbol{X}\|_2}{2\left(\lambda_{\min} - \lambda_{\max}\delta\right)} && \text{(D.56)} \\
&\leq 1/\lambda_{\min}. && \text{(D.57)}
\end{aligned}$$

Therefore,

$$\left\| \left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{1/2} \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2} - \boldsymbol{I} \right\|_2 \leq \kappa^2\delta/\sigma_{\min}. \tag{D.58}$$

■

Lemma D.3. *Suppose $\boldsymbol{A}_0$ has condition number κ and*

$$\delta = \left\| \frac{1}{\theta m} \boldsymbol{X}_0 \boldsymbol{X}_0^T - \boldsymbol{I} \right\|_2 \leq 1/\left(2\kappa^2\right) \tag{D.59}$$

then

$$\left\| \left(\frac{1}{\theta m} \boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} - \left(\boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \right\|_2 \leq 4\kappa^2 \delta / \sigma_{\min}^2. \tag{D.60}$$

Proof Denote symmetric matrix

$$\boldsymbol{M} = \boldsymbol{A}_0 \boldsymbol{A}_0^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T, \tag{D.61}$$

with $\lambda_{\max}$ and $\lambda_{\min}$ being its maximum and minimum eigenvalue. Then we have

$$\frac{1}{\theta m} \boldsymbol{Y}\boldsymbol{Y}^T = \boldsymbol{M} + \boldsymbol{\Delta}, \quad \|\boldsymbol{\Delta}\|_2 \leq \lambda_{\max} \delta. \tag{D.62}$$

Then

$$\begin{aligned}
&\left\| \left(\frac{1}{\theta m} \boldsymbol{Y}\boldsymbol{Y}^T \right)^{-1/2} - \left(\boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \right\|_2 \\
&= \left\| (\boldsymbol{M} + \boldsymbol{\Delta})^{-1/2} - \boldsymbol{M}^{-1/2} \right\|_2 && \text{(D.63)} \\
&\leq \|\boldsymbol{\Delta}\|_2 \cdot \sup_{0 \leq t \leq 1} \|Df\left(\boldsymbol{M} + t\boldsymbol{\Delta}\right)\|_2 . && \text{(D.64)}
\end{aligned}$$

Here, $f(t) = t^{-1/2}$ and Df is the derivative of function f. In addition, we define function $g(t) = t^{-2}$, $h(t) = t^{-1}$, $w(t) = t^2$, and following function compositions hold

$$f = g^{-1}, \quad g = h \circ w. \tag{D.65}$$

For differential function g and if $Dg\left(f\left(\boldsymbol{M}\right)\right) \neq 0$, we have

$$Df\left(\boldsymbol{M}\right) = \left[Dg\left(f\left(\boldsymbol{M}\right)\right)\right]^{-1}. \tag{D.66}$$

The derivative of function g satisfies the chain rule that

$$Dg\left(\boldsymbol{M}\right) = Dh\left(w\left(\boldsymbol{M}\right)\right)\left(Dw\left(\boldsymbol{M}\right)\right). \tag{D.67}$$

Plug in

$$\begin{aligned}
Dh\left(\boldsymbol{M}\right)\left(\boldsymbol{X}\right) &= -\boldsymbol{M}^{-1}\boldsymbol{X}\boldsymbol{M}^{-1}, && \text{(D.68)} \\
Dw\left(\boldsymbol{M}\right)\left(\boldsymbol{X}\right) &= \boldsymbol{M}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{M}, && \text{(D.69)}
\end{aligned}$$

we obtain that

$$\begin{aligned}
&Dg\left(\boldsymbol{M}\right)\left(\boldsymbol{X}\right) && \text{(D.70)} \\
&= Dh\left(w\left(\boldsymbol{M}\right)\right)\left(Dw\left(\boldsymbol{M}\right)\left(\boldsymbol{X}\right)\right) && \text{(D.71)} \\
&= Dh\left(w\left(\boldsymbol{M}\right)\right)\left[\boldsymbol{M}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{M}\right] && \text{(D.72)} \\
&= Dh\left(\boldsymbol{M}^2\right)\left[\boldsymbol{M}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{M}\right] && \text{(D.73)} \\
&= -\boldsymbol{M}^{-2}\left[\boldsymbol{M}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{M}\right]\boldsymbol{M}^{-2} && \text{(D.74)} \\
&= -\left[\boldsymbol{M}^{-1}\boldsymbol{X}\boldsymbol{M}^{-2} + \boldsymbol{M}^{-2}\boldsymbol{X}\boldsymbol{M}^{-1}\right]. && \text{(D.75)}
\end{aligned}$$

Since the function g is differentiable and $Dg(\boldsymbol{M}) \neq \mathbf{0}$, then

$$Df\left(\boldsymbol{M}\right) = \left[Dg\left(f\left(\boldsymbol{M}\right)\right)\right]^{-1} \tag{D.76}$$

$$= \left[Dg\left(\boldsymbol{M}^{-1/2}\right)\right]^{-1}. \tag{D.77}$$

Hence, directional derivative $\boldsymbol{Z} \doteq Df\left(\boldsymbol{M}\right)\left(\boldsymbol{X}\right)$ satisfies

$$\boldsymbol{M}^{1/2}\boldsymbol{Z}\boldsymbol{M} + \boldsymbol{M}\boldsymbol{Z}\boldsymbol{M}^{1/2} = -\boldsymbol{X}. \tag{D.78}$$

Denote $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ with $\boldsymbol{\Lambda} \succ 0$ and $\boldsymbol{U}$ orthogonal, without loss of generality

$$\boldsymbol{\Lambda}\boldsymbol{Z}\boldsymbol{\Lambda}^{1/2} + \boldsymbol{\Lambda}^{1/2}\boldsymbol{Z}\boldsymbol{\Lambda} = -\boldsymbol{X}. \tag{D.79}$$

Above equation can be reformulated as a Sylvester equation as following

$$\boldsymbol{\Lambda}^{1/2}\boldsymbol{Z} - \boldsymbol{Z}\left(-\boldsymbol{\Lambda}^{1/2}\right) = -\boldsymbol{\Lambda}^{-1/2}\boldsymbol{X}\boldsymbol{\Lambda}^{-1/2}. \tag{D.80}$$

From Theorem VII.2.3 of [Bha97], when there are no common eigenvalues of $\boldsymbol{\Lambda}^{1/2}$ and $-\boldsymbol{\Lambda}^{1/2}$, then there exists a closed form solution for matrix $\boldsymbol{Z}$ that

$$\boldsymbol{Z} = \int_{t=0}^{\infty} e^{-\boldsymbol{\Lambda}^{1/2}t}\left(-\boldsymbol{\Lambda}^{-1/2}\boldsymbol{X}\boldsymbol{\Lambda}^{-1/2}\right)e^{-\boldsymbol{\Lambda}^{1/2}t}dt \tag{D.81}$$

Therefore, the operator norm of $Df\left(\boldsymbol{M}\right)$ can be obtained as

$$\left\|Df(\boldsymbol{M})(\boldsymbol{X})\right\|_2 = \sup_{\|\boldsymbol{X}\|_2 \leq 1} \|\boldsymbol{Z}\|_2 \tag{D.82}$$

$$\leq \int_{t=0}^{\infty} \left\| e^{-\boldsymbol{\Lambda}^{1/2}t}\left(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{X}\boldsymbol{\Lambda}^{-1/2}\right)e^{-\boldsymbol{\Lambda}^{1/2}t}\right\| dt \tag{D.83}$$

$$\leq \int_{t=0}^{\infty} e^{-\lambda_{\min}t}\left\|\boldsymbol{\Lambda}^{-1/2}\boldsymbol{X}\boldsymbol{\Lambda}^{-1/2}\right\| dt \tag{D.84}$$

$$\leq \frac{\|\boldsymbol{X}\|}{\lambda_{\min}^2}. \tag{D.85}$$

Therefore

$$\left\|(\boldsymbol{M}+\boldsymbol{\Delta})^{-1/2} - \boldsymbol{M}^{-1/2}\right\|_2$$

$$\leq \frac{\|\boldsymbol{\Delta}\|_2}{\left(\lambda_{\min} - \|\boldsymbol{\Delta}\|_2\right)^2} \tag{D.86}$$

$$\leq \frac{4\|\boldsymbol{\Delta}\|_2}{\lambda_{\min}^2} \qquad \left(\delta \leq 1/\left(2\kappa^2\right)\right) \tag{D.87}$$

$$\leq \frac{4\lambda_{\max}\delta}{\lambda_{\min}^2} \tag{D.88}$$

$$= \frac{4\kappa^2\delta}{\sigma_{\min}^2} \tag{D.89}$$

■

E Concentration for Gradient (Lemma 4.2)

Lemma E.1. *Suppose* $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$. *There exists a positive constant* C *such that whenever*

$$m \geq C \frac{\min\left\{(2C_\star\mu)^{-1}; \kappa^2 k^2\right\}}{(1-\theta)^2 \sigma_{\min}^2} \kappa^8 k^4 \log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right) \tag{E.1}$$

and $\theta > \log k/k$, *then with probability no smaller than* $1 - c_1\exp(-k) - c_2 k^{-4} - 2\exp(-\theta k) - 24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right) - 48k^{-7} - 48m^{-5}$,

$$\left\|\mathrm{grad}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\mathrm{grad}[\varphi](\boldsymbol{q})\right\|_2 \leq c\frac{1-\theta}{\theta m^2}\frac{\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6}{\kappa^2}, \tag{E.2}$$

holds for all $\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}$ *with positive constant* $c \leq 3/(2C_\star)$.

Proof Denote $\boldsymbol{\eta} = \boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}$ and $\bar{\boldsymbol{\eta}} = \boldsymbol{Y}^T\left(\theta m \boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{q} = (\theta m)^{-1/2}\boldsymbol{X}_0^T\boldsymbol{\zeta}$, then

$$\begin{aligned}
&\left\|\mathrm{grad}\left[\psi\right](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\mathrm{grad}\left[\varphi\right](\boldsymbol{q})\right\|_2 \\
&= \left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\frac{1}{m}\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3} - \frac{3(1-\theta)}{\theta m^2}\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3}\right]\right\|_2 \\
&\leq \underbrace{\frac{1}{m}\left\|\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3} - (\theta m)^{-1/2}\boldsymbol{A}\boldsymbol{X}_0\boldsymbol{\eta}^{\circ 3}\right\|_2}_{\Delta_1^g} \\
&\quad + \underbrace{\frac{1}{\theta^{1/2}m^{3/2}}\left\|\boldsymbol{A}\boldsymbol{X}_0\boldsymbol{\eta}^{\circ 3} - \boldsymbol{A}\boldsymbol{X}_0\bar{\boldsymbol{\eta}}^{\circ 3}\right\|_2}_{\Delta_2^g} \\
&\quad + \underbrace{\left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\frac{1}{\theta^{1/2}m^{3/2}}\boldsymbol{A}\boldsymbol{X}_0\bar{\boldsymbol{\eta}}^{\circ 3} - \frac{3(1-\theta)}{\theta m^2}\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3}\right]\right\|_2}_{\Delta_3^g}.
\end{aligned}$$

First, let us note that

$$\begin{aligned}
&C(1-\theta)^{-2}\sigma_{\min}^{-2}\kappa^{10}k^6\log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right) \\
&\leq C\left(\frac{\kappa k}{\sigma_{\min}(1-\theta)}\right)^{10}\log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right) &\text{(E.3)}\\
&\leq C\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right)^{13}, &\text{(E.4)}
\end{aligned}$$

hence

$$\begin{aligned}
&\frac{\log^3\left(C(1-\theta)^{-2}\sigma_{\min}^{-2}\kappa^{10}k^6\log^3\left((1-\theta)^{-1}\sigma_{\min}^{-1}\kappa k\right)\right)}{C\log^3\left((1-\theta)^{-1}\sigma_{\min}^{-1}\kappa k\right)} \\
&\leq \left(\frac{\log C + 13\log\left((1-\theta)^{-1}\sigma_{\min}^{-1}\kappa k\right)}{C^{1/3}\log\left((1-\theta)^{-1}\sigma_{\min}^{-1}\kappa k\right)}\right)^3 &\text{(E.5)}
\end{aligned}$$

$$\leq \left(\frac{\log C}{C^{1/3} \log\left((1-\theta)^{-1} \sigma_{\min}^{-1} \kappa k \right)} + \frac{13}{C^{1/3}} \right)^3 \tag{E.6}$$

$$\leq \left(\frac{1}{C^{1/6}} + \frac{1}{2} \frac{1}{C^{1/6}} \right)^3 \qquad \left(C \geq 10^8 \right) \tag{E.7}$$

$$\leq \frac{4}{C^{1/2}}. \tag{E.8}$$

Given

$$m \geq C \frac{\min\left\{ (2C_\star \mu)^{-1}, \kappa^2 k^2 \right\}}{(1-\theta)^2 \sigma_{\min}^2} \kappa^8 k^4 \log^3\left(\frac{\kappa k}{\sigma_{\min}(1-\theta)} \right), \tag{E.9}$$

as the ratio $\log^3 m/m$ decreases with increasing m, then

$$\frac{\log^3 m}{m} \leq \frac{\log^3\left(\frac{C\kappa^{10} k^6}{(1-\theta)^2 \sigma_{\min}^2} \log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}} \right) \right)}{C \log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}} \right)} \times \frac{(1-\theta)^2 \sigma_{\min}^2}{\min\left\{ (2C_\star \mu)^{-1}, \kappa^2 k^2 \right\} \kappa^8 k^4} \tag{E.10}$$

$$\leq \frac{4}{C^{1/2}} \frac{(1-\theta)^2 \sigma_{\min}^2}{\min\left\{ (2C_\star \mu)^{-1}, \kappa^2 k^2 \right\} \kappa^8 k^4} \tag{E.11}$$

According to Lemma D.1, following inequality always holds

$$\left\| \frac{1}{\theta m} \boldsymbol{X}_0 \boldsymbol{X}_0^T - \boldsymbol{I} \right\|_2 \leq \delta \tag{E.12}$$

$$\leq 10\sqrt{k \log m/m} \tag{E.13}$$

$$\leq \frac{20(1-\theta)\sigma_{\min} \max\left\{ (2C_\star \mu)^{1/2}, (\kappa k)^{-1} \right\}}{C^{1/4} \kappa^4 k^{3/2} \log m} \tag{E.14}$$

$$\leq \frac{20\sigma_{\min}}{C^{1/4}\kappa^3} \frac{(1-\theta)\left\| \boldsymbol{A}^T \boldsymbol{q} \right\|_4^6}{\kappa^2 k \log m}, \qquad \forall \boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}. \tag{E.15}$$

with probability no smaller than $1 - \varepsilon_0$ with $\varepsilon_0 = 2\exp(-\theta k) + 24k \exp\left(-\frac{1}{144} \min\left\{ k, 3\sqrt{\theta m} \right\} \right) + 48k^{-7} + 48m^{-5}$.

