# Geometry and Symmetry in Short-and-Sparse Deconvolution

Han-Wen Kuo[1,2], Yenson Lau[1,2], Yuqian Zhang[3], John Wright[1,2,4]

[1]Department of Electrical Engineering, Columbia University
[2]Data Science Institute, Columbia University
[3]Department of Computer Science, Cornell University
[4]Department of Applied Physics and Applied Mathematics, Columbia University

January 3, 2019 Revised January 6, 2019

**Abstract**

We study the *Short-and-Sparse (SaS) deconvolution* problem of recovering a short signal $\boldsymbol{a}_0$ and a sparse signal $\boldsymbol{x}_0$ from their convolution. We propose a method based on nonconvex optimization, which under certain conditions recovers the target short and sparse signals, up to a signed shift symmetry which is intrinsic to this model. This symmetry plays a central role in shaping the optimization landscape for deconvolution. We give a *regional analysis*, which characterizes this landscape geometrically, on a union of subspaces. Our geometric characterization holds when the length-$p_0$ short signal $\boldsymbol{a}_0$ has shift coherence $\mu$, and $\boldsymbol{x}_0$ follows a random sparsity model with sparsity rate $\theta \in \left[\frac{c_1}{p_0}, \frac{c_2}{p_0\sqrt{\mu}+\sqrt{p_0}}\right] \cdot \frac{1}{\log^2 p_0}$. Based on this geometry, we give a provable method that successfully solves SaS deconvolution with high probability.

## 1 Introduction

Datasets in a wide range of areas, including neuroscience [Lew98], microscopy [CLC+17] and astronomy [Sah07], can be modeled as superpositions of translations of a basic motif. Data of this nature can be modeled mathematically as a convolution $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$, between a *short* signal $\boldsymbol{a}_0$ (the motif) and a longer *sparse* signal $\boldsymbol{x}_0$, whose nonzero entries indicate where in the sample the motif is present. A very similar structure arises in image deblurring [CW98], where $\boldsymbol{y}$ is a blurry image, $\boldsymbol{a}_0$ the blur kernel, and $\boldsymbol{x}_0$ the (edge map) of the target sharp image.

Motivated by these and related problems in imaging and scientific data analysis, we study the *Short-and-Sparse (SaS) Deconvolution* problem of recovering a short signal $\boldsymbol{a}_0 \in \mathbb{R}^{p_0}$ and a sparse signal $\boldsymbol{x}_0 \in \mathbb{R}^n$ ($n \gg p_0$) from their length-$n$ cyclic convolution $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 \in \mathbb{R}^n$. This SaS model exhibits a basic *scaled shift symmetry*: for any nonzero scalar $\alpha$ and cyclic shift $s_\ell[\cdot]$,

$$\Big(\alpha\, s_\ell[\boldsymbol{a}_0]\Big) \;*\; \Big(\tfrac{1}{\alpha}\, s_{-\ell}[\boldsymbol{x}_0]\Big) \;=\; \boldsymbol{y}. \tag{1.1}$$

Because of this symmetry, we only expect to recover $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ up to a signed shift (see Figure 1). Our problem of interest can be stated more formally as:

**Problem 1.1** (Short-and-Sparse Deconvolution). *Given the cyclic convolution $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0 \in \mathbb{R}^n$ of $\boldsymbol{a}_0 \in \mathbb{R}^{p_0}$ short ($p_0 \ll n$), and $\boldsymbol{x}_0 \in \mathbb{R}^n$ sparse, recover $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$, up to a scaled shift.*

Despite a long history and many applications, until recently very little algorithmic theory was available for SaS deconvolution. Much of this difficulty can be attributed to the scale-shift symmetry: natural convex relaxations fail, and nonconvex formulations exhibit a complicated optimization landscape, with many

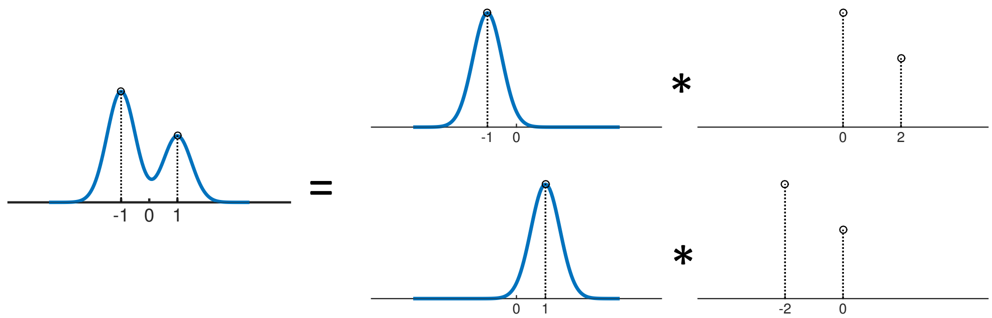


**Figure 1: Shift symmetry in Short-and-Sparse deconvolution.** An observation $\boldsymbol{y}$ (left) which is a convolution of a short signal $\boldsymbol{a}_0$ and a sparse signal $\boldsymbol{x}_0$ (top right) can be equivalently expressed as a convolution of $s_\ell[\boldsymbol{a}_0]$ and $s_{-\ell}[\boldsymbol{x}_0]$, where $s_\ell[\cdot]$ denotes a shift $\ell$ samples. The ground truth signals $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ can only be identified up to a scaled shift.

equivalent global minimizers (scaled shifts of the ground truth) and additional local minimizers (scaled shift truncations of the ground truth), and a variety of critical points [ZLK+17, ZKW18]. Currently available theory guarantees approximate recovery of a truncation[1] of a shift $s_\ell[\boldsymbol{a}_0]$, rather than guaranteeing recovery of $\boldsymbol{a}_0$ as a whole, and requires certain (complicated) conditions on the convolution matrix associated with $\boldsymbol{a}_0$ [ZKW18].

In this paper, describe an algorithm which, under simpler conditions, *exactly* recovers a scaled shift of the pair $(\boldsymbol{a}_0, \boldsymbol{x}_0)$. Our algorithm is based on a formulation first introduced in [ZLK+17], which casts the deconvolution problem as (nonconvex) optimization over the sphere. We characterize the geometry of this objective function, and show that near a certain union of subspaces, every local minimizer is very close to a signed shift of $\boldsymbol{a}_0$. Based on this geometric analysis, we give provable methods for SaS deconvolution that exactly recover a scaled shift of $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ whenever $\boldsymbol{a}_0$ is *shift-incoherent* and $\boldsymbol{x}_0$ is a sufficiently sparse random vector. Our geometric analysis highlights the role of symmetry in shaping the objective landscape for SaS deconvolution.

**Organization of this paper.** The remainder of this paper is organized as follows. Section 2 introduces our optimization approach and modeling assumptions. Section 3 introduces our main results — both geometric and algorithmic — and compares them to the literature. Section 4-5 describes the main ideas of our analysis. Finally, Section 7 discusses two main limitations of our analysis and describes directions for future work.

## 2 Formulation and Assumptions

### 2.1 Nonconvex SaS over the Sphere

**Bilinear Lasso.** Our starting point is the (natural) formulation

$$\min_{\boldsymbol{a},\boldsymbol{x}} \ \underbrace{\tfrac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x} - \boldsymbol{y}\right\|_2^2}_{\text{Data Fidelity}} + \underbrace{\lambda \left\|\boldsymbol{x}\right\|_1}_{\text{Sparsity}} \quad \text{s.t.} \quad \left\|\boldsymbol{a}\right\|_2 = 1. \tag{2.1}$$

We term this optimization problem the *Bilinear Lasso*, for its resemblance to the Lasso estimator in statistics. Indeed, letting

$$\varphi_{\text{lasso}}(\boldsymbol{a}) \equiv \min_{\boldsymbol{x}} \left\{ \tfrac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x} - \boldsymbol{y}\right\|_2^2 + \lambda \left\|\boldsymbol{x}\right\|_1 \right\} \tag{2.2}$$

denote the optimal Lasso cost, we see that (2.1) simply optimizes $\varphi_{\text{lasso}}$ with respect to $\boldsymbol{a}$:

$$\min_{\boldsymbol{a}} \ \varphi_{\text{lasso}}(\boldsymbol{a}) \quad \text{s.t.} \quad \left\|\boldsymbol{a}\right\|_2 = 1. \tag{2.3}$$

[1] I.e., the portion of the shifted signal $s_\ell[\boldsymbol{a}_0]$ that falls in the window $\{0, \dots, p_0 - 1\}$.

In (2.1)-(2.3), we constrain $\boldsymbol{a}$ to have unit $\ell^2$ norm. This constraint breaks the scale ambiguity between $\boldsymbol{a}$ and $\boldsymbol{x}$. Moreover, the choice of constraint manifold has surprisingly strong implications for computation: if $\boldsymbol{a}$ is instead constrained to the simplex, the problem admits trivial global minimizers. In contrast, local minima of the sphere-constrained formulation often correspond to shifts (or shift truncations [ZLK+17]) of the ground truth $\boldsymbol{a}_0$.

**Simplifications and approximations.** The problem (2.3) is defined in terms of the optimal Lasso cost. This function is challenging to analyze, especially far away from $\boldsymbol{a}_0$. [ZLK+17] analyzes the local minima of a simplification of (2.3), obtained by approximating[2] the data fidelity term as

$$\begin{aligned} \tfrac{1}{2}\left\|\boldsymbol{a}*\boldsymbol{x}-\boldsymbol{y}\right\|_2^2 &= \tfrac{1}{2}\left\|\boldsymbol{a}*\boldsymbol{x}\right\|_2^2-\langle\boldsymbol{a}*\boldsymbol{x},\boldsymbol{y}\rangle+\tfrac{1}{2}\left\|\boldsymbol{y}\right\|_2^2, \\ &\approx \tfrac{1}{2}\left\|\boldsymbol{x}\right\|_2^2-\langle\boldsymbol{a}*\boldsymbol{x},\boldsymbol{y}\rangle+\tfrac{1}{2}\left\|\boldsymbol{y}\right\|_2^2. \end{aligned} \tag{2.4}$$

This yields a simpler objective function

$$\varphi_{\ell^1}(\boldsymbol{a})=\min_{\boldsymbol{x}}\left\{\tfrac{1}{2}\left\|\boldsymbol{x}\right\|_2^2-\langle\boldsymbol{a}*\boldsymbol{x},\boldsymbol{y}\rangle+\tfrac{1}{2}\left\|\boldsymbol{y}\right\|_2^2+\lambda\left\|\boldsymbol{x}\right\|_1\right\}. \tag{2.5}$$

We make one further simplification to this problem, replacing the nondifferentiable penalty $\|\cdot\|_1$ with a smooth approximation $\rho(\boldsymbol{x})$.[3] Our analysis allows for a variety of smooth sparsity surrogates $\rho(\boldsymbol{x})$; for concreteness, we state our main results for the particular penalty[4]

$$\rho(\boldsymbol{x})=\textstyle\sum_i\left(\boldsymbol{x}_i^2+\delta^2\right)^{1/2}. \tag{2.6}$$

For $\delta>0$, this is a smooth function of $\boldsymbol{x}$; as $\delta\searrow 0$ it approaches $\|\boldsymbol{x}\|_1$. Replacing $\|\cdot\|_1$ with $\rho(\cdot)$, we obtain the objective function which will be our main object of study,

$$\varphi_{\rho}(\boldsymbol{a})=\min_{\boldsymbol{x}}\left\{\tfrac{1}{2}\left\|\boldsymbol{x}\right\|_2^2-\langle\boldsymbol{a}*\boldsymbol{x},\boldsymbol{y}\rangle+\tfrac{1}{2}\left\|\boldsymbol{y}\right\|_2^2+\lambda\rho(\boldsymbol{x})\right\}. \tag{2.7}$$

**Core optimization problem.** As in [ZLK+17], we optimize $\varphi_\rho(\boldsymbol{a})$ over the sphere $\mathbb{S}^{p-1}$:

$$\boxed{\min_{\boldsymbol{a}}\ \varphi_\rho(\boldsymbol{a})\quad \text{s.t.}\quad \boldsymbol{a}\in\mathbb{S}^{p-1}.} \tag{2.8}$$

Here, we set $p=3p_0-2$. As we will see, optimizing over this slightly higher dimensional sphere enables us to recover a (full) shift of $\boldsymbol{a}_0$, rather than a *truncated* shift. Our approach will leverage the following fact: if we view $\boldsymbol{a}\in\mathbb{S}^{p-1}$ as indexed by coordinates $W=\{-p_0+1,\dots,2p_0-1\}$ , then for any shifts $\ell\in\{-p_0+1,\dots,p_0-1\}$, the support of $\ell$-shifted short signal $s_\ell[\boldsymbol{a}_0]$ is entirely contained in interval $W$. We will give a provable method which recovers a scaled version of one of these canonical shifts.

## 2.2 Analysis Setting and Assumptions

For convenience, we assume that $\boldsymbol{a}_0$ has unit $\ell^2$ norm, i.e., $\boldsymbol{a}_0\in\mathbb{S}^{p_0-1}$.[5] Our analysis makes two main assumptions, on the short motif $\boldsymbol{a}_0$ and the sparse map $\boldsymbol{x}_0$, respectively:

**Shift incoherence of $\boldsymbol{a}_0$.** The first is that distinct shifts $\boldsymbol{a}_0$ have small inner product. We define the *shift coherence* of $\mu(\boldsymbol{a}_0)$ to be the largest inner product between distinct shifts:

$$\mu(\boldsymbol{a}_0)=\max_{\ell\neq 0}\left|\langle\boldsymbol{a}_0,s_\ell[\boldsymbol{a}_0]\rangle\right| \tag{2.9}$$

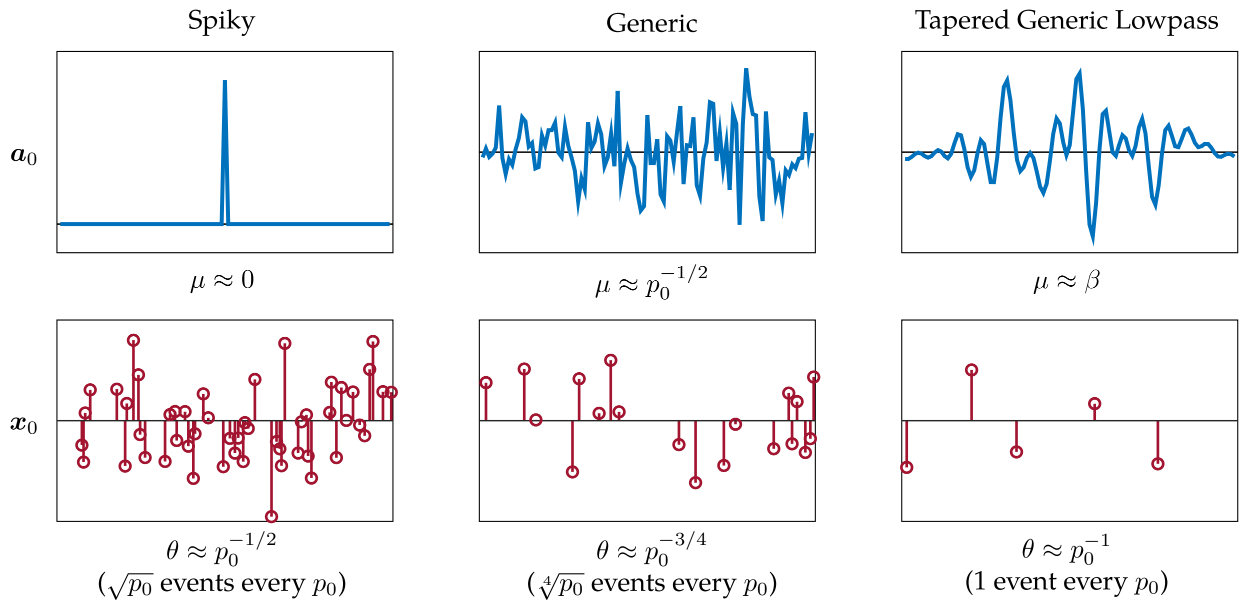


**Figure 2: Sparsity-coherence tradeoff:** Top: three families of motifs $\boldsymbol{a}_0$ with varying coherence $\mu$. Bottom: maximum allowable sparsity $\theta$ and number of copies $\theta p_0$ within each length-$p_0$ window. Here, we suppress constants and logarithmic factors. When the target motif has smaller shift-coherence $\mu$, our result allows larger $\theta$, and vise versa. This sparsity-coherence tradeoff is made precise in our main result Theorem 3.1, which, loosely speaking, asserts that when $\theta \lessapprox 1/(p_0\sqrt{\mu} + \sqrt{p_0})$, our method succeeds.

The quantity $\mu(\boldsymbol{a}_0)$ is bounded between $0$ and $1$. Our theory allows any $\mu$ smaller than some numerical constant. Figure 2 shows three examples of families of $\boldsymbol{a}_0$ that satisfy this assumption:

- *Spiky.* When $\boldsymbol{a}_0$ is close to the Dirac delta $\boldsymbol{\delta}_0$, the shift coherence $\mu(\boldsymbol{a}_0) \approx 0$.[6] Here, the observed signal $\boldsymbol{y}$ consists of a superposition of sharp pulses. This is arguably the easiest instance of SaS deconvolution.
- *Generic.* If $\boldsymbol{a}_0$ is chosen uniformly at random from the sphere $\mathbb{S}^{p_0-1}$, its coherence is bounded as $\mu(\boldsymbol{a}_0) \lessapprox \sqrt{1/p_0}$ with high probability.
- *Tapered Generic Lowpass.* Here, $\boldsymbol{a}_0$ is generated by taking a random conjugate symmetric superposition of the first $L$ length-$p_0$ Discrete Fourier Transform (DFT) basis signals, windowing (e.g., with a Hamming window) and normalizing to unit $\ell^2$ norm. When $L = p_0\sqrt{1-\beta}$, with high probability $\mu(\boldsymbol{a}_0) \lessapprox \beta$. In this model, $\mu$ does not have to diminish as $p_0$ grows – it can be a fixed constant.[7]

Intuitively speaking, problems with smaller $\mu$ are easier to solve, a claim which will be made precise in our technical results.

[2]For a generic $\boldsymbol{a}$, we have $\langle s_i[\boldsymbol{a}], s_j[\boldsymbol{a}]\rangle \approx 0$ and hence $\|\boldsymbol{a} * \boldsymbol{x}\|_2^2 = \boldsymbol{x}^*\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}}\boldsymbol{x} \approx \boldsymbol{x}^*\boldsymbol{I}\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$.

[3]The objective $\varphi_{\ell^1}$ is not twice differentiable everywhere, and hence cannot be minimized using conventional second order methods.

[4]This particular surrogate is sometimes being named as the pseudo-Huber function.

[5]This is purely a technical convenience. Our theory guarantees recovery of a signed shift $(\pm s_\ell[\boldsymbol{a}_0], \pm s_{-\ell}[\boldsymbol{x}_0])$ of the truth. If $\boldsymbol{a}_0$ does not have unit norm, identical reasoning implies that our method recovers a scaled shift $\left(\alpha s_\ell[\boldsymbol{a}_0], \alpha^{-1} s_{-\ell}[\boldsymbol{x}_0]\right)$ with $\alpha = \pm\frac{1}{\|\boldsymbol{a}_0\|_2}$.

[6]The use of "$\approx$" here suppresses constant and logarithmic factors.

[7]The upper right panel of Figure 2 is generated using random DFT components with frequencies smaller then one-third Nyquist. Such a kernel is incoherent, with high probability. Many commonly occurring low-pass kernels have $\mu(\boldsymbol{a}_0)$ larger – very close to one. One of the most important limitations of our results is that they do not provide guarantees in this highly coherent situation.

**Random sparsity model on $\boldsymbol{x}_0$.** We assume that $\boldsymbol{x}_0$ is a sparse random vector. More precisely, we assume that $\boldsymbol{x}_0$ is Bernoulli-Gaussian, with rate $\theta$:

$$\boldsymbol{x}_{0i} = \boldsymbol{\omega}_i \boldsymbol{g}_i, \tag{2.10}$$

where $\boldsymbol{\omega}_i \sim \mathrm{Ber}(\theta)$, $\boldsymbol{g}_i \sim \mathcal{N}(0,1)$ and all random variables are jointly independent. We write this as

$$\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta). \tag{2.11}$$

Here, $\theta$ is the probability that a given entry $\boldsymbol{x}_{0i}$ is nonzero. Problems with smaller $\theta$ are easier to solve. In the extreme case, when $\theta \ll 1/p_0$, the observation $\boldsymbol{y}$ contains many isolated copies of the motif $\boldsymbol{a}_0$, and $\boldsymbol{a}_0$ can be determined by direct inspection. Our analysis will focus on the nontrivial scenario, when $\theta \gtrapprox 1/p_0$.