Moreover, $4\kappa^3 \delta/\sigma_{\min} \leq 1/2$ whenever

$$C \geq \left(\frac{160(1-\theta)}{k \log m} \right)^4, \tag{E.16}$$

whence $\delta \leq 1/\left(8\kappa^2\right)$, and Lemma D.3 implies that

$$\left\| \left(\frac{1}{\theta m} \boldsymbol{Y} \boldsymbol{Y}^T \right)^{-1/2} \boldsymbol{A}_0 - \left(\boldsymbol{A}_0 \boldsymbol{A}_0^T \right)^{-1/2} \boldsymbol{A}_0 \right\|_2$$

$$\leq 4\kappa^3 \delta/\sigma_{\min} \tag{E.17}$$

$$\leq \frac{80(1-\theta)}{C^{1/4} k \log m} \frac{\left\| \boldsymbol{A}^T \boldsymbol{q} \right\|_4^6}{\kappa^2}, \qquad \forall \boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}. \tag{E.18}$$

At the same time,

$$\|\boldsymbol{X}_0\|_2 \leq (\theta m)^{1/2} \sqrt{1+\delta} \leq (\theta m)^{1/2} (1+\delta/2)\,. \tag{E.19}$$

Moreover, Lemma A.5 implies that with probability no smaller than $1-\varepsilon_B$, we have

$$\|\boldsymbol{x}_0\|_\infty \leq \sqrt{2}\log^{1/2}\left(\frac{2\theta m}{\varepsilon_B}\right). \tag{E.20}$$

Upper Bound for Δ_1^g. Using Lemma A.7, on the an event of probability at least $1-\varepsilon_0-\varepsilon_B$,

$$\begin{aligned}
\left\|\boldsymbol{\eta}^{\circ 3}\right\|_2 &= \|\boldsymbol{\eta}\|_6^3 && \text{(E.21)}\\
&\leq \left(1+\frac{4\kappa^3\delta}{\sigma_{\min}}\right)^2 \frac{2k}{\theta m}\|\boldsymbol{x}_0\|_\infty^2 && \text{(E.22)}\\
&\leq \frac{9k}{\theta m}\log\left(\frac{2\theta m}{\varepsilon_B}\right). && \text{(E.23)}
\end{aligned}$$

Therefore, we can obtain following upper bound

$$\begin{aligned}
\Delta_1^g &= \frac{1}{m}\left\|\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}\boldsymbol{\eta}^{\circ 3} - (\theta m)^{-1/2}\boldsymbol{A}\boldsymbol{X}_0\boldsymbol{\eta}^{\circ 3}\right\|_2 && \text{(E.24)}\\
&\leq \frac{1}{\theta^{1/2}m^{3/2}}\|\boldsymbol{X}_0\|_2\left\|\boldsymbol{\eta}^{\circ 3}\right\|_2 \times \\
&\qquad \left\|\left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{A}_0 - \boldsymbol{A}\right\|_2 && \text{(E.25)}\\
&\leq \frac{5}{4m}\cdot\frac{4\kappa^3\delta}{\sigma_{\min}}\cdot\frac{9k}{\theta m}\log\left(\frac{2\theta m}{\varepsilon_B}\right) && \text{(E.26)}\\
&\leq \frac{900\,(1-\theta)\log\left(2\theta m/\varepsilon_B\right)}{C^{1/4}\theta m^2\log m}\frac{\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6}{\kappa^2} \quad \forall \boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}.
\end{aligned}$$

Upper Bound for Δ_2^g. Similarly, with probability no smaller than $1-\varepsilon_0-\varepsilon_B$, together with Lemma A.7, following upper bound can be obtained

$$\begin{aligned}
&\left\|\boldsymbol{\eta}^{\circ 3}-\bar{\boldsymbol{\eta}}^{\circ 3}\right\|_2 \\
&= \left\|\boldsymbol{\eta}^{\circ 3} - \operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2}\right)\bar{\boldsymbol{\eta}} + \operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2}\right)\bar{\boldsymbol{\eta}} - \bar{\boldsymbol{\eta}}^{\circ 3}\right\|_2 && \text{(E.27)}\\
&\leq \|\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}\|_2\left\|\operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2}\right)\right\|_2 + \|\bar{\boldsymbol{\eta}}\|_2\left\|\operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2}-\bar{\boldsymbol{\eta}}^{\circ 2}\right)\right\|_2 && \text{(E.28)}\\
&= \|\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}\|_2\|\boldsymbol{\eta}\|_\infty^2 + \|\bar{\boldsymbol{\eta}}\|_2\left\|\boldsymbol{\eta}^{\circ 2}-\bar{\boldsymbol{\eta}}^{\circ 2}\right\|_\infty && \text{(E.29)}\\
&\leq \|\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}\|_2\|\boldsymbol{\eta}\|_\infty^2 + \|\bar{\boldsymbol{\eta}}\|_2\|\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}\|_\infty\|\boldsymbol{\eta}+\bar{\boldsymbol{\eta}}\|_\infty && \text{(E.30)}\\
&\leq 4\,(1+\delta/2)\,\frac{4\kappa^3\delta}{\sigma_{\min}}\frac{k}{\theta m}\log\left(2\theta m/\varepsilon_B\right)\times \\
&\qquad \left[\left(1+\frac{4\kappa^3\delta}{\sigma_{\min}}\right)^2 + \left(2+\frac{4\kappa^3\delta}{\sigma_{\min}}\right)\right] && \text{(E.31)}\\
&\leq \frac{24k}{\theta m}\log\left(2\theta m/\varepsilon_B\right)\cdot\frac{4\kappa^3\delta}{\sigma_{\min}}. && \text{(E.32)}
\end{aligned}$$

Therefore, we can obtain following upper bound

$$\Delta_2^g = \frac{1}{\theta^{1/2}m^{3/2}}\left\|\boldsymbol{A}\boldsymbol{X}_0^T\boldsymbol{\eta}^{\circ 3} - \boldsymbol{A}\boldsymbol{X}_0^T\bar{\boldsymbol{\eta}}^{\circ 3}\right\|_2 \tag{E.33}$$

$$\leq \frac{1}{\theta^{1/2}m^{3/2}} \|\boldsymbol{A}\|_2 \|\boldsymbol{X}_0\|_2 \left\|\boldsymbol{\eta}^{\circ 3} - \bar{\boldsymbol{\eta}}^{\circ 3}\right\|_2 \tag{E.34}$$

$$\leq \frac{5}{4m} \cdot \frac{24k}{\theta m} \log\left(2\theta m/\varepsilon_B\right) \cdot \frac{4\kappa^3\delta}{\sigma_{\min}} \tag{E.35}$$

$$\leq \frac{2400}{C^{1/4}} \frac{1-\theta}{\theta m^2} \frac{\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6}{\kappa^2} \cdot \frac{\log\left(2\theta m/\varepsilon_B\right)}{\log m}. \tag{E.36}$$

For both Δ_1^g and Δ_2^g to be bounded by $\frac{1}{2C_\star}\frac{1-\theta}{\theta m^2}\frac{\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6}{\kappa^2}$, we set

$$C \geq \left(4800 C_\star \frac{\log\left(2\theta m/\varepsilon_B\right)}{\log m}\right)^4. \tag{E.37}$$

Notice that the right hand side is indeed bounded by a numerical constant for all m.

Tail Bound for Δ_3^g. Note that

$$\left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{Y}\bar{\boldsymbol{\eta}}^{\circ 3}$$

$$= \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{A}_0\boldsymbol{X}_0\left(\boldsymbol{Y}^T\left(\theta m\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{q}\right)^{\circ 3} \tag{E.38}$$

$$= (\theta m)^{-3/2}\boldsymbol{A}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 3}, \tag{E.39}$$

and its expectation with respect to $\boldsymbol{x}_0$

$$\mathbb{E}\left[\frac{1}{m}\boldsymbol{A}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 3}\right]$$

$$= \mathbb{E}\left[\boldsymbol{A}\boldsymbol{x}_i\left(\boldsymbol{x}_i^T\boldsymbol{A}^T\boldsymbol{q}\right)^3\right] \tag{E.40}$$

$$= 3\theta\left(1-\theta\right)\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} + 3\theta^2\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_2^2\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{q} \tag{E.41}$$

$$= 3\theta\left(1-\theta\right)\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} + 3\theta^2\boldsymbol{q}, \tag{E.42}$$

hence

$$\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\mathbb{E}\left[\frac{1}{m}\boldsymbol{A}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 3}\right]\right]$$

$$= \boldsymbol{P}_{\boldsymbol{q}^\perp}\left[3\theta\left(1-\theta\right)\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3}\right]. \tag{E.43}$$

Therefore, the Δ_3^g term can be simplified as

$$\Delta_3^g = \left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\frac{1}{\theta^{1/2}m^{3/2}}\boldsymbol{A}\boldsymbol{X}_0^T\bar{\boldsymbol{\eta}}^{\circ 3} - \frac{3\left(1-\theta\right)}{\theta m^2}\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3}\right]\right\|_2 \tag{E.44}$$

$$= \frac{1}{\theta^2 m^2}\left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\frac{\boldsymbol{A}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{\zeta}\right)^{\circ 3}}{m} - 3\theta\left(1-\theta\right)\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3}\right]\right\|_2 \tag{E.45}$$

$$\leq \frac{1}{\theta^2 m^2}\left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\frac{\boldsymbol{A}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{\zeta}\right)^{\circ 3}}{m} - \mathbb{E}\left[\cdot\right]\right]\right\|_2$$

$$+ \frac{1}{\theta^2 m^2}\left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[3\theta^2\boldsymbol{q}\right]\right\|_2 \tag{E.46}$$

$$\leq \frac{1}{\theta^2 m^2}\left\|\frac{1}{m}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{\zeta}\right)^{\circ 3} - \mathbb{E}\left[\cdot\right]\right\|_2. \tag{E.47}$$

Under the assumption that

$$m \geq \frac{C}{(1-\theta)^2} \min\left\{\mu^{-1}, \kappa^2 k^2\right\} \kappa^2 k^4 \log^3(\kappa k), \tag{E.48}$$

applying Lemma E.2, we have

$$\left\| \frac{1}{m} \boldsymbol{X}_0 \left(\boldsymbol{X}_0^T \boldsymbol{A}^T \boldsymbol{q}\right)^{\circ 3} - \mathbb{E}[\cdot] \right\|_2 \leq c\theta(1-\theta) \frac{\left\|\boldsymbol{A}^T \boldsymbol{q}\right\|_4^6}{\kappa^2}. \tag{E.49}$$

with probability larger than $1 - c_2 \exp(-k) - c_2 k^{-4}$. At last, taking $\varepsilon_B = \theta^2 k^{-4}$, we obtain that

$$\begin{aligned} &\left\| \operatorname{grad}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2} \operatorname{grad}[\varphi](\boldsymbol{q}) \right\|_2 \\ &\leq c \frac{1-\theta}{\theta m^2} \frac{\left\|\boldsymbol{A}^T \boldsymbol{q}\right\|_4^6}{\kappa^2}, \qquad \forall \boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star} \end{aligned} \tag{E.50}$$

with probability larger than $1 - c_2 \exp(-k) - c_2 k^{-4} - \varepsilon_B - \varepsilon_0$ as desired. ■

E.1 Proof of Lemma E.2

Lemma E.2. *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^m$. There exist positive constant C such that whenever*

$$m \geq \frac{C}{(1-\theta)^2} \min\left\{(2C_\star \mu)^{-1}, \kappa^2 k^2\right\} \kappa^2 k^4 \log^3(\kappa k) \tag{E.51}$$

and $\theta k \geq 1$, then with probability no smaller than $1 - c_1 \exp(-k) - c_2 k^{-4}$,

$$\left\| \frac{1}{m} \boldsymbol{X}_0 \left(\boldsymbol{X}_0^T \boldsymbol{A}^T \boldsymbol{q}\right)^{\circ 3} - \mathbb{E}[\cdot] \right\|_2 \leq c\theta(1-\theta) \frac{\left\|\boldsymbol{A}^T \boldsymbol{q}\right\|_4^6}{\kappa^2} \tag{E.52}$$

holds for all $\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}$ with positive constant $c \leq 1/(2C_\star)$.