**Sparsity-Coherence tradeoffs.** Our technical results will articulate *sparsity-coherence* tradeoffs, in which smaller coherence $\mu$ enables larger $\theta$, and vice-versa. More specifically, in our main theorem, the sparsity-coherence relationship is captured in the form

$$\theta \lessapprox 1/(p_0\sqrt{\mu} + \sqrt{p_0}). \tag{2.12}$$

When the target $\boldsymbol{a}_0$ is highly shift-incoherent ($\mu \approx 0$), our method succeeds when each length-$p_0$ window contains about $\sqrt{p_0}$ copies of $\boldsymbol{a}_0$. When $\mu$ is larger (as in the generic lowpass model), our method succeeds as long as relatively few copies of $\boldsymbol{a}_0$ overlap in the observed signal. In Figure 2, we illustrate these tradeoffs for the three models described above.

# 3 Main Results: Geometry and Algorithms

In this section, we introduce our main results – on the geometry of $\varphi_\rho$ (Section 3.1) and its algorithmic implications (Section 3.2). Finally, in Section 3.3, we compare these results with the literature on deconvolution.

## 3.1 Geometry of the Objective $\varphi_\rho$

The goal in SaS deconvolution is to recover $\boldsymbol{a}_0$ (and $\boldsymbol{x}_0$) up to a signed shift — i.e., we wish to recover some $\pm s_\ell[\boldsymbol{a}_0]$. The shifts $\pm s_\ell[\boldsymbol{a}_0]$ play a key role in shaping the landscape of $\varphi_\rho$. In particular, we will argue that over a certain subset of the sphere, *every local minimum of $\varphi_\rho$ is close to some $\pm s_\ell[\boldsymbol{a}_0]$*.

**Geometry near a single shift.** To gain intuition into the properties of $\varphi_\rho$, we first visualize this function in the vicinity of a single shift $s_\ell[\boldsymbol{a}_0]$ of the ground truth $\boldsymbol{a}_0$. In Figure 3, we plot the function value of $\varphi_\rho$ over

$$\mathcal{B}_{\ell^2,r}(s_\ell[\boldsymbol{a}_0]) \cap \mathbb{S}^{p-1},$$

where $\mathcal{B}_{\ell^2,r}(\boldsymbol{a})$ is a ball of radius $r$ around $\boldsymbol{a}$. We make two observations:

- The objective function $\varphi_\rho$ is strongly convex on this neighborhood of $s_\ell[\boldsymbol{a}_0]$.
- There is a local minimizer very close to $s_\ell[\boldsymbol{a}_0]$.

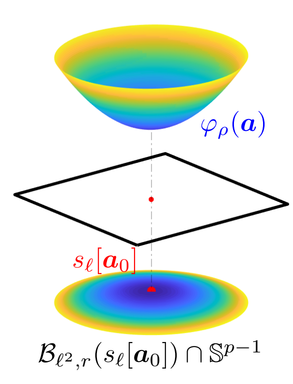


**Figure 3: Geometry of $\varphi_\rho$ near a shift of $\boldsymbol{a}_0$.** Bottom: a portion of the sphere $\mathbb{S}^{p-1}$, colored according to $\varphi_\rho$. Top: $\varphi_\rho$ visualized as height. $\varphi_\rho$ is strongly convex in this region, and it has a minimizer very close to $s_\ell[\boldsymbol{a}_0]$.

**Geometry near the span of two shifts.** We next visualize the objective function $\varphi_\rho$ near the linear span of *two* different shifts $s_{\ell_1}[\boldsymbol{a}_0]$ and $s_{\ell_2}[\boldsymbol{a}_0]$. More precisely, we plot $\varphi_\rho$ near the intersection (Figure 4, left) of the sphere $\mathbb{S}^{p-1}$ and the linear subspace

$$\mathcal{S}_{\{\ell_1,\ell_2\}} = \{\, \boldsymbol{\alpha}_1 s_{\ell_1}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_2 s_{\ell_2}[\boldsymbol{a}_0] \mid \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R} \,\}.$$

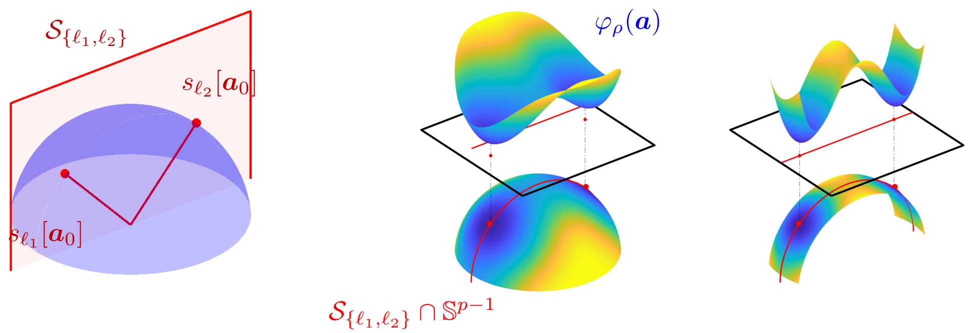


**Figure 4: Geometry of $\varphi_\rho$ near the span $\mathcal{S}_{\{\ell_1,\ell_2\}}$ of two shifts of $\boldsymbol{a}_0$.** Left: each pair of shifts $s_{\ell_1}[\boldsymbol{a}_0]$, $s_{\ell_2}[\boldsymbol{a}_0]$ defines a linear subspace $\mathcal{S}_{\{\ell_1,\ell_2\}}$ of $\mathbb{R}^p$. Center/right: every local minimum of $\varphi_\rho$ near $\mathcal{S}_{\{\ell_1,\ell_2\}}$ (red line) is close to either $s_{\ell_1}[\boldsymbol{a}_0]$ or $s_{\ell_2}[\boldsymbol{a}_0]$; there is a negative curvature in the middle of $s_{\ell_1}[\boldsymbol{a}_0]$, $s_{\ell_2}[\boldsymbol{a}_0]$, and $\varphi_\rho$ is convex in direction away from $\mathcal{S}_{\ell_1,\ell_2}$.

We make three observations:

- Again, there is a local minimizer near each shift $s_\ell[\boldsymbol{a}_0]$.
- These are the *only* local minimizers in the vicinity of $\mathcal{S}_{\{\ell_1,\ell_2\}}$. In particular, the objective function $\varphi$ exhibits *negative curvature* along $S_{\{\ell_1,\ell_2\}}$ at any superposition $\boldsymbol{\alpha}_1 s_{\ell_1}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_2 s_{\ell_2}[\boldsymbol{a}_0]$ whose weights $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are balanced, i.e., $|\boldsymbol{\alpha}_1| \approx |\boldsymbol{\alpha}_2|$.
- Furthermore, the function $\varphi_\rho$ exhibits *positive curvature* in directions away from the subspace $\mathcal{S}_{\ell_1,\ell_2}$.

**Geometry in the span of multiple shifts.** Finally, we visualize $\varphi_\rho$ over the intersection (Figure 5, left) of the sphere $\mathbb{S}^{p-1}$ with the linear span of three shifts $s_{\ell_1}[\boldsymbol{a}_0]$, $s_{\ell_2}[\boldsymbol{a}_0]$, $s_{\ell_3}[\boldsymbol{a}_0]$ of the true kernel $\boldsymbol{a}_0$:

$$\mathcal{S}_{\{\ell_1,\ell_2,\ell_3\}} = \{\, \boldsymbol{\alpha}_1 s_{\ell_1}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_2 s_{\ell_2}[\boldsymbol{a}_0] + \boldsymbol{\alpha}_3 s_{\ell_3}[\boldsymbol{a}_0] \mid \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3 \in \mathbb{R} \,\}$$

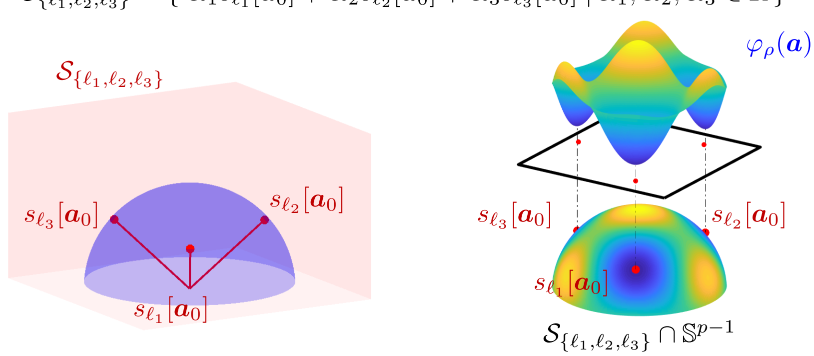


**Figure 5: Geometry of $\varphi_\rho$ over the span $\mathcal{S}_{\{\ell_1,\ell_2,\ell_3\}}$ of three shifts of $\boldsymbol{a}_0$.** The subspace $\mathcal{S}_{\{\ell_1,\ell_2,\ell_3\}}$ is three-dimensional; its intersection with the sphere $\mathbb{S}^{p-1}$ is isomorphic to a two-dimensional sphere. On this set, $\varphi_\rho$ has local minimizers near each of the $s_{\ell_i}[\boldsymbol{a}_0]$, and are the only minimizers near $\mathcal{S}_{\ell_1,\ell_2,\ell_3}$.

Again, *there is a local minimizer near each signed shift*. At roughly balanced superpositions of shifts, the objective function exhibits negative curvature. As a result, again, the *only* local minimizers are close to signed shifts.

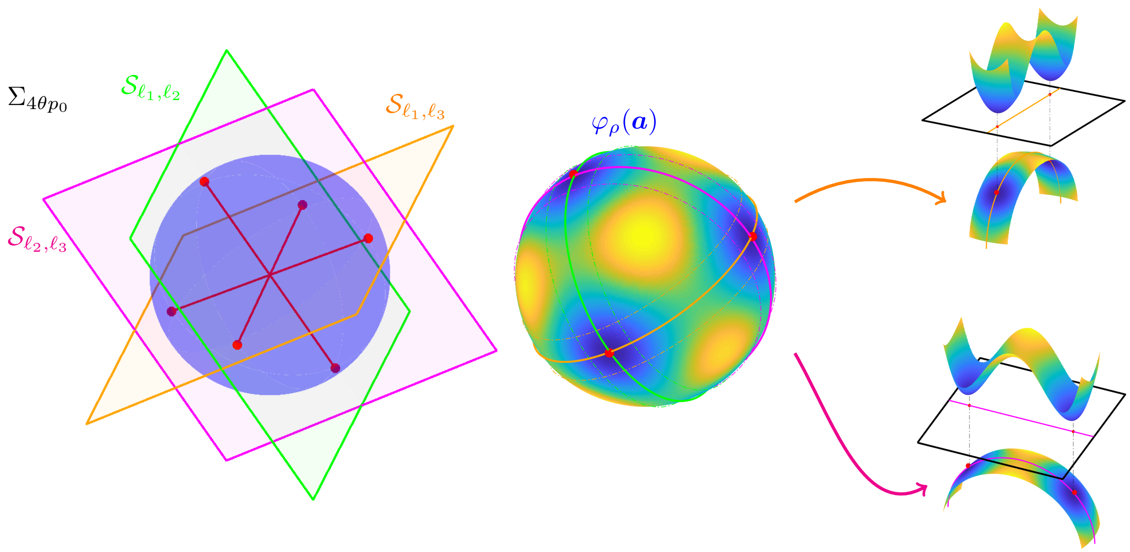


**Figure 6: Geometry of $\varphi_\rho$ over the union of subspaces $\Sigma_{4\theta p_0}$.** Left: schematic representation of the union of subspaces $\Sigma_{4\theta p_0}$. For each set $\boldsymbol{\tau}$ of at most $4\theta p_0$ shifts, we have a subspace $\mathcal{S}_{\boldsymbol{\tau}}$. Right: $\varphi_\rho$ has good geometry near this union of subspaces.