Proof Let $\bar{\boldsymbol{x}}_i \in \mathbb{R}^{2k-1}$ be generated via

$$\bar{\boldsymbol{x}}_i = \begin{cases} \boldsymbol{x}_i & \|\boldsymbol{x}_i\|_\infty \leq B \text{ and } \|\boldsymbol{x}_i\|_0 \leq 4\theta k \log m \\ \boldsymbol{0} & \text{else} \end{cases} \tag{E.53}$$

Let $\bar{\boldsymbol{X}}_0 \in \mathbb{R}^{(2k-1)\times m}$ denote the circulant submatrix generated by $\bar{\boldsymbol{x}}_0$. Then $\bar{\boldsymbol{X}}_0 = \boldsymbol{X}_0$ obtains whenever

1. $\|\boldsymbol{x}_0\|_\infty \leq B$, which happens with probability no smaller than $1 - 2\theta m e^{-B^2/2}$ according to Lemma A.5;
2. $\|\boldsymbol{x}_i\|_0 \leq 4\theta k \log m$ holds for any index i, applying Lemma A.4 and Boole's inequality we have

$$\begin{aligned} &\mathbb{E}\left[\mathbf{1}_{\bigcup_i \|\boldsymbol{x}_i\|_0 > 4\theta k \log m}\right] \\ &\leq m\mathbb{P}\left[\|\boldsymbol{x}_i\|_0 > 4\theta k \log m\right] &\text{(E.54)} \\ &\leq 2m \exp\left(-\tfrac{3}{4}\theta k \log m\right). &\text{(E.55)} \end{aligned}$$

Denote $\boldsymbol{\zeta} = \boldsymbol{A}^T \boldsymbol{q}$ and

$$\boldsymbol{g}_E = \mathbb{E}\left[\frac{1}{m} \boldsymbol{X}_0 \left(\boldsymbol{X}_0^T \boldsymbol{A}^T \boldsymbol{q}\right)^{\circ 3}\right], \tag{E.56}$$

$$\bar{\boldsymbol{g}}_E = \mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 3}\right], \tag{E.57}$$

then,

$$\begin{aligned}
&\mathbb{P}\left[\left\|\frac{1}{m}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{\zeta}\right)^{\circ 3} - \boldsymbol{g}_E\right\|_2 \geq c\theta\left(1-\theta\right)\frac{\|\boldsymbol{\zeta}\|_4^6}{\kappa^2}\right] \\
&\leq \mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3} - \boldsymbol{g}_E\right\|_2 \geq c\theta\left(1-\theta\right)\frac{\|\boldsymbol{\zeta}\|_4^6}{\kappa^2}\right] \\
&\quad + \ 2\theta m e^{-B^2/2} + 2m\exp\left(-\tfrac{3}{4}\theta k\log m\right)
\end{aligned} \tag{E.58}$$

With triangle inequality, we have

$$\begin{aligned}
&\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3} - \boldsymbol{g}_E\right\|_2 \\
&\leq \left\|\mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3}\right] - \bar{\boldsymbol{g}}_E\right\|_2 + \left\|\bar{\boldsymbol{g}}_E - \boldsymbol{g}_E\right\|_2.
\end{aligned} \tag{E.59}$$

Hence, provided

$$\left\|\bar{\boldsymbol{g}}_E - \boldsymbol{g}_E\right\|_2 \leq \frac{c}{2}\theta(1-\theta)\frac{\|\boldsymbol{\zeta}\|_4^6}{\kappa^2}, \tag{E.60}$$

we have

$$\begin{aligned}
&\mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3} - \boldsymbol{g}_E\right\|_2 \geq c\theta\left(1-\theta\right)\frac{\|\boldsymbol{\zeta}\|_4^6}{\kappa^2}\right] \\
&\leq \mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3} - \bar{\boldsymbol{g}}_E\right\|_2 \geq \frac{c}{2}\theta\left(1-\theta\right)\frac{\|\boldsymbol{\zeta}\|_4^6}{\kappa^2}\right].
\end{aligned} \tag{E.61}$$

Truncation Level Next, we choose a large enough entry-wise truncation level B such that the expectation of the gradient $\mathbb{E}\left[\frac{1}{m}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{\zeta}\right)^{\circ 3}\right]$ is close to that of its truncation $\mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3}\right]$.

Moreover, we introduce following events notation

$$\mathcal{E}_i \doteq \left\{\|\boldsymbol{x}_i\|_\infty > B \ \cup \ \|\boldsymbol{x}_i\|_0 > 4\theta k\log m\right\}, \tag{E.62}$$

then

$$\begin{aligned}
&\left\|\bar{\boldsymbol{g}}_E - \boldsymbol{g}_E\right\|_2 \\
&= \left\|\mathbb{E}\left[\frac{1}{m}\sum_i \boldsymbol{x}_i\left\langle\boldsymbol{x}_i,\boldsymbol{\zeta}\right\rangle^3\cdot\mathbf{1}_{\mathcal{E}_i}\right]\right\|_2 && \text{(E.63)} \\
&\leq \frac{1}{m}\sum_i\left\|\mathbb{E}\left[\boldsymbol{x}_i\left\langle\boldsymbol{x}_i,\boldsymbol{\zeta}\right\rangle^3\cdot\mathbf{1}_{\mathcal{E}_i}\right]\right\|_2 && \text{(E.64)} \\
&\leq \frac{1}{m}\sum_i\left(\mathbb{E}\left[\left\|\boldsymbol{x}_i\left(\boldsymbol{x}_i^T\boldsymbol{\zeta}\right)^{\circ 3}\right\|_2^2\right]\cdot\mathbb{E}\left[\mathbf{1}_{\mathcal{E}_i}\right]\right)^{1/2} && \text{(E.65)} \\
&\leq \left(\mathbb{E}\left[\left\|\boldsymbol{x}_i\right\|_2^8\right]\right)^{1/2}\times \\
&\quad\sqrt{\mathbb{E}\left[\mathbf{1}_{\|\boldsymbol{x}_i\|_\infty>B}\right]+\mathbb{E}\left[\mathbf{1}_{\|\boldsymbol{x}_i\|_0>4\theta k\log m}\right]} && \text{(E.66)}
\end{aligned}$$

$$\leq 50k^2\sqrt{4\theta k e^{-B^2/2}+\exp\left(-\tfrac{3}{4}\theta k\log m\right)} \tag{E.67}$$

By setting

$$B \geq C'\log^{1/2}\left(\frac{\kappa^4 k^8}{\theta\left(1-\theta\right)^2}\right), \tag{E.68}$$

we have

$$\theta k e^{-B^2/2} \leq \frac{1}{2}\left(\frac{c}{100}\right)^2\theta^2\left(1-\theta\right)^2\frac{\|\boldsymbol{\zeta}\|_4^{12}}{\kappa^4 k^4}. \tag{E.69}$$

In addition, whenever

$$\theta k \geq \frac{4}{3\log m}\log\left(\frac{400^2\kappa^4 k^4}{c^2\theta^2\left(1-\theta\right)^2\|\boldsymbol{\zeta}\|_4^{12}}\right), \tag{E.70}$$

we have

$$\exp\left(-\tfrac{3}{4}\theta k\log m\right) \leq \frac{1}{2}\left(\frac{c}{100}\right)^2\theta^2\left(1-\theta\right)^2\frac{\|\boldsymbol{\zeta}\|_4^{12}}{\kappa^4 k^4}. \tag{E.71}$$

Therefore,

$$\sqrt{4\theta k e^{-B^2/2}+\exp\left(-\tfrac{3}{4}\theta k\log m\right)} \leq \frac{c}{2}\theta\left(1-\theta\right)\frac{\|\boldsymbol{\zeta}\|_4^6}{50\kappa^2 k^2}. \tag{E.72}$$

In addition,

$$\left(\mathbb{E}\left[\|\boldsymbol{x}_i\|_2^8\right]\right)^{1/2} \leq \left(7!!\cdot 2^4 k^4\right)^{1/2} < 50k^2. \tag{E.73}$$

Plugging in Eq (E.73) and (E.72) back to (E.67), we obtain that

$$\|\bar{\boldsymbol{g}}_E-\boldsymbol{g}_E\|_2 \leq \frac{c}{2}\theta\left(1-\theta\right)\frac{\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6}{\kappa^2}, \tag{E.74}$$

and hence

$$\begin{aligned}&\mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3}-\boldsymbol{g}_E\right\|_2 \geq c\theta\left(1-\theta\right)\frac{\|\boldsymbol{\zeta}\|_4^6}{\kappa^2}\right]\\ &\leq \mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3}-\bar{\boldsymbol{g}}_E\right\|_2 \geq \frac{c}{2}\theta\left(1-\theta\right)\frac{\|\boldsymbol{\zeta}\|_4^6}{\kappa^2}\right].\end{aligned} \tag{E.75}$$

Independent Submatrices. To deal with the complicated dependence within the random circulant matrix $\boldsymbol{X}_0$, we break $\boldsymbol{X}_0$ into submatrices $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{2k-1}$, each of which is (marginally) distributed as a $(2k-1)\times\frac{m}{2k-1}$ i.i.d. $\mathrm{BG}(\theta)$ random matrix. Indeed, there exists a permutation $\boldsymbol{\Pi}$ such that

$$\boldsymbol{X}_0\boldsymbol{\Pi} = \left[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{2k-1}\right], \tag{E.76}$$

with

$$\boldsymbol{X}_i = \left[\boldsymbol{x}_i,\boldsymbol{x}_{i+(2k-1)},\cdots,\boldsymbol{x}_{i+(m-2k-1)}\right]. \tag{E.77}$$

We apply similar matrix breaking approach for the truncated matrix $\bar{\boldsymbol{X}}$. The summands within each term $\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 3}$ are mutually independent and hence is amenable to classical concentration results.

$$\frac{1}{m}\bar{\boldsymbol{X}}_0\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 3} = \frac{1}{m}\sum_{l=1}^{m}\left\langle\bar{\boldsymbol{x}}_l,\boldsymbol{\zeta}\right\rangle^3\bar{\boldsymbol{x}}_l \tag{E.78}$$

$$= \sum_{i=1}^{2k-1}\frac{1}{m}\left(\sum_{j=0}^{\frac{m}{2k-1}-1}\left\langle\bar{\boldsymbol{x}}_{i+(2k-1)j},\boldsymbol{\zeta}\right\rangle^3\bar{\boldsymbol{x}}_{i+(2k-1)j}\right) \tag{E.79}$$

$$= \sum_{i=1}^{2k-1} \frac{1}{m} \bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 3}. \tag{E.80}$$

We conservatively bound the quantity of interest, $\frac{1}{m}\bar{\boldsymbol{X}}_0 \left(\bar{\boldsymbol{X}}_0^T \boldsymbol{\zeta}\right)^{\circ 3}$, by ensuring that for each k, $\bar{\boldsymbol{X}}_k \left(\bar{\boldsymbol{X}}_k^T \boldsymbol{\zeta}\right)^{\circ 3}$ be close to its expectation.

$$\begin{aligned}
&\mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0 \left(\bar{\boldsymbol{X}}_0^T \boldsymbol{\zeta}\right)^{\circ 3} - \bar{\boldsymbol{g}}_E\right\|_2 \geq \frac{c}{2}\theta(1-\theta)\frac{\|\boldsymbol{\zeta}\|_4^6}{\kappa^2}\right] \\
&\leq \sum_{i=1}^{2k-1} \mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 3} - \frac{\bar{\boldsymbol{g}}_E}{2k-1}\right\|_2 \geq \frac{c}{2}\frac{\theta(1-\theta)\|\boldsymbol{\zeta}\|_4^6}{\kappa^2(2k-1)}\right] \\
&= \sum_{i=1}^{2k-1} \mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 3} - \bar{\boldsymbol{g}}_E\right\|_2 \geq \frac{c}{2}\frac{\theta(1-\theta)\|\boldsymbol{\zeta}\|_4^6}{\kappa^2(2k-1)}\right]
\end{aligned}$$

Applying Bernstein inequality for matrix variables as in Lemma G.7, with $d_1 = 2k-1$, $d_2 = 1$, we can obtain that for independent random vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ with

$$\sigma^2 = \sum_{i=1}^{n} \mathbb{E}[\|\boldsymbol{v}_i\|_2^2] \tag{E.81}$$

and ensuring that

$$\|\boldsymbol{v}_i\|_2 \leq R \qquad a.s. \tag{E.82}$$

we obtain that

$$\mathbb{P}\left[\left\|\sum_i \boldsymbol{v}_i - \mathbb{E}\left[\cdot\right]\right\| > t\right] \leq 4k \exp\left(\frac{-t^2/2}{\sigma^2 + 2Rt/3}\right) \tag{E.83}$$

Here, we have used that

$$\left\|\sum_{i=1}^{n} \mathbb{E}[\boldsymbol{v}_i \boldsymbol{v}_i^*]\right\| \leq \operatorname{tr} \sum_{i=1}^{n} \mathbb{E}[\boldsymbol{v}_i \boldsymbol{v}_i^*] \tag{E.84}$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[\|\boldsymbol{v}_i\|_2^2\right]. \tag{E.85}$$

and

$$\boldsymbol{w}_i = \bar{\boldsymbol{x}}_i \left\langle \bar{\boldsymbol{x}}_i, \boldsymbol{\zeta} \right\rangle^3. \tag{E.86}$$

Notice that

$$\|\boldsymbol{w}_i\|_2 \leq \|\bar{\boldsymbol{x}}_i\|_2^4 \tag{E.87}$$

$$\leq \left(4B^2\theta k \log m\right)^2 \tag{E.88}$$

$$= 16B^4\theta^2 k^2 \log m. \tag{E.89}$$

Let us further note that

$$\begin{aligned}
&\sum_{\substack{j_1, \\ j_2 \neq j_3 \neq j_4}} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2 \bar{\boldsymbol{x}}_i(j_2)^2 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i(j_3)^2 \boldsymbol{\zeta}_{j_3}^2 \bar{\boldsymbol{x}}_i(j_4)^2 \boldsymbol{\zeta}_{j_4}^2\right] \\
&= 3 \sum_{j_1 \neq j_2 \neq j_3} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^4 \boldsymbol{\zeta}_{j_1}^2 \bar{\boldsymbol{x}}_i(j_2)^2 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i(j_3)^2 \boldsymbol{\zeta}_{j_3}^2\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{j_1=1}^{2k-1} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2\right] \times \\
&\quad \sum_{j_1 \neq j_2 \neq j_3 \neq j_4} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_2)^2 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i(j_3)^2 \boldsymbol{\zeta}_{j_3}^2 \bar{\boldsymbol{x}}_i(j_4)^2 \boldsymbol{\zeta}_{j_4}^2\right] && \text{(E.90)} \\
&\leq 2\theta k \times \theta^3 \left\|\boldsymbol{\zeta}\right\|_2^6 + 3 \times 3\theta^3 \left\|\boldsymbol{\zeta}\right\|_2^6 && \text{(E.91)}
\end{aligned}
$$

In similar vein, we can obtain that

$$
\begin{aligned}
&\sum_{j_1, j_2 \neq j_3} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2 \bar{\boldsymbol{x}}_i(j_2)^2 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i(j_3)^4 \boldsymbol{\zeta}_{j_3}^4\right] \\
&= \sum_{j_1} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2\right] \sum_{j_2 \neq j_3 \neq j_1} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_2)^2 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i(j_3)^4 \boldsymbol{\zeta}_{j_3}^4\right] \\
&\quad + \sum_{j_1 \neq j_2} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^4 \boldsymbol{\zeta}_{j_1}^2 \bar{\boldsymbol{x}}_i(j_2)^4 \boldsymbol{\zeta}_{j_2}^4\right] \\
&\quad + \sum_{j_1 \neq j_2} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2 \boldsymbol{\zeta}_{j_1}^2 \bar{\boldsymbol{x}}_i(j_2)^6 \boldsymbol{\zeta}_{j_2}^4\right] && \text{(E.92)} \\
&\leq 2\theta k \times 3\theta^2 \left\|\boldsymbol{\zeta}\right\|_2^2 \left\|\boldsymbol{\zeta}\right\|_4^4 + (9+15)\,\theta^2 \left\|\boldsymbol{\zeta}\right\|_2^2 \left\|\boldsymbol{\zeta}\right\|_4^4 && \text{(E.93)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{j_1, j_2} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2 \bar{\boldsymbol{x}}_i(j_2)^6 \boldsymbol{\zeta}_{j_2}^6\right] \\
&= \sum_{j_1} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2\right] \sum_{j_2 \neq j_1} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_2)^6 \boldsymbol{\zeta}_{j_2}^6\right] \\
&\quad + \sum_{j_1} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^8 \boldsymbol{\zeta}_{j_1}^6\right] && \text{(E.94)} \\
&\leq 2\theta k \times 15\theta \left\|\boldsymbol{\zeta}\right\|_6^6 + 105\theta \left\|\boldsymbol{\zeta}\right\|_6^6 && \text{(E.95)}
\end{aligned}
$$