**Geometry of $\varphi_\rho$ over a union of subspaces.** Our main geometric result will show that these properties obtain on *every* subspace spanned by a few shifts of $\boldsymbol{a}_0$. Indeed, for each subset

$$\boldsymbol{\tau} \subseteq \{-p_0+1, \dots, p_0-1\}, \tag{3.1}$$

define a linear subspace

$$\mathcal{S}_{\boldsymbol{\tau}} = \left\{ \sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] \,\middle|\, \boldsymbol{\alpha}_{-p_0+1}, \dots, \boldsymbol{\alpha}_{p_0-1} \in \mathbb{R} \right\}. \tag{3.2}$$

The subspace $\mathcal{S}_{\boldsymbol{\tau}}$ is the linear span of the shifts $s_\ell[\boldsymbol{a}_0]$ indexed by $\ell$ in the set $\boldsymbol{\tau}$. Our geometric theory will show that with high probability the function $\varphi_\rho$ has no spurious local minimizers near any $\mathcal{S}_{\boldsymbol{\tau}}$ for which $\boldsymbol{\tau}$ is not too large – say, $|\boldsymbol{\tau}| \le 4\theta p_0$. Combining all of these subspaces into a single geometric object, define the union of subspaces

$$\Sigma_{4\theta p_0} = \bigcup_{|\boldsymbol{\tau}| \le 4\theta p_0} \mathcal{S}_{\boldsymbol{\tau}}. \tag{3.3}$$

Figure 6 (left) gives a schematic representation of this set. We claim:

- In the neighborhood of $\Sigma_{4\theta p_0}$, all local minimizers are near signed shifts.
- The value of $\varphi_\rho$ grows in any direction away from $\Sigma_{4\theta p_0}$.

**Main Geometric Result.** Our main result formalizes the above observations, under two key assumptions: first, that the sparsity rate $\theta$ is sufficiently small (relative to the shift coherence $\mu$ of $p_0$), and, second, the signal length $n$ is sufficiently large:

**Theorem 3.1** (Main Geometric Theorem). *Let $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ with $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$ $\mu$-shift coherent and $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$ with sparsity rate*

$$\theta \ \in \ \left[\frac{c_1}{p_0}, \frac{c_2}{p_0\sqrt{\mu} + \sqrt{p_0}}\right] \cdot \frac{1}{\log^2 p_0}. \tag{3.4}$$

*Choose $\rho(x) = \sqrt{x^2 + \delta^2}$ and set $\lambda = 0.1/\sqrt{p_0\theta}$ in $\varphi_\rho$. Then there exists $\delta > 0$ and numerical constant $c$ such that if $n \geq \mathrm{poly}(p_0)$, with high probability, every local minimizer $\bar{\boldsymbol{a}}$ of $\varphi_\rho$ over $\Sigma_{4\theta p_0}$ satisfies $\|\bar{\boldsymbol{a}} - \sigma s_\ell[\boldsymbol{a}_0]\|_2 \leq c \max\left\{\mu, p_0^{-1}\right\}$ for some signed shift $\sigma s_\ell[\boldsymbol{a}_0]$ of the true kernel. Above, $c_1, c_2 > 0$ are positive numerical constants.*

**Proof.** This follows from Theorem 4.1. ∎

The upper bound on $\theta$ in (3.4) yields the tradeoff between coherence and sparsity described in Figure 2. Simply put, when $\boldsymbol{a}_0$ is better conditioned (as a kernel), its coherence $\mu$ is smaller and $\boldsymbol{x}_0$ can be denser.

At a technical level, our proof of Theorem 3.1 shows that (i) $\varphi_\rho(\boldsymbol{a})$ is strongly convex in the vicinity of each signed shift, and that at every other point $\boldsymbol{a}$ near $\Sigma_{4\theta p_0}$, there is either (ii) a nonzero gradient or (iii) a direction of strict negative curvature; furthermore (iv) the function $\varphi_\rho$ grows away from $\Sigma_{4\theta p_0}$. Points (ii)-(iii) imply that near $\Sigma_{4\theta p_0}$ there are no "flat" saddles: every saddle point has a direction of strict negative curvature. We will leverage these properties to propose an efficient algorithm for finding a local minimizer near $\Sigma_{4\theta p_0}$. Moreover, this minimizer is close enough to a shift (here, $\|\bar{\boldsymbol{a}} - s_\ell[\boldsymbol{a}_0]\|_2 \lessapprox \mu$) for us to exactly recover $s_\ell[\boldsymbol{a}_0]$: we will give a refinement algorithm that produces $(\pm s_\ell[\boldsymbol{a}_0], \pm s_{-\ell}[\boldsymbol{x}_0])$.

### 3.2 Provable Algorithm for SaS Deconvolution

The objective function $\varphi_\rho$ has good geometric properties on (and near!) the union of subspaces $\Sigma_{4\theta p_0}$. In this section, we show how to use give an efficient method that exactly recovers $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$, up to shift symmetry. Although our geometric analysis only controls $\varphi_\rho$ near $\Sigma_{4\theta p_0}$, we will give a descent method which, with appropriate initialization $\boldsymbol{a}^{(0)}$, produces iterates $\boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(k)}, \dots$ that remain close to $\Sigma_{4\theta p_0}$ for all $k$. In short, it is easy to *start* near $\Sigma_{4\theta p_0}$ and easy to *stay* near $\Sigma_{4\theta p_0}$. After finding a local minimizer $\bar{\boldsymbol{a}}$, we refine it to produce a signed shift of $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ using alternating minimization.

The next two paragraphs give the main ideas behind the main steps of the algorithm. We then describe its components in more detail (Algorithm 1) and state our main algorithmic result (Theorem 3.2), which asserts that under appropriate conditions this method produces a signed shift of $(\boldsymbol{a}_0, \boldsymbol{x}_0)$.

**Minimization: Starting and staying near $\Sigma_{4\theta p_0}$.** Our algorithm starts with a initialization scheme which generates $\boldsymbol{a}^{(0)}$ near the union of subspaces $\Sigma_{4\theta p_0}$, which consists of linear combinations of just a few shifts of $\boldsymbol{a}_0$. How can we find a point near this union? Notice that *the data $\boldsymbol{y}$ also consists of a linear combination of just a few shifts of $\boldsymbol{a}_0$* Indeed:

$$\boldsymbol{y} \ = \ \boldsymbol{a}_0 * \boldsymbol{x}_0 \ = \sum_{\ell \in \mathrm{supp}(\boldsymbol{x}_0)} \boldsymbol{x}_{0\ell} s_\ell[\boldsymbol{a}_0]. \tag{3.5}$$

A length-$p_0$ segment of data $\boldsymbol{y}_{0,\dots,p_0-1} = [\boldsymbol{y}_0, \dots, \boldsymbol{y}_{p_0-1}]^*$ captures portions of roughly $2\theta p_0 \ll 4\theta p_0$ shifts $s_\ell[\boldsymbol{a}_0]$.

Many of these copies of $\boldsymbol{a}_0$ are truncated by the restriction to $\{0, \dots, p_0 - 1\}$. A relatively simple remedy is as follows: first, we zero-pad $\boldsymbol{y}_{0,\dots,p_0-1}$ to length $p = 3p_0 - 2$, giving

$$\left[\mathbf{0}^{p_0-1}; \boldsymbol{y}_0; \cdots; \boldsymbol{y}_{p_0-1}; \mathbf{0}^{p_0-1}\right]. \tag{3.6}$$

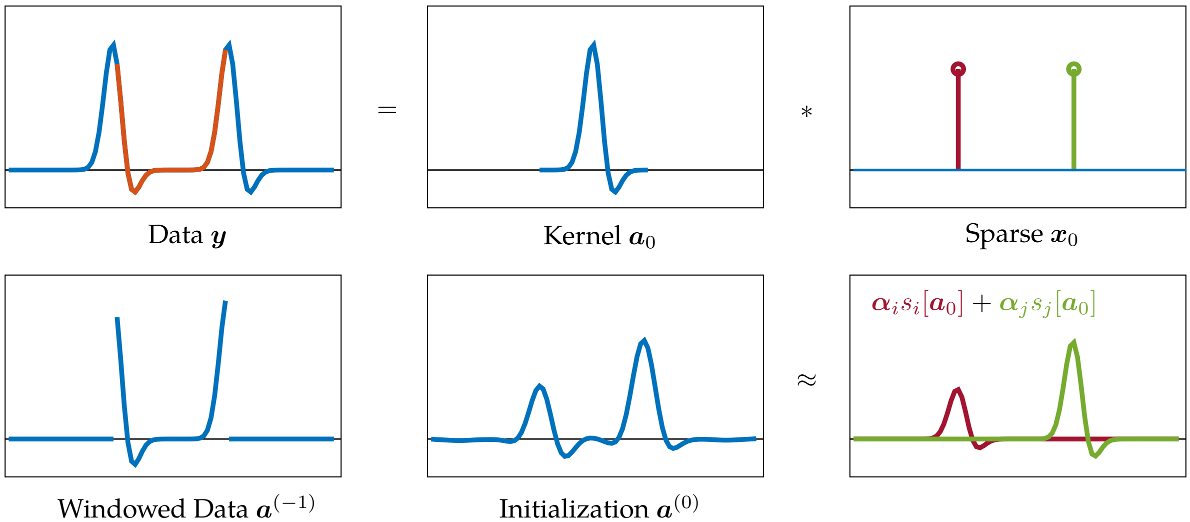


**Figure 7: Data-driven initialization:** using a piece of the observed data $\boldsymbol{y}$ to generate an initial point $\boldsymbol{a}^{(0)}$ that is close to a superposition of shifts $s_\ell[\boldsymbol{a}_0]$ of the ground truth. Top: data $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ is a superposition of shifts of the true kernel $\boldsymbol{a}_0$. Bottom: a length-$p_0$ window contains pieces of just a few shifts. Bottom middle: one step of the generalized power method approximately fills in the missing pieces, yielding a near superposition of shifts of $\boldsymbol{a}_0$ (right).

Zero padding provides enough space to accommodate any shift $s_\ell[\boldsymbol{a}_0]$ with $\ell \in \boldsymbol{\tau}$. We then perform one step of the generalized power method[8], writing

$$\boldsymbol{a}^{(0)} = -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla \varphi_{\ell^1} \left( \boldsymbol{P}_{\mathbb{S}^{p-1}} \left[ \mathbf{0}^{p_0-1}; \boldsymbol{y}_0; \cdots; \boldsymbol{y}_{p_0-1}; \mathbf{0}^{p_0-1} \right] \right), \tag{3.7}$$

where $\boldsymbol{P}_{\mathbb{S}^{p-1}}$ projects onto the sphere. The reasoning behind this construction may seem obscure. We will explain it at a more technical level in Section 5 after interpreting the gradient $\nabla \varphi_\rho$ in terms of its action on the shifts $s_\ell[\boldsymbol{a}_0]$ in Section 4. For now, we note that this operation has the effect of (approximately) filling in the missing pieces of the truncated shifts $s_\ell[\boldsymbol{a}_0]$ – see Figure 7 for an example. We will prove that with high probability $\boldsymbol{a}^{(0)}$ is indeed close to $\Sigma_{4\theta p_0}$.

The next key observation is that the function $\varphi_\rho$ grows as we move away from the subspace $\mathcal{S}_{\boldsymbol{\tau}}$ – see Figure 8. Because of this, a small-stepping descent method will not move far away from $\Sigma_{4\theta p_0}$. For concreteness, we will analyze a variant of the curvilinear search method [Gol80, GMWZ17], which moves in a linear combination of the negative gradient direction $-\boldsymbol{g}$ and a negative curvature direction $-\boldsymbol{v}$. At the $k$-th iteration, the algorithm updates $\boldsymbol{a}^{(k+1)}$ as

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}} \left[ \boldsymbol{a}^{(k)} - t\boldsymbol{g}^{(k)} - t^2 \boldsymbol{v}^{(k)} \right] \tag{3.8}$$

with appropriately chosen step size $t$. The inclusion of a negative curvature direction allows the method to avoid stagnation near saddle points. Indeed, we will prove that starting from initialization $\boldsymbol{a}^{(0)}$, this method produces a sequence $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \ldots$ which efficiently converges to a local minimizer $\bar{\boldsymbol{a}}$ that is near some signed shift $\pm s_\ell[\boldsymbol{a}_0]$ of the ground truth.

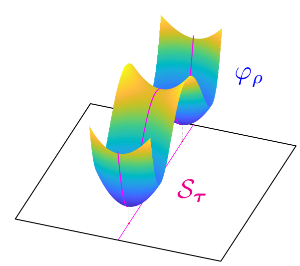


**Figure 8: Growth of $\varphi_\rho$ away from $\mathcal{S}_{\boldsymbol{\tau}}$.** Because $\varphi_\rho$ grows away from $\mathcal{S}_{\boldsymbol{\tau}}$, small-stepping descent methods stay near $\mathcal{S}_{\boldsymbol{\tau}}$.

[8] The power method for minimizing a quadratic form $\xi(\boldsymbol{a}) = \frac{1}{2}\boldsymbol{a}^* \boldsymbol{M} \boldsymbol{a}$ over the sphere consists of the iteration $\boldsymbol{a} \mapsto -\boldsymbol{P}_{\mathbb{S}^{p-1}} \boldsymbol{M} \boldsymbol{a}$. Notice that in this mapping, $-\boldsymbol{M}\boldsymbol{a} = -\nabla \xi(\boldsymbol{a})$. The generalized power method, for minimizing a function $\varphi$ over the sphere consists of repeatedly projecting $-\nabla \varphi$ onto the sphere, giving the iteration $\boldsymbol{a} \mapsto -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla \varphi(\boldsymbol{a})$. (3.7) can be interpreted as one step of the generalized power method for the objective function $\varphi_\rho$.

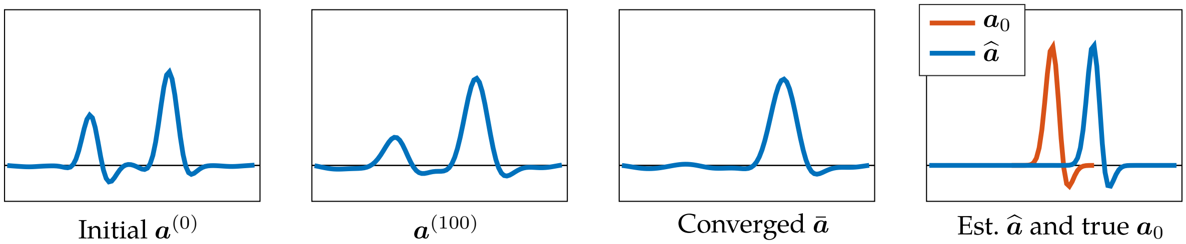


**Figure 9: Local minimization and refinement.** Left: data-driven initialization $\boldsymbol{a}^{(0)}$ consisting of a near-superposition of two shifts. Middle: minimizing $\varphi_\rho$ produces a near shift of $\boldsymbol{a}_0$. Right: rounded solution $\widehat{\boldsymbol{a}}$ using the Lasso. $\widehat{\boldsymbol{a}}$ is very close to a shift of $\boldsymbol{a}_0$.

**Refinement: Rounding a near-solution with homotopy alternating minimization.** The second step of our algorithm *rounds* the local minimizer $\bar{\boldsymbol{a}} \approx \sigma s_\ell[\boldsymbol{a}_0]$ to produce an exact solution $\widehat{\boldsymbol{a}} = \sigma s_\ell[\boldsymbol{a}_0]$. As a byproduct, it also exactly recovers the corresponding signed shift of the true sparse signal, $\widehat{\boldsymbol{x}} = \sigma s_{-\ell}[\boldsymbol{x}_0]$.

Our rounding algorithm is an alternating minimization scheme, which alternates between minimizing the Lasso cost over $\boldsymbol{a}$ with $\boldsymbol{x}$ fixed, and minimizing the Lasso cost over $\boldsymbol{x}$ with $\boldsymbol{a}$ fixed. We make two modifications to this basic idea, both of which are important for obtaining exact recovery. First, unlike the standard Lasso cost, which penalizes all of the entries of $\boldsymbol{x}$, we maintain a running estimate $I^{(k)}$ of the support of $\boldsymbol{x}_0$, and only penalize those entries that are not in $I^{(k)}$:

$$\tfrac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x} - \boldsymbol{y}\right\|_2^2 + \lambda \sum_{i \notin I^{(k)}} |\boldsymbol{x}_i| . \tag{3.9}$$

This can be viewed as an extreme form of *reweighting* [CWB08]. Second, our algorithm gradually decreases penalty variable $\lambda$ to $0$, so that eventually

$$\widehat{\boldsymbol{a}} * \widehat{\boldsymbol{x}} \approx \boldsymbol{y}. \tag{3.10}$$

This can be viewed as a *homotopy* or *continuation* method [OPT00, EHJ+04]. For concreteness, at $k$-th iteration the algorithm reads:

$$\begin{aligned}
\text{Update } \boldsymbol{x}\text{:} \qquad & \boldsymbol{x}^{(k+1)} \leftarrow \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)} \sum_{i \notin I^{(k)}} |\boldsymbol{x}_i| , && (3.11)\\
\text{Update } \boldsymbol{a}\text{:} \qquad & \boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\operatorname*{argmin}_{\boldsymbol{a}} \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2\right], && (3.12)\\
\text{Update } \lambda \text{ and } I\text{:} \qquad & \lambda^{(k+1)} \leftarrow \tfrac{1}{2}\lambda^{(k)}, \qquad I^{(k+1)} \leftarrow \operatorname{supp}\left(\boldsymbol{x}^{(k+1)}\right). && (3.13)
\end{aligned}$$

We prove that the iterates produced by this sequence of operations converge to the ground truth at a linear rate, as long as the initializer $\bar{\boldsymbol{a}}$ is sufficiently nearby.

**Algorithm and Main Algorithmic Result.** Our overall algorithm is summarized as Algorithm 1. Figure 9 illustrates the main steps of this algorithm. Our main algorithmic result states that under closely related hypotheses as above, Algorithm 1 produces a signed shift of the ground truth $(\boldsymbol{a}_0, \boldsymbol{x}_0)$:

**Algorithm 1** Short and Sparse Deconvolution

**Input:** Observation $\boldsymbol{y}$, motif length $p_0$, sparsity $\theta$, shift-coherence $\mu$, and curvature threshold $-\eta_v$.

<u>**Minimization**</u>:

Set $\boldsymbol{a}^{(0)} \leftarrow -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla \varphi_\rho \left( \boldsymbol{P}_{\mathbb{S}^{p-1}} \left[ \boldsymbol{0}^{p_0-1}; \boldsymbol{y}_0; \cdots; \boldsymbol{y}_{p_0-1}; \boldsymbol{0}^{p_0-1} \right] \right)$.

Set $\lambda = 0.1/\sqrt{p_0\theta}$ [9] and $\delta > 0$ in $\varphi_\rho$. For $k = 1, 2, \ldots, K_1$, let

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}[\boldsymbol{a}^{(k)} - t\boldsymbol{g}^{(k)} - t^2\boldsymbol{v}^{(k)}] \tag{3.14}$$

where $\boldsymbol{g}^{(k)}$ is the Riemannian gradient; $\boldsymbol{v}^{(k)}$ is the eigenvector of smallest Riemannian Hessian eigenvalue if less then $-\eta_v$ with $\langle \boldsymbol{v}^{(k)}, \boldsymbol{g}^{(k)} \rangle \geq 0$, otherwise let $\boldsymbol{v}^{(k)} = \boldsymbol{0}$; and $t \in (0, 0.1/n\theta]$ satisfies

$$\varphi_\rho(\boldsymbol{a}^{(k+1)}) < \varphi_\rho(\boldsymbol{a}^{(k)}) - \tfrac{1}{2}t\|\boldsymbol{g}^{(k)}\|_2^2 - \tfrac{1}{4}t^4\eta_v\|\boldsymbol{v}^{(k)}\|_2^2 \tag{3.15}$$

to obtain a near local minimizer $\bar{\boldsymbol{a}} \leftarrow \boldsymbol{a}^{(K_1)}$.