Now we calculate

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{w}_i\right\|_2^2\right] = \mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_i\right\|_2^2 \left\langle \bar{\boldsymbol{x}}_i, \boldsymbol{\zeta}\right\rangle^6\right] && \text{(E.96)} \\
&= \mathbb{E}\left[\sum_{j_1, \ldots, j_7} \bar{\boldsymbol{x}}_i(j_1)^2 \prod_{\ell=2}^{7} \bar{\boldsymbol{x}}_i(j_\ell) \boldsymbol{\zeta}_{j_\ell}\right] && \text{(E.97)} \\
&= 15 \sum_{\substack{j_1, \\ j_2 \neq j_3 \neq j_4}} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2 \bar{\boldsymbol{x}}_i(j_2)^2 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i(j_3)^2 \boldsymbol{\zeta}_{j_3}^2 \bar{\boldsymbol{x}}_i(j_4)^2 \boldsymbol{\zeta}_{j_4}^2\right] \\
&\quad + 15 \sum_{j_1, j_2 \neq j_3} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2 \bar{\boldsymbol{x}}_i(j_2)^2 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i(j_3)^4 \boldsymbol{\zeta}_{j_3}^4\right] \\
&\quad + \sum_{j_1, j_2} \mathbb{E}\left[\bar{\boldsymbol{x}}_i(j_1)^2 \bar{\boldsymbol{x}}_i(j_2)^6 \boldsymbol{\zeta}_{j_2}^6\right] && \text{(E.98)} \\
&\leq 15\theta^3 \left\|\boldsymbol{\zeta}\right\|_2^6 (2\theta k + 9) \\
&\quad + 15\theta^2 \left\|\boldsymbol{\zeta}\right\|_4^4 (6 + 24) \\
&\quad + \theta \left\|\boldsymbol{\zeta}\right\|_6^6 (30\theta k + 105) && \text{(E.99)} \\
&\leq 150\theta^2 k + 600\theta && \text{(E.100)}
\end{aligned}
$$

whence for $\theta > 1/k$,

$$\mathbb{E}\left[\|\boldsymbol{w}_i\|_2^2\right] \;\leq\; C\theta^2 k, \tag{E.101}$$

and hence

$$\sigma^2 \;\leq\; C'\theta^2 m. \tag{E.102}$$

Matrix Bernstein gives that

$$\begin{aligned}&\mathbb{P}\left[\left\|\bar{\boldsymbol{X}}_i(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta})^{\circ 3} - \mathbb{E}\left[\cdot\right]\right\|_2 \geq t\right]\\ &\leq 4k\exp\left(\frac{-t^2/2}{C\theta^2 m + C'B^4\theta^2k^2\log^2 kt}\right).\end{aligned} \tag{E.103}$$

Setting $t = \frac{c}{4}\frac{m\theta(1-\theta)\|\boldsymbol{\zeta}\|_4^6}{\kappa^2(2k-1)}$, we obtain that

$$\begin{aligned}&\mathbb{P}\left[\left\|\frac{1}{m}\bar{\boldsymbol{X}}_i(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta})^{\circ 3} - \mathbb{E}\left[\cdot\right]\right\|_2 \geq \frac{c}{4}\frac{\theta(1-\theta)\|\boldsymbol{\zeta}\|_4^6}{\kappa^2\left(2k-1\right)}\right]\\ &\leq 4k\exp\left(-\frac{c''m\left(1-\theta\right)^2\|\boldsymbol{\zeta}\|_4^{12}}{\kappa^4k^2 + \theta\left(1-\theta\right)B^4\kappa^2k^3\|\boldsymbol{\zeta}\|_4^6}\right)\end{aligned} \tag{E.104}$$

ε-Net Covering To obtain a probability bound for all $\boldsymbol{q} \in \mathbb{S}^{k-1}$, we choose a set of $\boldsymbol{\zeta}_n = \boldsymbol{A}^T\boldsymbol{q}_n$ with $n = 1, \cdots, N$. Suppose for any $\boldsymbol{q} \in \mathbb{S}^{k-1}$, there exists $\boldsymbol{q}_n$ such that $\|\boldsymbol{q} - \boldsymbol{q}_n\|_2 \leq \varepsilon$, then

$$\left\|\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 3} - \frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}_n\right)^{\circ 3}\right\|_2 \leq L\left\|\boldsymbol{q} - \boldsymbol{q}_n\right\|_2. \tag{E.105}$$

For entry wise bounded $\bar{\boldsymbol{X}}_i \in \mathbb{R}^{(2k-1)\times\frac{m}{2k-1}}$, we have

$$\left\|\bar{\boldsymbol{X}}_i\right\|_2 \leq \sqrt{2\theta m}B, \quad \left\|\bar{\boldsymbol{X}}_i\boldsymbol{e}_j\right\|_2 \leq \sqrt{4\theta k}B, \tag{E.106}$$

then the Lipschitz constant L can be bounded as

$$\begin{aligned}L &\leq \frac{1}{m}\left\|\bar{\boldsymbol{X}}_i\right\|_2\left\|\mathrm{diag}\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 2}\right\|_2\left\|\bar{\boldsymbol{X}}_i^T\boldsymbol{A}^T\right\|_2 &\text{(E.107)}\\ &\leq 8\theta^2kB^4. &\text{(E.108)}\end{aligned}$$

With triangle inequality, we have

$$\begin{aligned}&\left\|\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 3} - \mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 3}\right]\right\|_2\\ &\leq \left\|\mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 3}\right] - \mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}_n\right)^{\circ 3}\right]\right\|_2\\ &\quad + \left\|\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}_n\right)^{\circ 3} - \mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}_n\right)^{\circ 3}\right]\right\|_2\\ &\quad + \left\|\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 3} - \frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}_n\right)^{\circ 3}\right\|_2 &\text{(E.109)}\\ &\leq \left\|\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}_n\right)^{\circ 3} - \mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}_n\right)^{\circ 3}\right]\right\|_2\\ &\quad + 2L\varepsilon. &\text{(E.110)}\end{aligned}$$

Hence we need to choose the ε-net to cover the sphere of $\boldsymbol{q}$ with

$$\varepsilon = \frac{c}{4}\frac{\theta\left(1-\theta\right)}{\kappa^2\left(2k-1\right)L}\min_{\boldsymbol{q}\in\mathbb{S}^{k-1}}\|\boldsymbol{\zeta}\|_4^6, \tag{E.111}$$

plug in $L \leq 4\theta^2 k B^4$ and number of sample N suffice

$$N \leq \left(\frac{3}{\varepsilon}\right)^k \tag{E.112}$$

$$\leq \exp\left(k \ln\left(\frac{3}{\varepsilon}\right)\right) \tag{E.113}$$

$$\leq \exp\left[k \ln\left(C \frac{\theta^2 \kappa^2 k^4 B^4}{\theta\left(1-\theta\right)}\right)\right] \tag{E.114}$$

For $n = 1, \cdots, N$, denote

$$P_i\left(\boldsymbol{q}_n\right) = \mathbb{P}\left[\left\| \frac{\bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}_n\right)^{\circ 3}}{m} - \mathbb{E}[\cdot] \right\|_2 \geq \frac{c\theta(1-\theta)\|\boldsymbol{\zeta}_n\|_4^6}{4\kappa^2\left(2k-1\right)}\right], \tag{E.115}$$

then together with union bound over all $\boldsymbol{q}_n$, we obtain that,

$$\mathbb{P}\left[\sup_{\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}} \frac{\left\| \frac{1}{m} \bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 3} - \mathbb{E}\left[\cdot\right] \right\|_2}{\|\boldsymbol{\zeta}\|_4^6} \geq \frac{c}{2} \frac{\theta\left(1-\theta\right)}{\kappa^2\left(2k-1\right)}\right]$$

$$\leq \sum_{\boldsymbol{q}_n \in \hat{\mathcal{R}}_{2C_\star}} P_i\left(\boldsymbol{q}_n\right) \tag{E.116}$$

$$\leq N \max_{\boldsymbol{q}_n \in \hat{\mathcal{R}}_{2C_\star}} P_i\left(\boldsymbol{q}_n\right) \tag{E.117}$$

$$\leq 4k \sup_{\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}} \exp\left(-\frac{cm\left(1-\theta\right)^2 \|\boldsymbol{\zeta}\|_4^{12}}{\kappa^4 k^2 + \theta\left(1-\theta\right) B^4 \kappa^2 k^3 \|\boldsymbol{\zeta}\|_4^6}\right) \times$$

$$\exp\left(k \ln\left(\frac{3}{\varepsilon}\right)\right). \tag{E.118}$$

Hence

$$\mathbb{P}\left[\sup_{\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}} \frac{\left\| \frac{1}{m} \bar{\boldsymbol{X}}_0 \left(\bar{\boldsymbol{X}}_0^T \boldsymbol{\zeta}\right)^{\circ 3} - \mathbb{E}\left[\cdot\right] \right\|_2}{\|\boldsymbol{\zeta}\|_4^6} \geq \frac{c}{2} \frac{\theta\left(1-\theta\right)}{\kappa^2}\right]$$

$$\leq \sum_i \mathbb{P}\left[\sup_{\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}} \frac{\left\| \frac{1}{m} \bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 3} - \mathbb{E}\left[\cdot\right] \right\|_2}{\|\boldsymbol{\zeta}\|_4^6} \geq \frac{c\theta\left(1-\theta\right)}{2\kappa^2\left(2k-1\right)}\right] \tag{E.119}$$

$$\leq \left(2k-1\right) \max_i$$

$$\mathbb{P}\left[\sup_{\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}} \frac{\left\| \frac{1}{m} \bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 3} - \mathbb{E}\left[\cdot\right] \right\|_2}{\|\boldsymbol{\zeta}\|_4^6} \geq \frac{c\theta\left(1-\theta\right)}{2\kappa^2\left(2k-1\right)}\right] \tag{E.120}$$

$$\leq 8k^2 \sup_{\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}} \exp\left(-\frac{cm\left(1-\theta\right)^2 \|\boldsymbol{\zeta}\|_4^{12}}{\kappa^4 k^2 + \theta\left(1-\theta\right) B^4 \kappa^2 k^3 \|\boldsymbol{\zeta}\|_4^6}\right) \times$$

$$\exp\left(k \ln\left(\frac{3}{\varepsilon}\right)\right), \tag{E.121}$$

which is bounded by $\exp\left(-k\right)$ as long as

$$m \geq C \frac{\min\left\{\left(2C_\star \mu\right)^{-2}, \kappa^2 k^2\right\}}{\left(1-\theta\right)^2} \kappa^2 k^4 \log^3\left(\kappa k\right) \tag{E.122}$$

$$\geq C' k \log\left(\frac{\theta\kappa^2 k^2 B^4}{(1-\theta)\left\|\boldsymbol{\zeta}\right\|_4^6}\right) \times \max\left\{\frac{\kappa^4 k^2}{(1-\theta)^2\left\|\boldsymbol{\zeta}\right\|_4^{12}}, \frac{\theta B^4 \kappa^2 k^3}{(1-\theta)\left\|\boldsymbol{\zeta}\right\|_4^6}\right\}. \tag{E.123}$$

To sum up, we obtain that for all $\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}$, inequality

$$\left\|\frac{1}{m}\boldsymbol{X}_0\left(\boldsymbol{X}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 3} - \mathbb{E}\left[\cdot\right]\right\|_2 \leq c\theta(1-\theta)\frac{\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^6}{\kappa^2} \tag{E.124}$$

holds with probability no smaller than $1 - c_1\exp(-k) - c_2 k^{-4} - c_3\exp(-\theta k)$. ∎

F Concentration for Hessian (Lemma 4.3)

Lemma F.1. *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta)$. There exists positive constant C that whenever*

$$m \geq C\theta\frac{\min\left\{\left(2C_\star\mu\kappa^2\right)^{-4/3}, k^2\right\}}{(1-\theta)^2\sigma_{\min}^2}\kappa^6 k^4\log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right) \tag{F.1}$$

and $\theta \geq \log k/k$, then with probability no smaller than $1 - c_1\exp(-k) - c_2 k^{-4} - 48k^{-7} - 48m^{-5} - 24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right)$,

$$\left\|\text{Hess}[\psi](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\text{Hess}[\varphi](\boldsymbol{q})\right\|_2 \leq c\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4, \tag{F.2}$$

holds for all $\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}$ with positive constant $c \leq 0.048 \leq 3\left(1 - 6c_\star - 36c_\star^2 - 24c_\star^3\right)$.