<u>**Refinement**</u>:

Set $\boldsymbol{a}^{(0)} \leftarrow \bar{\boldsymbol{a}},\ \lambda^{(0)} \leftarrow 10(p\theta + \log n)(\mu + 1/p)$ and $I^{(0)} \leftarrow \mathcal{S}_{\lambda^{(0)}}\left[\mathrm{supp}(\check{\boldsymbol{y}} * \bar{\boldsymbol{a}}\right])$. For $k = 1, 2, \ldots, K_2$, let

$$\boldsymbol{x}^{(k+1)} \leftarrow \mathrm{argmin}_{\boldsymbol{x}}\ \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)} \textstyle\sum_{i \notin I^{(k)}} |\boldsymbol{x}_i|\,, \tag{3.16}$$

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\mathrm{argmin}_{\boldsymbol{a}}\ \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2\right], \tag{3.17}$$

$$\lambda^{(k+1)} \leftarrow \lambda^{(k)}/2, \qquad I^{(k+1)} \leftarrow \mathrm{supp}(\boldsymbol{x}^{(k+1)}), \tag{3.18}$$

to obtain $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}}) \leftarrow (\boldsymbol{a}^{(K_2)}, \boldsymbol{x}^{(K_2)})$.

**Output:** Return $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}})$.

---

**Theorem 3.2** (Main Algorithmic Theorem)**.** *Suppose $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ where $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$ is $\mu$-truncated shift coherent such that $\max_{i \neq j} \left| \left\langle \boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0], \boldsymbol{\iota}_{p_0}^* s_j[\boldsymbol{a}_0] \right\rangle \right| \leq \mu$ and $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$ with $\theta, \mu$ satisfying*

$$\theta \in \left[ \frac{c_1}{p_0}, \frac{c_2}{\left(p_0\sqrt{\mu} + \sqrt{p_0}\right)\log^2 p_0} \right], \qquad \mu \leq \frac{c_3}{\log^2 n} \tag{3.19}$$

*for some constant $c_1, c_2, c_3 > 0$. If the signal lengths $n, p_0$ satisfy $n > \mathrm{poly}(p_0)$ and $p_0 > \mathrm{polylog}(n)$, then there exist $\delta, \eta_v > 0$ such that with high probability, Algorithm 1 produces $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}})$ that are equal to the ground truth up to signed shift symmetry:*

$$\left\| \left(\widehat{\boldsymbol{a}},\ \widehat{\boldsymbol{x}}\right) - \sigma\left(s_\ell[\boldsymbol{a}_0],\ s_{-\ell}[\boldsymbol{x}_0]\right) \right\|_2 \leq \varepsilon \tag{3.20}$$

*for some $\sigma \in \{-1, 1\}$ and $\ell \in \{-p_0 + 1, \ldots, p_0 - 1\}$ if $K_1 > \mathrm{poly}(n, p_0)$ and $K_2 > \mathrm{polylog}(n, p_0, \varepsilon^{-1})$.*

**Proof.** See Theorem 5.1 and Theorem 5.2. ■

## 3.3 Relationship to the Literature

Blind deconvolution is a classical problem in signal processing [SCI75, Can76], and has been studied under a variety of hypotheses. In this section, we first discuss the relationship between our results and the existing literature on the short-and-sparse version of this problem, and then briefly discuss other deconvolution variants in the theoretical literature.

**Applications of SaS Deconvolution.** The short-and-sparse model arises in a number of applications. One class of applications involves finding basic motifs (repeated patterns) in datasets. This *motif discovery* problem arises in extracellular spike sorting [Lew98, ETS11] and calcium imaging [PSG+16], where the observed signal exhibits repetitive *short* neuron excitation patterns occurring *sparsely* across time and/or space. Similarly, electron microscopy images [CLC+17] arising in study of nanomaterials often exhibit repeated motifs.

[9] In practice, we suggest setting $\lambda = c_\lambda/\sqrt{p_0\theta}$ with $c_\lambda \in [0.5, 0.8]$.

Another significant application of SaS deconvolution is *image deblurring*. Typically, the blur kernel is small relative to the image size (*short*) [AD88, YK96, Car01, LFDF07, LWDF11]. In natural image deblurring, the target image is often assumed to have relatively few sharp edges [FSH+06, JSK08, LWDF11], and hence have *sparse* derivatives. In scientific image deblurring, e.g., in astronomy [Lan92, HHSS09, BDH+13] and geophysics [KT98], the target image is often sparse, either in the spatial or wavelet domains, again leading to variants of the SaS model. The literature on blind image deconvolution is large; see, e.g., [KH96, CE16] for surveys.

Variants of the SaS deconvolution problem arise in many other areas of engineering as well. Examples include *blind equalization* in comunications [Sat75, SW90, JSE+98], *dereverberation* in sound engineering [MK88, NG10] and image *super-resolution* [BK02, SGG+09, YWHM10].

**Algorithmic theory for SaS deconvolution.** These applications have motivated a great deal of algorithmic work on variants of the SaS problem [LB87, BPSW95, BS95, KH96, MC99, CE16, WJPH17]. In contrast, relatively little theory is available to explain when and why algorithms succeed. Our algorithm minimizes $\varphi_\rho$ as an approximation to the Lasso cost over the sphere. Our formulation and results have strong precedent in the literature. Lasso-like objective functions have been widely used in image deblurring [YK96, CW98, FSH+06, LFDF07, SJA08, XJ10, DZSW11, KTF11, LWDF11, WZ14, PF14, ZLK+17]. A number of insights have been obtained into the geometry of sparse deconvolution – in particular, into the effect of various constraints on $\boldsymbol{a}$ on the presence or absence of spurious local minimizers. In image deblurring, a simplex constraint ($\boldsymbol{a} \geq \mathbf{0}$ and $\|\boldsymbol{a}\|_1 = 1$) arises naturally from the physical structure of the problem [YK96, CW98]. Perhaps surprisingly, simplex-constrained deconvolution admits trivial global minimizers, at which the recovered kernel $\boldsymbol{a}$ is a spike, rather than the target blur kernel [LWDF11, BVG13].

[WZ14] imposes the $\ell^2$ regularization on $\boldsymbol{a}$ and observes that this alternative constraint gives more reliable algorithm. [ZLK+17] studies the geometry of the simplified objective $\varphi_{\ell^1}$ over the sphere, and proves that in the dilute limit in which $\boldsymbol{x}_0$ has one nonzero entry, all strict local minima of $\varphi_{\ell^1}$ are close to signed shifts truncations of $\boldsymbol{a}_0$. By adopting a different objective function (based on $\ell^4$ maximization) over the sphere, [ZKW18] proves that on a certain region of the sphere every local minimum is near a *truncated* signed shift of $\boldsymbol{a}_0$, i.e., the restriction of $s_\ell[\boldsymbol{a}_0]$ to the window $\{0, \ldots, p_0 - 1\}$. The analysis of [ZKW18] allows the sparse sequence $\boldsymbol{x}_0$ to be denser ($\theta \sim p_0^{-2/3}$ for a generic kernel $\boldsymbol{a}_0$, as opposed to $\theta \lesssim p_0^{-3/4}$ in our result). Both [ZLK+17] and [ZKW18] guarantee *approximate* recovery of a portion of $s_\ell[\boldsymbol{a}_0]$, under complicated conditions on the kernel $\boldsymbol{a}_0$. Our core optimization problem is very similar to [ZLK+17]. However, we obtains *exact* recovery of both $\boldsymbol{a}_0$ and relatively dense $\boldsymbol{x}_0$, under the much simpler assumption of shift incoherence.

**Identifiability in SaS deconvolution.** Other aspects of the SaS problem have been studied theoretically. One basic question is under what circumstances the problem is identifiable, up to the scaled shift ambiguity. [CM15] shows that the problem ill-posed for worst case $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ – in particular, for certain support patterns in which $\boldsymbol{x}_0$ does not have any isolated nonzero entries. This demonstrates that *some* modeling assumptions on the support of the sparse term are needed. At the same time, this worst case structure is unlikely to occur, either under the Bernoulli model, or in practical deconvolution problems.

**Other low dimensional deconvolution models.** Motivated by a variety of applications, many low-dimensional deconvolution models have been studied in the theoretical literature. In communication applications, the signals $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ either live in known low-dimensional subspaces, or are sparse in some known dictionary [ARR14, LLB16, Chi16, LS15, LLB17, LS17, KK17]. These theoretical works assume that the subspace / dictionary are chosen at random. This low-dimensional deconvolution model does not exhibit the signed shift ambiguity; nonconvex formlations for this model exhibit a different structure from that studied here. In fact, the variant in which both signals belong to known subspaces can be solved by convex relaxation [ARR14]. The SaS model does not appear to be amenable to convexification, and exhibits a more complicated nonconvex geometry, due to the shift ambiguity. The main motivation for tackling this model lies in the aforementioned applications in imaging and data analysis.

[WC16, LB18] study the related *multi-instance* sparse blind deconvolution problem (MISBD), where there are $K$ observations $\boldsymbol{y}_i = \boldsymbol{a}_0 * \boldsymbol{x}_i$ consisting of multiple convolutions $i = 1, \dots, K$ of a kernel $\boldsymbol{a}_0$ and different sparse vectors $\boldsymbol{x}_i$. Both works develop provable algorithms. There are several key differences with our work. First, both the proposed algorithms and their analysis require the kernel to be invertible. Second, despite the apparent similarity between the SaS model and MISBD, these problems are not equivalent. It might seem possible to reduce SaS to MISBD by dividing the single observation $\boldsymbol{y}$ into $K$ pieces; this apparent reduction fails due to boundary effects.

### 3.4 Notations

**Vectors and indices.** All vectors/matrices are written in bold font $\boldsymbol{a}/\boldsymbol{A}$; indexed values are written as $\boldsymbol{a}_i$, $\boldsymbol{A}_{ij}$. Zeros or ones vectors are defined as $\mathbf{0}$ or $\mathbf{1}$, and $i$-th canonical basis vector defined as $\boldsymbol{e}_i$. The indices for vectors/matrices all start from 0 and is taking modulo-$n$, thus a vector of length $n$ should has its indices labeled as $\{0, 1, \dots, n-1\}$. We write $[n] = \{0, \dots, n-1\}$. We often use captial italic symbols $I, J$ for subsets of $[n]$. We abuse notation slightly and write $[-p] = \{n-p+1, \dots, n-1, 0\}$ and $[\pm p] = \{n-p+1, \dots, n-1, 0, 1, \dots, p-1\}$. Index sets can be labels for vectors; $\boldsymbol{a}_I \in \mathbb{R}^{|I|}$ denotes the restriction of the vector $\boldsymbol{a}$ to coordinates $I$. Also, we use check symbol for reversal operator on index set $\check{I} = -I$ and vectors $\check{\boldsymbol{a}}_i = \boldsymbol{a}_{-i}$.

**Operators.** We let $\boldsymbol{P}_C$ denote the projection operator associated with a compact set $C$. The zero-filling operator $\boldsymbol{\iota}_I : \mathbb{R}^{|I|} \to \mathbb{R}^n$ injects the input vector to higher dimensional Euclidean space, via $(\boldsymbol{\iota}_I \boldsymbol{x})_i = \boldsymbol{x}_{I^{-1}(i)}$ for $i \in I$ and 0 otherwise. Its adjoint operator $\boldsymbol{\iota}_I^*$ can be understood as subset selection operator which picks up entries of coordinates $I$. A common zero-filling operator through out this paper $\boldsymbol{\iota}$ is abbreviation of $\boldsymbol{\iota}_{[p]}$, which is often being addressed as zero-padding operator and its adjoint $\boldsymbol{\iota}^*$ as truncation operator.

**Convolution** The convolution operator are all circular with modulo-$n$: $(\boldsymbol{a} * \boldsymbol{x})_i = \sum_{j \in [n]} \boldsymbol{a}_j \boldsymbol{x}_{i-j}$, also, the convolution operator works on index set: $I * J = \operatorname{supp}(\mathbf{1}_I * \mathbf{1}_J)$. Similarly, the shift operator $s_\ell[\cdot] : \mathbb{R}^p \to \mathbb{R}^n$ is circular with modulo-n without specification: $(s_\ell[\boldsymbol{a}])_j = (\boldsymbol{\iota}_{[p]}\boldsymbol{a})_{j-\ell}$. Notice that here $\boldsymbol{a}$ can be shorter $p \le n$. Let $\boldsymbol{C}_{\boldsymbol{a}} \in \mathbb{R}^{n \times n}$ denote a circulant matrix (with modulo-$n$) for vector $\boldsymbol{a}$, whose $j$-th column is the cyclic shift of $\boldsymbol{a}$ by $j$: $\boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{e}_j = s_j[\boldsymbol{a}]$. It satisfies for any $b \in \mathbb{R}^n$,

$$\boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{b} = \boldsymbol{a} * \boldsymbol{b}. \tag{3.21}$$

The correlation between $\boldsymbol{a}$ and $\boldsymbol{b}$ can be also written in similar form of convolution operator which reverse one vector before convolution. Define two correlation matrices $\boldsymbol{C}_{\boldsymbol{a}}^*$ and $\check{\boldsymbol{C}}_{\boldsymbol{a}}$ as $\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{e}_j = s_j[\check{\boldsymbol{a}}]$ and $\check{\boldsymbol{C}}_{\boldsymbol{a}} \boldsymbol{e}_j = s_{-j}[\boldsymbol{a}]$. The two operators will satisfy

$$\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{b} = \check{\boldsymbol{a}} * \boldsymbol{b}, \quad \check{\boldsymbol{C}}_{\boldsymbol{a}} \boldsymbol{b} = \boldsymbol{a} * \check{\boldsymbol{b}}. \tag{3.22}$$

## 4 Geometry of $\varphi_\rho$ in Shift Space

Underlying our main geometric and algorithmic results is a relationship between the geometry of the function $\varphi_\rho$ and the symmetries of the deconvolution problem. In this section, we describe this relationship at a more technical level, by interpreting the gradient and hessian of the function $\varphi_\rho$ in terms of the shifts $s_\ell[\boldsymbol{a}_0]$ and stating a key lemma which asserts that a certain neighborhood of the union of subspaces $\Sigma_{4\theta p_0}$ can be decomposed into regions of negative curvature, strong gradient, and strong convexity near the target solutions $\pm s_\ell[\boldsymbol{a}_0]$.

## 4.1 Shifts and Correlations

The set $\Sigma_{4\theta p_0}$ is a union of subspaces. Any point $\boldsymbol{a}$ in one of these subspaces $\mathcal{S}_{\boldsymbol{\tau}}$ is a superposition of shifts of $\boldsymbol{a}_0$:

$$\boldsymbol{a} = \sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0]. \tag{4.1}$$

This representation can be extended to a general point $\boldsymbol{a} \in \mathbb{S}^{p-1}$ by writing

$$\boldsymbol{a} = \sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] + \sum_{\ell \notin \boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0]. \tag{4.2}$$

The vector $\boldsymbol{\alpha}$ can be viewed as the coefficients of a decomposition of $\boldsymbol{a}$ into different shifts of $\boldsymbol{a}_0$. This representation is not unique. For $\boldsymbol{a}$ close to $\mathcal{S}_{\boldsymbol{\tau}}$, we can choose a particular $\boldsymbol{\alpha}$ for which $\boldsymbol{\alpha}_{\tau^c}$ is small, a notion that we will formalize below.

For convenience, we introduce a closely related vector $\boldsymbol{\beta} \in \mathbb{R}^n$, whose entries are the inner products between $\boldsymbol{a}$ and the shifts of $\boldsymbol{a}_0$: $\boldsymbol{\beta}_\ell = \langle \boldsymbol{a}, s_\ell[\boldsymbol{a}_0] \rangle$. Since the columns of $\boldsymbol{C}_{\boldsymbol{a}_0}$ are the shifts of $\boldsymbol{a}_0$, we can write

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{a} && (4.3) \\ &= \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha} =: \boldsymbol{M} \boldsymbol{\alpha}. && (4.4) \end{aligned}$$

The matrix $\boldsymbol{M}$ is the Gram matrix of the truncated shifts $\boldsymbol{\iota}^* s_\ell[\boldsymbol{a}_0]$: $\boldsymbol{M}_{ij} = \langle \boldsymbol{\iota}^* s_i[\boldsymbol{a}_0], \boldsymbol{\iota}^* s_j[\boldsymbol{a}_0] \rangle$. When $\mu$ is small, the off-diagonal elements of $\boldsymbol{M}$ are small. In particular, on $\mathcal{S}_\tau$ we may take $\boldsymbol{\alpha}_{\tau^c} = \mathbf{0}$, and $\boldsymbol{\beta} \approx \boldsymbol{\alpha}$, in the sense that $\boldsymbol{\beta}_\tau \approx \boldsymbol{\alpha}_\tau$ and the entries of $\boldsymbol{\beta}_{\tau^c}$ are small. For detailed elaboration, see Appendix B.

## 4.2 Shifts and the Calculus of $\varphi_{\ell^1}$

Our main geometric claims pertain to the function $\varphi_\rho$, which is based on a smooth sparsity surrogate $\rho(\cdot) \approx \|\cdot\|_1$. In this section, we sketch the main ideas of the proof as if $\rho(\cdot) = \|\cdot\|_1$, by relating the geometry of the function $\varphi_{\ell^1}$ to the vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ introduced above. Working with $\varphi_{\ell^1}$ simplifies the exposition; it is also faithful to the structure of our proof, which relates the derivatives of the smooth function $\varphi_\rho$ to similar quantities associated with the nonsmooth function $\varphi_{\ell^1}$.

The function $\varphi_{\ell^1}$ has a relatively simple closed form:

$$\varphi_{\ell^1}(\boldsymbol{a}) = -\tfrac{1}{2} \left\| \mathcal{S}_\lambda \left[ \breve{\boldsymbol{y}} * \boldsymbol{a} \right] \right\|_2^2. \tag{4.5}$$

Here, $\mathcal{S}_\lambda$ is the *soft thresholding operator*, which is defined for scalars $t$ as $\mathcal{S}_\lambda[t] = \operatorname{sign}(t) \max\{|t| - \lambda, 0\}$, and is extended to vectors by applying it elementwise. The operator $\mathcal{S}_\lambda[\boldsymbol{x}]$ shrinks the elements of $\boldsymbol{x}$ towards zero. Small elements become identically zero, resulting in a sparse vector.

## Gradient: Sparsifying the Correlations $\boldsymbol{\beta}$

**Gradient over Euclidean space.** Our goal is to understand the local minimizers of the function $\varphi_{\ell^1}$ over the sphere. The function $\varphi_{\ell^1}$ is differentiable. Clearly, any point $\boldsymbol{a}$ at which its gradient (over the sphere) is nonzero cannot be a local minimizer. We first give an expression for the gradient of $\varphi_{\ell^1}$ over Euclidean space $\mathbb{R}^p$, and then extend it to the sphere $\mathbb{S}^{p-1}$. Using $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ and calculus gives

$$\begin{aligned} \nabla \varphi_{\ell^1}(\boldsymbol{a}) &= -\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \breve{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda \left[ \breve{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{a} \right] \\ &= -\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \breve{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda \left[ \breve{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{\beta} \right] \\ &= -\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\chi}[\boldsymbol{\beta}], && (4.6) \end{aligned}$$

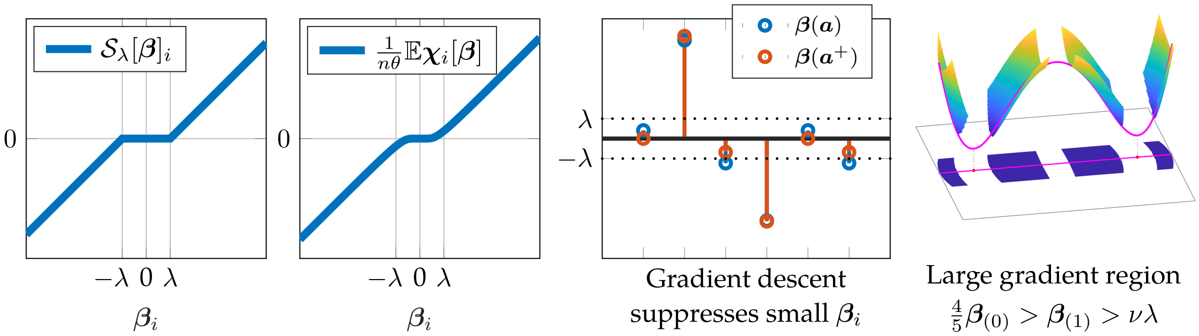


**Figure 10: Gradient Sparsifies Correlations.** Left: the soft thresholding operator $\mathcal{S}_\lambda[\boldsymbol{\beta}]$ shrinks the entries of $\boldsymbol{\beta}$ towards zero, making it sparser. Middle left: the negative gradient $-\nabla\varphi_{\ell^1}$ is a superposition of shifts $s_\ell[\boldsymbol{a}_0]$, with coefficients $\boldsymbol{\chi}_\ell[\boldsymbol{\beta}] \approx \mathcal{S}_\lambda[\boldsymbol{\beta}]_\ell$. Because of this, gradient descent sparsifies $\boldsymbol{\beta}$. Middle right: $\boldsymbol{\beta}(\boldsymbol{a})$ before, and $\boldsymbol{\beta}(\boldsymbol{a}^+)$ after, one projected gradient step $\boldsymbol{a}^+ = \boldsymbol{P}_{\mathbb{S}^{p-1}}[\boldsymbol{a} - t \cdot \mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a})]$. Notice that the small entries of $\boldsymbol{\beta}$ are shrunk towards zero. Right: the gradient $\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a})$ is large whenever it is easy to sparsify $\boldsymbol{\beta}$; in particular, when the largest entry $\boldsymbol{\beta}_{(0)} \gg \boldsymbol{\beta}_{(1)} \gg 0$.