Proof Denote $\boldsymbol{\eta} = \boldsymbol{Y}^T\left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{q}$ and $\bar{\boldsymbol{\eta}} = \boldsymbol{Y}^T\left(\theta m\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{q} = (\theta m)^{-1/2}\boldsymbol{X}_0^T\boldsymbol{\zeta}$, and

$$\boldsymbol{W} = \left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} - \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}, \tag{F.3}$$

$$\widehat{\boldsymbol{Y}} = \left(\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{Y}. \tag{F.4}$$

Then we have

$$\begin{aligned}
&\left\|\text{Hess}\left[\psi\right](\boldsymbol{q}) - \frac{3(1-\theta)}{\theta m^2}\text{Hess}\left[\varphi\right](\boldsymbol{q})\right\|_2 \\
&= \left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\frac{3}{m}\widehat{\boldsymbol{Y}}\,\text{diag}\left(\boldsymbol{\eta}^{\circ 2}\right)\widehat{\boldsymbol{Y}}^T - \langle\boldsymbol{q}, \nabla\psi(\boldsymbol{q})\rangle\boldsymbol{I}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}\right. \\
&\left. - \frac{3(1-\theta)}{\theta m^2}\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[3\boldsymbol{A}\,\text{diag}\left(\boldsymbol{\zeta}^{\circ 2}\right)\boldsymbol{A}^T - \left\|\boldsymbol{\zeta}\right\|_4^4\boldsymbol{I}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}\right\|_2 \qquad \text{(F.5)} \\
&\leq \left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\frac{3}{m}\widehat{\boldsymbol{Y}}\,\text{diag}\left(\boldsymbol{\eta}^{\circ 2}\right)\widehat{\boldsymbol{Y}}^T\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}\right. \\
&\left. - \boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\frac{9(1-\theta)}{\theta m^2}\boldsymbol{A}\,\text{diag}\left(\boldsymbol{\zeta}^{\circ 2}\right)\boldsymbol{A}^T - \frac{3}{m^2}\boldsymbol{I}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}\right\|_2 \\
&+ \left\|\left[\langle\boldsymbol{q}, \nabla\psi(\boldsymbol{q})\rangle - \frac{3(1-\theta)}{\theta m^2}\left\|\boldsymbol{\zeta}\right\|_4^4 - \frac{3}{m^2}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}\right\|_2 \qquad \text{(F.6)}
\end{aligned}$$

$$
\begin{aligned}
&\leq \underbrace{\frac{3}{\theta m^2}\left\|\boldsymbol{W}\boldsymbol{Y}\operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2}\right)\boldsymbol{Y}^T\left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\right\|_2}_{\Delta_1^H} \\
&+ \underbrace{\frac{3}{\theta m^2}\left\|\boldsymbol{A}\boldsymbol{X}_0\operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2}\right)\boldsymbol{Y}^T\boldsymbol{W}\right\|_2}_{\Delta_2^H} \\
&+ \underbrace{\frac{3}{\theta m^2}\left\|\boldsymbol{A}\boldsymbol{X}_0\operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2}-\bar{\boldsymbol{\eta}}^{\circ 2}\right)\boldsymbol{X}_0^T\boldsymbol{A}^T\right\|_2}_{\Delta_3^H} \\
&+ \underbrace{\frac{3}{\theta m^2}\left\|\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[\boldsymbol{A}\boldsymbol{X}_0\operatorname{diag}\left(\bar{\boldsymbol{\eta}}^{\circ 2}\right)\boldsymbol{X}_0^T\boldsymbol{A}^T\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}\right.}_{} \\
&\quad \underbrace{\left.-\boldsymbol{P}_{\boldsymbol{q}^\perp}\left[3\left(1-\theta\right)\boldsymbol{A}\operatorname{diag}\left(\boldsymbol{\zeta}^{\circ 2}\right)\boldsymbol{A}^T+\theta\boldsymbol{I}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}\right\|_2}_{\Delta_4^H} \\
&+ \underbrace{\left\|\left[\langle\boldsymbol{q},\nabla\psi\left(\boldsymbol{q}\right)\rangle-\frac{3\left(1-\theta\right)}{\theta m^2}\left\|\boldsymbol{\zeta}\right\|_4^4-\frac{3}{m^2}\right]\boldsymbol{P}_{\boldsymbol{q}^\perp}\right\|_2}_{\Delta_5^H}
\end{aligned}
\tag{F.7}
$$

In the rest of the proof, we prove that

$$
\Delta_i^H \leq \frac{c}{9}\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{\zeta}\right\|_4^4, \quad i=1,2,3. \tag{F.8}
$$

and

$$
\Delta_i^H \leq \frac{c}{3}\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{\zeta}\right\|_4^4, \quad i=4,5. \tag{F.9}
$$

First, let us note that

$$
C\left(1-\theta\right)^{-2}\sigma_{\min}^{-2}\kappa^6 k^5\log^3\left(\frac{\kappa k}{\left(1-\theta\right)\sigma_{\min}}\right) \tag{F.10}
$$

$$
\leq C\left(\frac{\kappa k}{\left(1-\theta\right)\sigma_{\min}}\right)^6\log^3\left(\frac{\kappa k}{\left(1-\theta\right)\sigma_{\min}}\right) \tag{F.11}
$$

$$
\leq C\left(\frac{\kappa k}{\left(1-\theta\right)\sigma_{\min}}\right)^9 \tag{F.12}
$$

or

$$
\frac{\log^3\left(C\left(1-\theta\right)^{-2}\sigma_{\min}^{-2}\kappa^6 k^5\log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right)\right)}{C\log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right)}
$$

$$
\leq \left(\frac{\log C+9\log\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right)}{C^{1/3}\log\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right)}\right)^3 \tag{F.13}
$$

$$
\leq \left(\frac{\log C}{C^{1/3}\log\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}}\right)}+\frac{9}{C^{1/3}}\right)^3 \tag{F.14}
$$

$$\leq \left(\frac{1}{C^{1/6}} + \frac{1}{2}\frac{1}{C^{1/6}}\right)^3 \qquad \left(C \geq 10^8\right) \tag{F.15}$$

$$\leq \frac{4}{C^{1/2}}. \tag{F.16}$$

Since

$$m \geq C\frac{\min\left\{\left(2C_\star\mu\kappa^2\right)^{-4/3}, k^2\right\}}{\left(1-\theta\right)^2\sigma_{\min}^2}\kappa^6 k^4 \log^3\left(\frac{\kappa k}{\left(1-\theta\right)\sigma_{\min}}\right), \tag{F.17}$$

as the ratio $\log^3 m/m$ decreases with increasing m, then

$$\begin{aligned}
&\frac{\log^3 m}{m} \\
&\leq \frac{\log^3\left(C\frac{\kappa^6k^5}{(1-\theta)^2\sigma_{\min}^2}\log^3\left(\frac{\kappa k}{\sigma_{\min}(1-\theta)}\right)\right)}{C\log^3\left(\frac{\kappa k}{\sigma_{\min}(1-\theta)}\right)} \\
&\quad\times \frac{\left(1-\theta\right)^2\sigma_{\min}^2}{\min\left\{\left(2C_\star\mu\kappa^2\right)^{-2/3}, k\right\}\kappa^6k^4} \qquad \text{(F.18)} \\
&\leq \frac{4}{C^{1/2}}\frac{\left(1-\theta\right)^2\sigma_{\min}^2}{\min\left\{\left(2C_\star\mu\kappa^2\right)^{-2/3}, k\right\}\kappa^6k^4} \qquad \text{(F.19)}
\end{aligned}$$

According to Lemma D.1, following inequality obtains

$$\left\|\frac{1}{\theta m}\boldsymbol{X}_0\boldsymbol{X}_0^T - \boldsymbol{I}\right\|_2 \leq \delta \tag{F.20}$$

$$\leq 10\sqrt{k\log m/m} \tag{F.21}$$

$$\leq \frac{20\left(1-\theta\right)\sigma_{\min}\max\left\{\left(2C_\star\mu\kappa^2\right)^{2/3}, k^{-1}\right\}}{C^{1/4}\kappa^3k^{3/2}\log m} \tag{F.22}$$

$$\leq \frac{20\sigma_{\min}}{C^{1/4}\kappa^3}\cdot\frac{\left(1-\theta\right)\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4}{k^{3/2}\log m}, \quad \forall \boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star} \tag{F.23}$$

with probability no smaller than $1-\varepsilon_0$ with $\varepsilon_0 = 2\exp\left(-\theta k\right) + 24k\exp\left(-\frac{1}{144}\min\left\{k, 3\sqrt{\theta m}\right\}\right) + 48k^{-7} + 48m^{-5}$.

We have $4\kappa^3\delta/\sigma_{\min} \leq 1/2$ whenever

$$C \geq \left(\frac{160\left(1-\theta\right)}{k^{3/2}\log m}\right)^4 \tag{F.24}$$

whence $\delta \leq 1/\left(8\kappa^2\right)$, and Lemma D.3 implies that

$$\begin{aligned}
&\left\|\left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2}\boldsymbol{A}_0 - \left(\boldsymbol{A}_0\boldsymbol{A}_0^T\right)^{-1/2}\boldsymbol{A}_0\right\|_2 \\
&\leq 4\kappa^3\delta/\sigma_{\min} \qquad \text{(F.25)} \\
&\leq \frac{80\left(1-\theta\right)\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4}{C^{1/4}k^{3/2}\log m}, \quad \forall \boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}. \qquad \text{(F.26)}
\end{aligned}$$

Moreover,

$$\|\boldsymbol{X}_0\|_2 \leq (\theta m)^{1/2} \sqrt{1+\delta} \tag{F.27}$$

$$\leq (\theta m)^{1/2} (1+\delta/2) \tag{F.28}$$

$$\leq \frac{17}{16} (\theta m)^{1/2}. \tag{F.29}$$

Finally, Lemma A.5 implies that with probability no smaller than $1-\varepsilon_B$, we have

$$\|\boldsymbol{x}_0\|_\infty \leq \sqrt{2} \log^{1/2}\left(\frac{2\theta m}{\varepsilon_B}\right). \tag{F.30}$$

Upper Bound for Δ_1^H and Δ_2^H. With probability no smaller than $1-\varepsilon_0-\varepsilon_B$, the norms of $\boldsymbol{\eta}$ are upper bounded as in Lemma A.7,

$$\Delta_1^H \leq \frac{3}{\theta m^2} \left\| \left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} \boldsymbol{A}_0 - \boldsymbol{A} \right\|_2 \times \|\boldsymbol{X}_0\|_2^2 \|\boldsymbol{\eta}\|_\infty^2 \left\| \boldsymbol{A}_0^T \left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} \right\|_2 \tag{F.31}$$

$$\leq \frac{3}{\theta m^2} \cdot \frac{4\kappa^3\delta}{\sigma_{\min}} \cdot (1+\delta/2)^2 \theta m \cdot \left(1+\frac{4\kappa^3\delta}{\sigma_{\min}}\right)^3 \frac{4k}{\theta m} \log\left(2\theta m/\varepsilon_B\right) \tag{F.32}$$

$$\leq \frac{3660}{C^{1/4}} \frac{1-\theta}{\theta m^2} \|\boldsymbol{\zeta}\|_4^4 \cdot \frac{\log\left(2\theta m/\varepsilon_B\right)}{k^{1/2}\log m}. \tag{F.33}$$

A similar result holds for

$$\Delta_2^H \leq \frac{3}{\theta m^2} \|\boldsymbol{X}_0\|_2^2 \left\|\mathrm{diag}\left(\boldsymbol{\eta}^{\circ 2}\right)\right\|_2 \times \left\| \left(\frac{1}{\theta m}\boldsymbol{Y}\boldsymbol{Y}^T\right)^{-1/2} \boldsymbol{A}_0 - \boldsymbol{A} \right\|_2 \tag{F.34}$$

$$\leq \frac{2440}{C^{1/4}} \frac{1-\theta}{\theta m^2} \|\boldsymbol{\zeta}\|_4^4 \cdot \frac{\log\left(2\theta m/\varepsilon_B\right)}{k^{1/2}\log m}. \tag{F.35}$$

To make $\Delta_1^H \leq \frac{c}{9}\frac{1-\theta}{\theta m^2}\|\boldsymbol{\zeta}\|_4^4$ and $\Delta_2^H \leq \frac{c}{9}\frac{1-\theta}{\theta m^2}\|\boldsymbol{\zeta}\|_4^4$, we require

$$C \geq \left(9 \times 3660 c^{-1} \frac{\log\left(2\theta m/\varepsilon_B\right)}{k^{1/2}\log m}\right)^4. \tag{F.36}$$

The right hand side is bounded by an absolute constant for all m.

Upper Bound for Δ_3^H. With probability no smaller than $1-\varepsilon_0-\varepsilon_B$, the difference between $\bar{\boldsymbol{\eta}}^{\circ 2}$ and $\boldsymbol{\eta}^{\circ 2}$ is upper bounded as in Lemma A.7,

$$\left\|\boldsymbol{\eta}^{\circ 2} - \bar{\boldsymbol{\eta}}^{\circ 2}\right\|_\infty \leq \|\boldsymbol{\eta}-\bar{\boldsymbol{\eta}}\|_\infty \|\boldsymbol{\eta}+\bar{\boldsymbol{\eta}}\|_\infty \tag{F.37}$$

$$\leq \frac{4\kappa^3\delta}{\sigma_{\min}} \left(2+\frac{4\kappa^3\delta}{\sigma_{\min}}\right) \frac{2k}{\theta m} \log\left(2\theta m/\varepsilon_B\right) \tag{F.38}$$

$$\leq \frac{5k}{\theta m} \log\left(2\theta m/\varepsilon_B\right) \cdot \frac{4\kappa^3 \delta}{\sigma_{\min}}. \tag{F.39}$$

Therefore

$$\Delta_3^H = \frac{3}{\theta m^2} \left\| \boldsymbol{A} \boldsymbol{X}_0 \operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2} - \bar{\boldsymbol{\eta}}^{\circ 2}\right) \boldsymbol{X}_0^T \boldsymbol{A}^T \right\|_2 \tag{F.40}$$

$$\leq \frac{}{\theta m^2} \left\|\boldsymbol{A}\right\|_2^2 \left\|\boldsymbol{X}_0\right\|_2^2 \left\|\operatorname{diag}\left(\boldsymbol{\eta}^{\circ 2} - \bar{\boldsymbol{\eta}}^{\circ 2}\right)\right\|_2 \tag{F.41}$$

$$\leq \frac{15k}{\theta m^2} \left(1+\delta/2\right)^2 \log\left(2\theta m/\varepsilon_B\right) \cdot \frac{4\kappa^3\delta}{\sigma_{\min}} \tag{F.42}$$

$$\leq \frac{1400\left(1-\theta\right)\log\left(2\theta m/\varepsilon_B\right)}{C^{1/4}\theta k^{1/2} m^2 \log m} \left\|\boldsymbol{\zeta}\right\|_4^4. \tag{F.43}$$

Again, Δ_3^H is bounded by $\frac{c}{9}\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{\zeta}\right\|_4^4$ whenever

$$C \geq \left(9 \times 1400 c^{-1} \frac{\log\left(2\theta m/\varepsilon_B\right)}{k^{1/2}\log m}\right)^4 \tag{F.44}$$

Upper Bound for Δ_4^H. Recall that

$$\bar{\boldsymbol{\eta}} = \boldsymbol{Y}^T \left(\theta m \boldsymbol{A}_0 \boldsymbol{A}_0^T\right)^{-1/2} \boldsymbol{q}, \tag{F.45}$$

then

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{X}_0 \operatorname{diag}\left(\bar{\boldsymbol{\eta}}^{\circ 2}\right) \boldsymbol{X}_0^T\right] \\
&= \mathbb{E}\left[\frac{1}{\theta m} \boldsymbol{X}_0 \operatorname{diag}\left(\boldsymbol{X}_0^T \boldsymbol{A}^T \boldsymbol{q}\right)^{\circ 2} \boldsymbol{X}_0^T\right] &\text{(F.46)}\\
&= 3\left(1-\theta\right) \operatorname{diag}\left(\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 2} + 2\theta \boldsymbol{A}^T\boldsymbol{q}\boldsymbol{q}^T\boldsymbol{A} \\
&\quad + \theta \left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_2^2 \boldsymbol{I}, &\text{(F.47)}
\end{aligned}$$

once including the projection $\boldsymbol{P}_{\boldsymbol{q}^\perp}$, we have

$$\begin{aligned}
&\boldsymbol{P}_{\boldsymbol{q}^\perp} \mathbb{E}\left[\boldsymbol{A}\boldsymbol{X}_0 \operatorname{diag}\left(\bar{\boldsymbol{\eta}}^{\circ 2}\right) \boldsymbol{X}_0^T \boldsymbol{A}^T\right] \boldsymbol{P}_{\boldsymbol{q}^\perp} &\text{(F.48)}\\
&= \boldsymbol{P}_{\boldsymbol{q}^\perp} \left[3\left(1-\theta\right) \boldsymbol{A} \operatorname{diag}\left(\boldsymbol{\zeta}^{\circ 2}\right) \boldsymbol{A}^T + \theta \boldsymbol{I}\right] \boldsymbol{P}_{\boldsymbol{q}^\perp}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\Delta_4^H = \frac{3}{\theta m^2} \Big\| &\boldsymbol{P}_{\boldsymbol{q}^\perp} \left[\boldsymbol{A}\boldsymbol{X}_0 \operatorname{diag}\left(\bar{\boldsymbol{\eta}}^{\circ 2}\right) \boldsymbol{X}_0^T \boldsymbol{A}^T\right] \boldsymbol{P}_{\boldsymbol{q}^\perp} \\
&- \boldsymbol{P}_{\boldsymbol{q}^\perp} \left[3\left(1-\theta\right) \boldsymbol{A} \operatorname{diag}\left(\boldsymbol{\zeta}^{\circ 2}\right) \boldsymbol{A}^T + \theta \boldsymbol{I}\right] \boldsymbol{P}_{\boldsymbol{q}^\perp} \Big\|_2 &\text{(F.49)}\\
\leq \frac{3}{\theta^2 m^2} \left\| \frac{1}{m} \boldsymbol{X}_0 \operatorname{diag}\left(\boldsymbol{X}_0^T\boldsymbol{\zeta}\right)^{\circ 2} \boldsymbol{X}_0^T - \mathbb{E}\left[\cdot\right] \right\|_2 &\text{(F.50)}
\end{aligned}$$

Under the assumption for sample size that $m \geq C\left(1-\theta\right)^{-2} \kappa^4 \min\left\{\left(2C_\star\mu\right)^{-2/3}, k\right\} k^3 \log^5\left(\kappa k\right)$, applying Lemma F.2, we have

$$\left\| \frac{1}{m} \boldsymbol{X}_0 \operatorname{diag}\left(\boldsymbol{X}_0^T\boldsymbol{\zeta}\right)^{\circ 2} \boldsymbol{X}_0^T - \mathbb{E}\left[\cdot\right] \right\|_2 \leq \frac{c}{9}\theta\left(1-\theta\right)\left\|\boldsymbol{\zeta}\right\|_4^4. \tag{F.51}$$

simultaneously at every $\boldsymbol{q} \in \hat{\mathcal{R}}_{2C_\star}$ with probability no smaller than $1 - c_1\exp\left(-k\right) - c_2 k^{-4}$.

Upper Bound for Δ_5^H. Note that this term is essentially the difference between

$$\Delta_5^H = \left\| \left[\langle \boldsymbol{q}, \nabla\psi(\boldsymbol{q}) \rangle - \frac{3(1-\theta)}{\theta m^2} \|\boldsymbol{\zeta}\|_4^4 - \frac{3}{m^2} \right] \boldsymbol{P}_{\boldsymbol{q}^\perp} \right\|_2 \tag{F.52}$$

$$\leq \left| \langle \boldsymbol{q}, \nabla\psi(\boldsymbol{q}) \rangle - \frac{3(1-\theta)}{\theta m^2} \|\boldsymbol{\zeta}\|_4^4 - \frac{3}{m^2} \right| \tag{F.53}$$

$$\leq \frac{1}{\theta^2 m^2} \left| \frac{1}{m} \left\| \boldsymbol{X}_0^T \boldsymbol{\zeta} \right\|_4^4 - 3\theta(1-\theta) \|\boldsymbol{\zeta}\|_4^4 - 3\theta^2 \right|$$

$$+ \left| \langle \boldsymbol{q}, \nabla\psi(\boldsymbol{q}) \rangle - \frac{1}{\theta^2 m^2} \left\| \boldsymbol{X}_0^T \boldsymbol{\zeta} \right\|_4^4 \right| \tag{F.54}$$

$$\leq \frac{1}{\theta^2 m^2} \left\| \frac{\boldsymbol{A}\boldsymbol{X}_0 \left(\boldsymbol{X}_0^T \boldsymbol{\zeta}\right)^{\circ 3}}{m} - 3\theta(1-\theta)\boldsymbol{A}^T \boldsymbol{\zeta}^{\circ 3} - 3\theta^2 \boldsymbol{q} \right\|_2$$

$$+ \frac{1}{m} \left| \|\boldsymbol{\eta}\|_4^4 - \|\bar{\boldsymbol{\eta}}\|_4^4 \right| \tag{F.55}$$

Recall that

$$\mathbb{E}\left[\frac{1}{m} \boldsymbol{A}\boldsymbol{X}_0 \left(\boldsymbol{X}_0^T \boldsymbol{A}^T \boldsymbol{q} \right)^{\circ 3} \right]$$

$$= \mathbb{E}\left[\boldsymbol{A}\boldsymbol{x}_i \left(\boldsymbol{x}_i^T \boldsymbol{A}^T \boldsymbol{q} \right)^3 \right] \tag{F.56}$$

$$= 3\theta(1-\theta)\boldsymbol{A}\boldsymbol{\zeta}^{\circ 3} + 3\theta^2 \boldsymbol{q}, \tag{F.57}$$

With similar argument as in Lemma 4.2, we can show that this term can be bounded by $\frac{c}{6}\frac{1-\theta}{\theta m^2}\|\boldsymbol{\eta}\|_4^4$ whenever

$$m \geq C' \frac{\min\left\{ \left(\mu\kappa^2\right)^{-4/3}, k^2 \right\}}{(1-\theta)^2 \sigma_{\min}^2} \kappa^6 k^4 \log^3\left(\frac{\kappa k}{(1-\theta)\sigma_{\min}} \right). \tag{F.58}$$

Moreover, with probability $1 - \varepsilon_0 - \varepsilon_B$

$$\frac{1}{m} \left| \|\boldsymbol{\eta}\|_4^4 - \|\bar{\boldsymbol{\eta}}\|_4^4 \right|$$

$$\leq \frac{1}{m} \left| \langle \boldsymbol{\eta} - \bar{\boldsymbol{\eta}}, 4\boldsymbol{\eta}^{\circ 3} \rangle \right| \tag{F.59}$$

$$\leq \frac{4}{m} \|\boldsymbol{\eta} - \bar{\boldsymbol{\eta}}\|_2 \|\boldsymbol{\eta}\|_6^3 \tag{F.60}$$

$$\leq \frac{16\kappa^3\delta}{\sigma_{\min} m} (1 + \delta/2) \left(1 + \frac{4\kappa^3\delta}{\sigma_{\min}} \right)^2 \frac{4k}{\theta m} \log\left(2\theta m/\varepsilon_B\right) \tag{F.61}$$

$$\leq \frac{153k}{\theta m^2} \log\left(2\theta m/\varepsilon_B\right) \cdot \frac{\kappa^3\delta}{\sigma_{\min}} \tag{F.62}$$

$$\leq \frac{3060}{C^{1/4}} \frac{(1-\theta)}{\theta m^2} \|\boldsymbol{\zeta}\|_4^4 \cdot \frac{\log\left(2\theta m/\varepsilon_B\right)}{k^{1/2}\log m}, \tag{F.63}$$

which is bounded by $\frac{c}{6}\frac{1-\theta}{\theta m^2}\|\boldsymbol{\zeta}\|_4^4$ whenever

$$C \geq \left(6 \times 3060 c^{-1} \frac{(1-\theta)\log\left(2\theta m/\varepsilon_B\right)}{k^{1/2}\log m} \right)^4. \tag{F.64}$$

The right hand side is bounded by an absolute constant for all m.

Adding up failure probabilities, we have that with probability larger than $1-c_2\exp(-k)-c_2k^{-4}-\varepsilon_0$,

$$\left\|\mathrm{Hess}\left[\psi\right](\boldsymbol{q})-\frac{3(1-\theta)}{\theta m^2}\mathrm{Hess}\left[\varphi\right](\boldsymbol{q})\right\|_2\le c\frac{1-\theta}{\theta m^2}\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4 \tag{F.65}$$

holds as desired for all $\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}$, where $\varepsilon_0=2\exp(-\theta k)+24k\exp\left(-\frac{1}{144}\min\left\{k,3\sqrt{\theta m}\right\}\right)+48k^{-7}+48m^{-5}$. ■

F.1 Proof of Lemma F.2

Lemma F.2. *Suppose $\boldsymbol{x}_0\sim_{\text{i.i.d.}}\mathrm{BG}(\theta)$. There exist constants $C>0$ that whenever*

$$m\ge C\frac{\min\left\{\left(2C_\star\mu\kappa^2\right)^{-4/3},k^2\right\}}{(1-\theta)^2}k^4\log^3(\kappa k), \tag{F.66}$$

and $\theta k>1$, then with probability no smaller than $1-c_1\exp(-k)-c_2k^{-4}$,

$$\begin{aligned}&\left\|\frac{1}{m}\boldsymbol{X}_0\,\mathrm{diag}\left(\boldsymbol{X}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 2}\boldsymbol{X}_0^T-\mathbb{E}\left[\cdot\right]\right\|_2\\&\quad\le c\theta(1-\theta)\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4,\end{aligned} \tag{F.67}$$

holds for all $\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}$ with positive constant $c\le 0.005\le\left(1-6c_\star-36c_\star^2-24c_\star^3\right)/3$.

Proof The proof strategy for the finite sample concentration of the Hessian is similar to that of the gradient as presented in Lemma E.2. For simplicity, we will only demonstrate some key steps here, please refer to Lemma E.2 for detailed arguments.

Again, from Lemma A.5, the coefficient satisfies $\left\|\boldsymbol{x}_0\right\|_\infty\le B$ with probability no smaller than $1-2\theta m e^{-B^2/2}$. We write $\bar{\boldsymbol{x}}_0(i)=\boldsymbol{x}_0(i)\mathbb{1}_{|\boldsymbol{x}_0(i)|\le B}$, and let $\bar{\boldsymbol{X}}_0$ denote the circulant matrix generated by the truncated vector $\bar{\boldsymbol{x}}_0$. Denote

$$\boldsymbol{H}_E=\mathbb{E}\left[\frac{1}{m}\boldsymbol{X}_0\,\mathrm{diag}\left(\boldsymbol{X}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 2}\boldsymbol{X}_0^T\right], \tag{F.68}$$

$$\bar{\boldsymbol{H}}_E=\mathbb{E}\left[\frac{1}{m}\bar{\boldsymbol{X}}_0\,\mathrm{diag}\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 2}\bar{\boldsymbol{X}}_0^T\right], \tag{F.69}$$

then

$$\begin{aligned}&\mathbb{P}\left[\left\|\frac{\boldsymbol{X}_0\,\mathrm{diag}\left(\boldsymbol{X}_0^T\boldsymbol{\zeta}\right)^{\circ 2}\boldsymbol{X}_0^T}{m}-\boldsymbol{H}_E\right\|_2\ge c\theta(1-\theta)\|\boldsymbol{\zeta}\|_4^4\right]\\&\le\mathbb{P}\left[\left\|\frac{\bar{\boldsymbol{X}}_0\,\mathrm{diag}\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 2}\bar{\boldsymbol{X}}_0^T}{m}-\boldsymbol{H}_E\right\|_2\ge c\theta(1-\theta)\|\boldsymbol{\zeta}\|_4^4\right]\\&\quad+2\theta m e^{-B^2/2}+2m\exp\left(-\frac{3}{4}\theta k\log m\right)\end{aligned} \tag{F.70}$$

while via triangle inequality,

$$\begin{aligned}&\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0\,\mathrm{diag}\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 2}\bar{\boldsymbol{X}}_0^T-\boldsymbol{H}_E\right\|_2\\&\le\left\|\frac{1}{m}\bar{\boldsymbol{X}}_0\,\mathrm{diag}\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 2}\bar{\boldsymbol{X}}_0^T-\bar{\boldsymbol{H}}_E\right\|_2\end{aligned}$$

$$+ \left\| \bar{\boldsymbol{H}}_E - \boldsymbol{H}_E \right\|_2 . \tag{F.71}$$

Truncation Level. Next, we choose a large enough entry-wise truncation level B such that the expectation of the Hessian $\mathbb{E}\left[\boldsymbol{X}_0 \operatorname{diag}\left(\boldsymbol{X}_0^T \boldsymbol{A}^T \boldsymbol{q}\right)^{\circ 2} \boldsymbol{X}_0^T\right]$ is close to that of its truncation $\mathbb{E}\left[\bar{\boldsymbol{X}}_0 \operatorname{diag}\left(\bar{\boldsymbol{X}}_0^T \boldsymbol{A}^T \boldsymbol{q}\right)^{\circ 2} \bar{\boldsymbol{X}}_0^T\right]$. Moreover, we introduce following events notation

$$\mathcal{E}_i \doteq \left\{ \|\boldsymbol{x}_i\|_\infty > B \ \cup \ \|\boldsymbol{x}_i\|_0 > 4\theta k \log m \right\}, \tag{F.72}$$

then

$$\begin{aligned}
&\left\| \bar{\boldsymbol{H}}_E - \boldsymbol{H}_E \right\|_2 \\
&= \left\| \mathbb{E}\left[\frac{1}{m} \sum_i \left\langle \boldsymbol{x}_i, \boldsymbol{\zeta} \right\rangle^2 \boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \mathbf{1}_{\boldsymbol{E}_i} \right] \right\|_F && \text{(F.73)} \\
&\le \frac{1}{m} \sum_i \left\| \mathbb{E}\left[\left\langle \boldsymbol{x}_i, \boldsymbol{\zeta} \right\rangle^2 \boldsymbol{x}_i \boldsymbol{x}_i^T \cdot \mathbf{1}_{\boldsymbol{E}_i} \right] \right\|_F && \text{(F.74)} \\
&\le \frac{1}{m} \sum_i \left(\mathbb{E}\left[\left\| \left\langle \boldsymbol{x}_i, \boldsymbol{\zeta} \right\rangle^2 \boldsymbol{x}_i \boldsymbol{x}_i^T \right\|_F^2 \right] \cdot \mathbb{E}\left[\mathbf{1}_{\boldsymbol{E}_i} \right] \right)^{1/2} && \text{(F.75)} \\
&\le \left(\mathbb{E}\left[\|\boldsymbol{x}_i\|_2^8 \right] \right)^{1/2} \times \\
&\quad \sqrt{\mathbb{E}\left[\mathbf{1}_{\|\boldsymbol{x}_i\|_\infty > B} \right] + \mathbb{E}\left[\mathbf{1}_{\|\boldsymbol{x}_i\|_0 > 4\theta k \log m} \right]} && \text{(F.76)} \\
&\le 50 k^2 \sqrt{4\theta k e^{-B^2/2} + \exp\left(-\tfrac{3}{4} \theta k \log m \right)} && \text{(F.77)}
\end{aligned}$$

By setting

$$B \ge C' \log^{1/2}\left(\frac{k^7}{\theta \left(1-\theta\right)^2} \right) \tag{F.78}$$

we have

$$\theta k e^{-B^2/2} \le c' \theta^2 \left(1-\theta\right)^2 \frac{\|\boldsymbol{\zeta}\|_4^8}{k^4} \tag{F.79}$$

In addition, whenever

$$\theta k \ge \frac{4}{3 \log m} \log\left(\frac{400^2 k^4}{c^2 \theta^2 \left(1-\theta\right)^2 \|\boldsymbol{\zeta}\|_4^8} \right), \tag{F.80}$$

we have

$$\exp\left(-\tfrac{3}{4} \theta k \log m \right) \le \frac{1}{2} \left(\frac{c\theta(1-\theta)}{100} \right)^2 \frac{\|\boldsymbol{\zeta}\|_4^8}{k^4}. \tag{F.81}$$

Hence,

$$\sqrt{4\theta k e^{-B^2/2} + \exp\left(-\tfrac{3}{4} \theta k \log m \right)} \le \frac{c\theta \left(1-\theta\right)}{100 k^2} \|\boldsymbol{\zeta}\|_4^4 . \tag{F.82}$$

Therefore, we can obtain that

$$\left\| \bar{\boldsymbol{H}}_E - \boldsymbol{H}_E \right\|_2 \le \frac{c}{2} \theta \left(1-\theta\right) \|\boldsymbol{\zeta}\|_4^4 \tag{F.83}$$

always holds, hence

$$\begin{aligned}&\mathbb{P}\left[\left\|\frac{\bar{\boldsymbol{X}}_0\operatorname{diag}\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 2}\bar{\boldsymbol{X}}_0^T}{m}-\boldsymbol{H}_E\right\|_2\geq c\theta(1-\theta)\left\|\boldsymbol{\zeta}\right\|_4^4\right]\\&\leq\mathbb{P}\left[\left\|\frac{\bar{\boldsymbol{X}}_0\operatorname{diag}\left(\bar{\boldsymbol{X}}_0^T\boldsymbol{\zeta}\right)^{\circ 2}\bar{\boldsymbol{X}}_0^T}{m}-\bar{\boldsymbol{H}}_E\right\|_2\geq\frac{c}{2}\theta(1-\theta)\left\|\boldsymbol{\zeta}\right\|_4^4\right].\end{aligned}\tag{F.84}$$

Independent Sub-matrices. As we did in Lemma E.2, we remove the dependence in $\boldsymbol{X}_0$ by sampling every $2k-1$ column such that

$$\boldsymbol{X}_0\boldsymbol{\Pi}=\left[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{2k-1}\right],\tag{F.85}$$

where

$$\boldsymbol{X}_i=\left[\boldsymbol{x}_i,\boldsymbol{x}_{i+(2k-1)},\cdots,\boldsymbol{x}_{i+(m-2k-1)}\right],\tag{F.86}$$

and $\boldsymbol{\Pi}$ is a certain permutation of the columns of $\boldsymbol{X}_0$.

Applying Bernstein inequality for matrix variables as in Lemma G.7, with $\boldsymbol{M}_i=\left\langle\bar{\boldsymbol{x}}_i,\boldsymbol{A}^T\boldsymbol{q}\right\rangle^2\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\in\mathbb{R}^{(2k-1)\times(2k-1)}$. Since

$$\begin{aligned}\left\|\boldsymbol{M}_i\right\|_2&=\left\|\left\langle\bar{\boldsymbol{x}}_i,\boldsymbol{A}^T\boldsymbol{q}\right\rangle^2\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right\|_2&\text{(F.87)}\\&\leq\left\|\bar{\boldsymbol{x}}_i\right\|_2^4&\text{(F.88)}\\&\leq 4B^4k^2&\text{(F.89)}\end{aligned}$$

and

$$\begin{aligned}\left\|\mathbb{E}\left[\boldsymbol{M}_i\boldsymbol{M}_i^*\right]\right\|&=\left\|\mathbb{E}\left[\boldsymbol{M}_i^*\boldsymbol{M}_i\right]\right\|&\text{(F.90)}\\&=\left\|\mathbb{E}\left[\left\langle\bar{\boldsymbol{x}}_i,\boldsymbol{A}^T\boldsymbol{q}\right\rangle^4\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right]\right\|&\text{(F.91)}\\&=\left\|\mathbb{E}\left[\left\langle\bar{\boldsymbol{x}}_i,\boldsymbol{\zeta}\right\rangle^4\left\|\bar{\boldsymbol{x}}_i\right\|_2^2\bar{\boldsymbol{x}}_i\bar{\boldsymbol{x}}_i^T\right]\right\|&\text{(F.92)}\\&\leq\mathbb{E}\left[\left\langle\bar{\boldsymbol{x}}_i,\boldsymbol{\zeta}\right\rangle^4\left\|\bar{\boldsymbol{x}}_i\right\|_2^4\right],&\text{(F.93)}\end{aligned}$$

we obtain the following upper bound:

$$\begin{aligned}&\mathbb{E}\left[\left\langle\bar{\boldsymbol{x}}_i,\boldsymbol{\zeta}\right\rangle^4\left\|\bar{\boldsymbol{x}}_i\right\|_2^4\right]\\&=\mathbb{E}\left[\sum_{j_1,j_2}^{2k-1}\bar{\boldsymbol{x}}_i\left(j_1\right)^2\bar{\boldsymbol{x}}_i\left(j_2\right)^2\sum_{j_3,\cdots,j_6}\prod_{\ell=3}^{6}\bar{\boldsymbol{x}}_i\left(j_\ell\right)\boldsymbol{\zeta}_{j_\ell}\right]&\text{(F.94)}\\&=3\mathbb{E}\left[\sum_{j_1,j_2}^{2k-1}\bar{\boldsymbol{x}}_i\left(j_1\right)^2\bar{\boldsymbol{x}}_i\left(j_2\right)^2\sum_{j_3\neq j_4}\bar{\boldsymbol{x}}_i\left(j_3\right)^2\boldsymbol{\zeta}_{j_3}^2\bar{\boldsymbol{x}}_i\left(j_4\right)^2\boldsymbol{\zeta}_{j_4}^2\right]\\&\quad+\mathbb{E}\left[\sum_{j_1,j_2}^{2k-1}\bar{\boldsymbol{x}}_i\left(j_1\right)^2\bar{\boldsymbol{x}}_i\left(j_2\right)^2\cdot\sum_{j_3}\bar{\boldsymbol{x}}_i\left(j_3\right)^4\boldsymbol{\zeta}_{j_3}^4\right]&\text{(F.95)}\\&=3\mathbb{E}\left[\sum_{\substack{j_1\neq j_2\\\neq j_3\neq j_4}}\bar{\boldsymbol{x}}_i\left(j_1\right)^2\bar{\boldsymbol{x}}_i\left(j_2\right)^2\bar{\boldsymbol{x}}_i\left(j_3\right)^2\boldsymbol{\zeta}_{j_3}^2\bar{\boldsymbol{x}}_i\left(j_4\right)^2\boldsymbol{\zeta}_{j_4}^2\right]\end{aligned}$$

$$
\begin{aligned}
&+ 3\mathbb{E}\left[\sum_{j_1\neq j_2\neq j_3} \bar{\boldsymbol{x}}_i\left(j_1\right)^4 \bar{\boldsymbol{x}}_i\left(j_2\right)^2 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i\left(j_3\right)^2 \boldsymbol{\zeta}_{j_3}^2\right] \\
&+ 6\mathbb{E}\left[\sum_{j_1\neq j_2} \bar{\boldsymbol{x}}_i\left(j_1\right)^6 \boldsymbol{\zeta}_{j_1}^2 \bar{\boldsymbol{x}}_i\left(j_2\right)^2 \boldsymbol{\zeta}_{j_2}^2\right] \\
&+ 6\mathbb{E}\left[\sum_{j_1\neq j_2\neq j_3} \bar{\boldsymbol{x}}_i\left(j_1\right)^2 \bar{\boldsymbol{x}}_i\left(j_2\right)^4 \boldsymbol{\zeta}_{j_2}^2 \bar{\boldsymbol{x}}_i\left(j_3\right)^2 \boldsymbol{\zeta}_{j_3}^2\right] \\
&+ 6\mathbb{E}\left[\sum_{j_1\neq j_2} \bar{\boldsymbol{x}}_i\left(j_1\right)^4 \boldsymbol{\zeta}_{j_1}^2 \bar{\boldsymbol{x}}_i\left(j_2\right)^4 \boldsymbol{\zeta}_{j_2}^2\right] \\
&+ 2\mathbb{E}\left[\sum_{j_1\neq j_2} \bar{\boldsymbol{x}}_i\left(j_1\right)^2 \bar{\boldsymbol{x}}_i\left(j_2\right)^6 \boldsymbol{\zeta}_{j_2}^4\right] \\
&+ \mathbb{E}\left[\sum_{j_1\neq j_2\neq j_3} \bar{\boldsymbol{x}}_i\left(j_1\right)^2 \bar{\boldsymbol{x}}_i\left(j_2\right)^2 \bar{\boldsymbol{x}}_i\left(j_3\right)^4 \boldsymbol{\zeta}_{j_3}^4\right] \\
&+ \mathbb{E}\left[\sum_{j_1\neq j_2} \bar{\boldsymbol{x}}_i\left(j_1\right)^4 \bar{\boldsymbol{x}}_i\left(j_2\right)^4 \boldsymbol{\zeta}_{j_2}^4\right] \\
&+ \mathbb{E}\left[\sum_{j} \bar{\boldsymbol{x}}_i\left(j\right)^8 \boldsymbol{\zeta}_{j}^4\right] \qquad \text{(F.96)} \\
&\leq \left(105\theta + 18\theta^2 k + 60\theta^2 k + 12\theta^3 k^2\right)\|\boldsymbol{\zeta}\|_4^4 \\
&\quad + 3\left(21\theta^2 + 30\theta^2 + 4\theta^4 k^2 + 12\theta^2 k\right)\|\boldsymbol{\zeta}\|_2^4 \qquad \text{(F.97)} \\
&\leq C\theta^3 k^2 \qquad \text{(F.98)}
\end{aligned}
$$

Assuming $\theta m \geq 1$, hence

$$
\sigma^2 = C\theta^3 k m. \qquad \text{(F.99)}
$$

Setting $t = \frac{c}{2}\frac{\theta(1-\theta)m\|\boldsymbol{\zeta}\|_4^4}{2k-1}$ in Matrix Bernstein gives that

$$
\begin{aligned}
&\mathbb{P}\left[\left\|\bar{\boldsymbol{X}}_i\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 3} - \mathbb{E}\left[\cdot\right]\right\|_2 > t\right] \\
&\qquad \leq 8k\exp\left(\frac{-t^2/2}{C\theta^3 k m + C' B^4\theta^2 k^2 t}\right), \qquad \text{(F.100)}
\end{aligned}
$$

we obtain that

$$
\begin{aligned}
&\mathbb{P}\left[\left\|\frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T\boldsymbol{\zeta}\right)^{\circ 2}\bar{\boldsymbol{X}}_i^T}{m} - \mathbb{E}\left[\cdot\right]\right\|_2 > c\frac{\theta\left(1-\theta\right)\|\boldsymbol{\zeta}\|_4^6}{2k-1}\right] \\
&\leq\ 8k\exp\left(-\frac{cm\left(1-\theta\right)^2\|\boldsymbol{\zeta}\|_4^8}{\theta k^3 + \theta\left(1-\theta\right)B^4 k^3\|\boldsymbol{\zeta}\|_4^4}\right). \qquad \text{(F.101)}
\end{aligned}
$$

ε-Net Covering To obtain a probability bound for all $\boldsymbol{q} \in \mathbb{S}^{k-1}$, we choose a set of $\boldsymbol{\zeta}_n = \boldsymbol{A}^T\boldsymbol{q}_n$ with $n =$

$1, \cdots, N$. Since for any $\boldsymbol{q}, \boldsymbol{q}' \in \mathbb{S}^{k-1}$ and $\boldsymbol{\zeta}' = \boldsymbol{A}^T \boldsymbol{q}'$, we have

$$
\begin{aligned}
&\left\| \frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}'\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} \right\|_2 \\
&= \frac{1}{m} \left\| \bar{\boldsymbol{X}}_i \operatorname{diag}\left[\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} - \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}'\right)^{\circ 2} \right] \bar{\boldsymbol{X}}_i^T \right\|_2 && \text{(F.102)} \\
&\leq \frac{\left\|\bar{\boldsymbol{X}}_i\right\|_2^2}{m} \left\| \operatorname{diag}\left[\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} - \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}'\right)^{\circ 2} \right] \right\|_2 && \text{(F.103)} \\
&\leq \frac{\left\|\bar{\boldsymbol{X}}_i\right\|_2^2}{m} \left\| \bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta} + \bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}' \right\|_\infty \left\| \bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta} - \bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}' \right\|_\infty && \text{(F.104)} \\
&\leq L \left\| \boldsymbol{q} - \boldsymbol{q}' \right\|_2 && \text{(F.105)}
\end{aligned}
$$

Then the Lipschitz constant L is upper bounded by

$$
\begin{aligned}
L &\leq \frac{\left\|\bar{\boldsymbol{X}}_i\right\|_2^2}{m} \left\| \bar{\boldsymbol{X}}_i^T \boldsymbol{A}^T \right\|_2 \left(\left\| \bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta} \right\|_\infty + \left\| \bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}' \right\|_\infty \right) && \text{(F.106)} \\
&\leq \frac{2}{m} \left\| \bar{\boldsymbol{X}}_i \right\|_2^4 && \text{(F.107)} \\
&\leq 8\theta^2 m B^4. && \text{(F.108)}
\end{aligned}
$$

With triangle inequality, we have

$$
\begin{aligned}
&\left\| \frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \mathbb{E}\left[\cdot\right] \right\|_2 \\
&\leq \left\| \frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}_n\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} \right\|_2 \\
&+ \left\| \mathbb{E}\left[\frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} \right] - \mathbb{E}\left[\frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}_n\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} \right] \right\|_2 \\
&+ \left\| \frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}_n\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \mathbb{E}\left[\frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}_n\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} \right] \right\|_2 && \text{(F.109)} \\
&\leq \left\| \frac{\bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}_n\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \mathbb{E}\left[\frac{\bar{\boldsymbol{X}}_i \left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}_n\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} \right] \right\|_2 \\
&+ 2L\varepsilon && \text{(F.110)}
\end{aligned}
$$

Next, we are going to choose the ε-net to cover the sphere of $\boldsymbol{q}$ with

$$
\varepsilon = \frac{c}{4} \frac{\theta\left(1-\theta\right)}{\left(2k-1\right) L} \min_{\boldsymbol{q} \in \mathbb{S}^{k-1}} \left\| \boldsymbol{\zeta} \right\|_4^4, \tag{F.111}
$$

hence the number of samples N is bounded by

$$
\begin{aligned}
N &= \left(\frac{3}{\varepsilon} \right)^k && \text{(F.112)} \\
&\leq \exp\left(-k \ln \varepsilon \right) && \text{(F.113)} \\
&\leq C \exp\left[k \log\left(\frac{\theta B^4 k^2 m}{1-\theta} \right) \right]. && \text{(F.114)}
\end{aligned}
$$

For $n = 1, \cdots, N$, denote

$$P_i\left(\boldsymbol{q}_n\right) = \mathbb{P}\left[\left\|\frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}_n\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \mathbb{E}[\cdot]\right\|_2 \geq \frac{c\theta(1-\theta)\|\boldsymbol{\zeta}_n\|_4^4}{4(2k-1)}\right] \tag{F.115}$$

together with union bound over all $\boldsymbol{q}_n$, we obtain

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \frac{\left\|\frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \mathbb{E}\left[\cdot\right]\right\|_2}{\|\boldsymbol{\zeta}\|_4^4} \geq \frac{c\theta\left(1-\theta\right)}{2(2k-1)}\right] \\
&\leq \sum_{\boldsymbol{q}_n\in\hat{\mathcal{R}}_{2C_\star}} P_i\left(\boldsymbol{q}_n\right) &\text{(F.116)}\\
&\leq N \max_{\boldsymbol{q}_n\in\hat{\mathcal{R}}_{2C_\star}} P_i\left(\boldsymbol{q}_n\right) &\text{(F.117)}\\
&\leq 8k \sup_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \exp\left(-\frac{cm\left(1-\theta\right)^2\|\boldsymbol{\zeta}\|_4^8}{\theta k^3 + \theta\left(1-\theta\right)B^4k^3\|\boldsymbol{\zeta}\|_4^4}\right) \times \\
&\quad \exp\left(k\ln\left(\frac{3}{\varepsilon}\right)\right). &\text{(F.118)}
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \frac{\left\|\frac{\bar{\boldsymbol{X}}_0 \operatorname{diag}\left(\bar{\boldsymbol{X}}_0^T \boldsymbol{\zeta}\right)^{\circ 2} \bar{\boldsymbol{X}}_0^T}{m} - \mathbb{E}\left[\cdot\right]\right\|_2}{\|\boldsymbol{\zeta}\|_4^4} \geq \frac{c}{2}\theta(1-\theta)\right] \\
&\leq \sum_i \mathbb{P}\left[\sup_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \frac{\left\|\frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \mathbb{E}\left[\cdot\right]\right\|_2}{\|\boldsymbol{\zeta}\|_4^4} \geq \frac{c\theta(1-\theta)}{2(2k-1)}\right] &\text{(F.119)}\\
&\leq (2k-1)\max_i \\
&\mathbb{P}\left[\sup_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \frac{\left\|\frac{\bar{\boldsymbol{X}}_i \operatorname{diag}\left(\bar{\boldsymbol{X}}_i^T \boldsymbol{\zeta}\right)^{\circ 2} \bar{\boldsymbol{X}}_i^T}{m} - \mathbb{E}\left[\cdot\right]\right\|_2}{\|\boldsymbol{\zeta}\|_4^4} \geq \frac{c\theta\left(1-\theta\right)}{2(2k-1)}\right] &\text{(F.120)}\\
&\leq 16k^2 \sup_{\boldsymbol{q}\in\hat{\mathcal{R}}_{2C_\star}} \exp\left(-\frac{c'm\left(1-\theta\right)^2\|\boldsymbol{\zeta}\|_4^8}{\theta k^3 + \theta(1-\theta)B^4k^3\|\boldsymbol{\zeta}\|_4^4}\right) \times \\
&\quad \exp\left(k\ln\left(\frac{3}{\varepsilon}\right)\right) &\text{(F.121)}
\end{aligned}$$

Therefore, by taking

$$m \geq \frac{C\theta}{\left(1-\theta\right)^2} \min\left\{\left(2C_\star\mu\kappa^2\right)^{-4/3}, k^2\right\} k^4 \log^3 k \tag{F.122}$$

$$\geq C'\theta k \log\left(\frac{\theta k m B^4}{(1-\theta)\left\|\boldsymbol{\zeta}\right\|_4^4}\right)\frac{k^3+(1-\theta)B^4k^3\left\|\boldsymbol{\zeta}\right\|_4^4}{(1-\theta)^2\left\|\boldsymbol{\zeta}\right\|_4^8} \tag{F.123}$$

and adding up failure probability, we obtain

$$\begin{aligned}&\left\|\frac{1}{m}\boldsymbol{X}_0\operatorname{diag}\left(\boldsymbol{X}_0^T\boldsymbol{A}^T\boldsymbol{q}\right)^{\circ 2}\boldsymbol{X}_0^T-\mathbb{E}\left[\cdot\right]\right\|_2\\ &\qquad\leq c\theta\left(1-\theta\right)\left\|\boldsymbol{A}^T\boldsymbol{q}\right\|_4^4\end{aligned} \tag{F.124}$$

with probability no smaller than $1-c_1\exp\left(-k\right)-c_2\theta\left(1-\theta\right)^2k^{-4}-c_3\exp\left(-\theta k\right)$. ■

G Tools

Lemma G.1 (Moments of the Gaussian Random Variables)**.** *If $X\sim\mathcal{N}\left(0,\sigma^2\right)$, then it holds for all integer $p\geq 1$ that*

$$\mathbb{E}\left[\left|X\right|^p\right]=\sigma^p\left(p-1\right)!!\left[\sqrt{\frac{2}{\pi}}\mathbb{1}_{p\text{ odd}}+\mathbb{1}_{p\text{ even}}\right] \tag{G.1}$$

$$\leq\sigma^p\left(p-1\right)!!. \tag{G.2}$$

Lemma G.2 (Moments of the χ^2 Random Variables)**.** *If $X\sim\chi^2\left(n\right)$, then it holds for all integer $p\geq 1$,*

$$\mathbb{E}\left[X^p\right]=2^p\frac{\Gamma\left(p+n/2\right)}{\Gamma\left(n/2\right)} \tag{G.3}$$

$$=\prod_{k=1}^{p}(n+2k-2)\leq p!(2n)^p/2 \tag{G.4}$$

Lemma G.3 (Moments of the χ Random Variables)**.** *If $X\sim\chi\left(n\right)$, then it holds for all integer $p\geq 1$,*

$$\mathbb{E}\left[X^p\right]=2^{p/2}\frac{\Gamma\left(p/2+n/2\right)}{\Gamma\left(n/2\right)}\leq p!!n^{p/2}. \tag{G.5}$$

Lemma G.4 (Moment-Control Bernstein's Inequality for Scalar RVs, Theorem 2.10 of [FR13])**.** *Let $X_1,\ldots,X_p$ be i.i.d. real-valued random variables. Suppose that there exist some positive number R and σ^2 such that*

$$\mathbb{E}\left[\left|X_k\right|^m\right]\leq\frac{m!}{2}\sigma^2R^{m-2},\ \text{ for all integers } m\geq 2.$$

Let $S\doteq\frac{1}{p}\sum_{k=1}^{p}X_k$, then for all $t>0$, it holds that

$$\mathbb{P}\left[\left|S-\mathbb{E}\left[S\right]\right|\geq t\right]\leq 2\exp\left(-\frac{pt^2}{2\sigma^2+2Rt}\right). \tag{G.6}$$

Corollary G.5 (Moment-Control Bernstein's Inequality for Vector RVs, Corollary A.10 of [SQW15])**.** *Let $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p\in\mathbb{R}^d$ be i.i.d. random vectors. Suppose there exist some positive number R and σ^2 such that*

$$\mathbb{E}\left[\left\|\boldsymbol{x}_k\right\|^m\right]\ \leq\ \frac{m!}{2}\sigma^2R^{m-2},\quad \text{for all integers } m\geq 2.$$

Let $\boldsymbol{s}=\frac{1}{p}\sum_{k=1}^{p}\boldsymbol{x}_k$, then for any $t>0$, it holds that

$$\mathbb{P}\left[\left\|\boldsymbol{s}-\mathbb{E}\left[\boldsymbol{s}\right]\right\|\geq t\right]\ \leq\ 2(d+1)\exp\left(-\frac{pt^2}{2\sigma^2+2Rt}\right). \tag{G.7}$$

Lemma G.6 (Moment-Control Bernstein's Inequality for Matrix RVs, Theorem 6.2 of [Tro12]). *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p \in \mathbb{R}^{d\times d}$ be i.i.d. random, symmetric matrices. Suppose there exist some positive number R and σ^2 such that*

$$\mathbb{E}\left[\boldsymbol{X}_k^m\right] \preceq \frac{m!}{2}\sigma^2 R^{m-2}\boldsymbol{I}, \tag{G.8}$$

$$-\mathbb{E}\left[\boldsymbol{X}_k^m\right] \preceq \frac{m!}{2}\sigma^2 R^{m-2}\boldsymbol{I}. \tag{G.9}$$

for all integers $m \geq 2$. Let $\boldsymbol{S} \doteq \frac{1}{p}\sum_{k=1}^{p}\boldsymbol{X}_k$, then for all $t > 0$, it holds that

$$\mathbb{P}\left[\|\boldsymbol{S} - \mathbb{E}\left[\boldsymbol{S}\right]\| \geq t\right] \leq 2d\exp\left(-\frac{pt^2}{2\sigma^2 + 2Rt}\right). \tag{G.10}$$

Lemma G.7 (Bernstein's Inequality for Uncentered Matrix RVs). *The matrix Bernstein inequality states that for independent random matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_n \in \mathbb{R}^{d_1\times d_2}$, if*

$$\sigma^2 = \max\left\{\left\|\sum_{i=1}^{n}\mathbb{E}[\boldsymbol{M}_i\boldsymbol{M}_i^*]\right\|, \left\|\sum_{i=1}^{n}\mathbb{E}[\boldsymbol{M}_i^*\boldsymbol{M}_i]\right\|\right\}, \tag{G.11}$$

and

$$\|\boldsymbol{M}_i\|_2 \leq R \qquad a.s., \tag{G.12}$$

then

$$\mathbb{P}\left[\left\|\sum_i \boldsymbol{M}_i - \mathbb{E}\left[\cdot\right]\right\| > t\right] \leq (d_1 + d_2)\exp\left(\frac{-t^2/2}{\sigma^2 + 2Rt/3}\right). \tag{G.13}$$

Proof For zero mean random matrices

$$\boldsymbol{M}_1 - \mathbb{E}\boldsymbol{M}_1, \ldots, \boldsymbol{M}_n - \mathbb{E}\boldsymbol{M}_n \in \mathbb{R}^{d_1\times d_2}, \tag{G.14}$$

we have that

$$\|\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i\|_2 \leq 2R, \tag{G.15}$$

and

$$\mathbf{0} \preceq \sum_{i=1}^{n}\mathbb{E}[(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)^*] \tag{G.16}$$

$$\preceq \sum_{i=1}^{n}\mathbb{E}[\boldsymbol{M}_i\boldsymbol{M}_i^*], \tag{G.17}$$

$$\mathbf{0} \preceq \sum_{i=1}^{n}\mathbb{E}[(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)^*(\boldsymbol{M}_i - \mathbb{E}\boldsymbol{M}_i)] \tag{G.18}$$

$$\preceq \sum_{i=1}^{n}\mathbb{E}[\boldsymbol{M}_i^*\boldsymbol{M}_i]. \tag{G.19}$$

Plugging corresponding quantities back to Theorem 1.6 of [Tro12], we obtain that

$$\mathbb{P}\left[\left\|\sum_i \boldsymbol{M}_i - \mathbb{E}\left[\cdot\right]\right\| > t\right] \leq (d_1 + d_2)\exp\left(\frac{-t^2/2}{\sigma^2 + 2Rt/3}\right). \tag{G.20}$$

■