where we have simplified the notation by introducing an operator $\boldsymbol{\chi} : \mathbb{R}^n \to \mathbb{R}^n$ as $\boldsymbol{\chi}[\boldsymbol{\beta}] = \breve{\boldsymbol{C}}_{\boldsymbol{x}_0}\mathcal{S}_\lambda\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{\beta}\right]$. This representation exhibits the (negative) gradient as a superposition of shifts of $\boldsymbol{a}_0$ with coefficients given by the entries of $\boldsymbol{\chi}[\boldsymbol{\beta}]$:

$$-\nabla\varphi_{\ell^1}(\boldsymbol{a}) = \sum_\ell \boldsymbol{\chi}[\boldsymbol{\beta}]_\ell \, s_\ell[\boldsymbol{a}_0]. \tag{4.7}$$

The operator $\boldsymbol{\chi}$ appears complicated. However, its effect is relatively simple: *when $\boldsymbol{x}_0$ is a long random vector, $\boldsymbol{\chi}[\boldsymbol{\beta}]$ acts like a soft thresholding operator on the vector $\boldsymbol{\beta}$.* That is,

$$\frac{1}{n\theta}\cdot\boldsymbol{\chi}[\boldsymbol{\beta}]_\ell \approx \begin{cases} \boldsymbol{\beta}_\ell - \lambda, & \boldsymbol{\beta}_\ell > \lambda \\ \boldsymbol{\beta}_\ell + \lambda, & \boldsymbol{\beta}_\ell < -\lambda \\ 0, & \text{otherwise} \end{cases}. \tag{4.8}$$

We show this rigorously below, in the proof of our main theorems. Here, we support this claim pictorially, by plotting the $\ell$-th entry $\boldsymbol{\chi}[\boldsymbol{\beta}]_\ell$ as $\boldsymbol{\beta}_\ell$ varies – see Figure 10 (middle left) and compare to Figure 10 (left). Because $\boldsymbol{\chi}[\boldsymbol{\beta}]$ suppresses small entries of $\boldsymbol{\beta}$, the strongest contributions to $-\nabla\varphi_{\ell^1}$ in (4.7) will come from shifts $s_\ell[\boldsymbol{a}_0]$ with large $\boldsymbol{\beta}_\ell$. *In particular, the Euclidean gradient is large whenever there is a single preferred shift $s_\ell[\boldsymbol{a}_0]$, i.e., the largest entry of $\boldsymbol{\beta}$ is significantly larger than the second largest entry.*

**Gradient over Sphere.** The (Euclidean) gradient $\nabla\varphi_{\ell^1}$ measures the slope of $\varphi_{\ell^1}$ over $\mathbb{R}^n$. We are interested in the slope of $\varphi_{\ell^1}$ over the sphere $\mathbb{S}^{p-1}$, which is measured by the Riemannian gradient

$$\begin{aligned} \mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}) &= \boldsymbol{P}_{\boldsymbol{a}^\perp}\nabla\varphi_{\ell^1}(\boldsymbol{a}) \\ &= -\boldsymbol{P}_{\boldsymbol{a}^\perp}\sum_\ell \boldsymbol{\chi}_\ell[\boldsymbol{\beta}]\, s_\ell[\boldsymbol{a}_0]. \end{aligned} \tag{4.9}$$

The Riemannian gradient simply projects the Euclidean gradient onto the tangent space $\boldsymbol{a}^\perp$ to $\mathbb{S}^{p-1}$ at $\boldsymbol{a}$. The Riemannian gradient is large whenever

(i) **Negative gradient points to one particular shift**: there is a single preferred shift $s_\ell[\boldsymbol{a}_0]$ so that the Euclidean gradient is large *and*

(ii) $\boldsymbol{a}$ **is not too close to any shift**: it is possible to move in the tangent space in the direction of this shift.[10] Since the tangent space consists of those vectors orthogonal to $\boldsymbol{a}$, this is possible whenever $s_\ell[\boldsymbol{a}_0]$ is not too aligned with $\boldsymbol{a}$, i.e., $\boldsymbol{a}$ is not too close to $s_\ell[\boldsymbol{a}_0]$.

Our technical lemma quantifies this situation in terms of the ordered entries of $\boldsymbol{\beta}$. Write $|\boldsymbol{\beta}_{(0)}| \geq |\boldsymbol{\beta}_{(1)}| \geq \dots$, with corresponding shifts $s_{(0)}[\boldsymbol{a}_0], s_{(1)}[\boldsymbol{a}_0], \dots$. There is a strong gradient whenever $|\boldsymbol{\beta}_{(0)}|$ is significantly larger than $|\boldsymbol{\beta}_{(1)}|$ and $|\boldsymbol{\beta}_{(1)}|$ is not too small compared to $\lambda$: in particular, when $\frac{4}{5}|\boldsymbol{\beta}_{(0)}| > |\boldsymbol{\beta}_{(1)}| > \frac{\lambda}{4\log^2\theta^{-1}}$. In this situation, gradient descent drives $\boldsymbol{a}$ toward $s_{(0)}[\boldsymbol{a}_0]$, reducing $|\boldsymbol{\beta}_{(1)}|, \dots$, and making the vector $\boldsymbol{\beta}$ sparser. We establish the technical claim that the (Euclidean) gradient of $\varphi_{\ell^1}$ sparsifies vectors in shift space in Appendix C.

## Hessian: Negative Curvature Breaks Symmetry

When there is no single preferred shift, i.e., when $|\boldsymbol{\beta}_{(1)}|$ is close to $|\boldsymbol{\beta}_{(0)}|$, the gradient can be small. Similarly, when $\boldsymbol{a}$ is very close to $\pm s_{(0)}[\boldsymbol{a}_0]$, the gradient can be small. In either of these situations, we need to study the curvature of the function $\varphi$ to determine whether there are local minimizers.

**Nonsmoothness.** Strictly speaking, the function $\varphi_{\ell^1}$ is not twice differentiable, due to the nonsmoothness of the soft thresholding operator $\mathcal{S}_\lambda[t]$ at $t = \pm\lambda$. Indeed, $\varphi_{\ell^1}$ is nonsmooth at any point $\boldsymbol{a}$ for which some entry of $\breve{\boldsymbol{y}} * \boldsymbol{a}$ has magnitude $\lambda$. At other points $\boldsymbol{a}$, $\varphi_{\ell^1}$ is twice differentiable, and its Hessian is given by

$$\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a}) = -\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\breve{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{P}_I\breve{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{\iota}, \tag{4.10}$$

with $I = \operatorname{supp}\left(\mathcal{S}_\lambda\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\right]\right)$. We (formally) extend this expression to *every* $\boldsymbol{a} \in \mathbb{R}^n$, terming $\widetilde{\nabla}^2\varphi_{\ell^1}$ the *pseudo-Hessian* of $\varphi_{\ell^1}$. For appropriately chosen smooth sparsity surrogate $\rho$, we will see that the (true) Hessian of the smooth function $\nabla^2\varphi_\rho$ is close to $\widetilde{\nabla}^2\varphi_{\ell^1}$, and so $\widetilde{\nabla}^2\varphi_{\ell^1}$ yields useful information about the curvature of $\varphi_\rho$.

**Curvature over Euclidean Space.** As with the gradient, the Hessian is complicated, but becomes simpler when the sample size is large. The following approximation

$$\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a}) \approx -\sum_\ell s_\ell[\boldsymbol{a}_0]s_\ell[\boldsymbol{a}_0]^*\left(\frac{\partial}{\partial\boldsymbol{\beta}_\ell}\boldsymbol{\chi}_\ell[\boldsymbol{\beta}]\right) \tag{4.11}$$

can be obtained from (4.7) noting that $\frac{\partial}{\partial\boldsymbol{a}}\boldsymbol{\chi}_\ell[\boldsymbol{\beta}] = \sum_j s_j[\boldsymbol{a}_0]\frac{\partial}{\partial\boldsymbol{\beta}_j}\boldsymbol{\chi}_\ell[\boldsymbol{\beta}]$, that $\frac{\partial}{\partial\boldsymbol{\beta}_j}\boldsymbol{\chi}_\ell[\boldsymbol{\beta}] \approx 0$ for $j \neq \ell$, and that

$$\frac{1}{n\theta}\cdot\frac{\partial\boldsymbol{\chi}_\ell[\boldsymbol{\beta}]}{\partial\boldsymbol{\beta}_\ell} \approx \begin{cases} 0 & |\boldsymbol{\beta}_\ell| \ll \lambda \\ 1 & |\boldsymbol{\beta}_\ell| \gg \lambda \end{cases} \tag{4.12}$$

Again, we corroborate this approximation pictorially – see Figure 11.

From this approximation, we can see that the quadratic form $\boldsymbol{v}^*\widetilde{\nabla}^2\varphi_{\ell^1}\boldsymbol{v}$ takes on a large negative value whenever $\boldsymbol{v}$ is a shift $s_\ell[\boldsymbol{a}_0]$ corresponding to some $|\boldsymbol{\beta}_\ell| \geq \lambda$, or whenever $\boldsymbol{v}$ is a linear combination of such shifts. *In particular, if for some $j$, $|\boldsymbol{\beta}_{(0)}|, |\boldsymbol{\beta}_{(1)}|, \dots, |\boldsymbol{\beta}_{(j)}| \gg \lambda$, then $\varphi_{\ell^1}$ will exhibit negative curvature in any direction $\boldsymbol{v} \in \operatorname{span}(s_{(0)}[\boldsymbol{a}_0], s_{(1)}[\boldsymbol{a}_0], \dots, s_{(j)}[\boldsymbol{a}_0])$.*

**Curvature over the Sphere.** The (Euclidean) Hessian measures the curvature of the function $\varphi_{\ell^1}$ over $\mathbb{R}^n$. The Riemannian Hessian

$$\widetilde{\operatorname{Hess}}[\varphi_{\ell^1}](\boldsymbol{a}) = \boldsymbol{P}_{\boldsymbol{a}^\perp}\left(\underbrace{\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a})}_{\text{Curvature of }\varphi_{\ell^1}} + \underbrace{\langle -\nabla\varphi_{\ell^1}(\boldsymbol{a}), \boldsymbol{a}\rangle \cdot \boldsymbol{I}}_{\text{Curvature of the sphere}}\right)\boldsymbol{P}_{\boldsymbol{a}^\perp}. \tag{4.13}$$

[10]...so the projection of the Euclidean gradient onto the tangent space does not vanish.

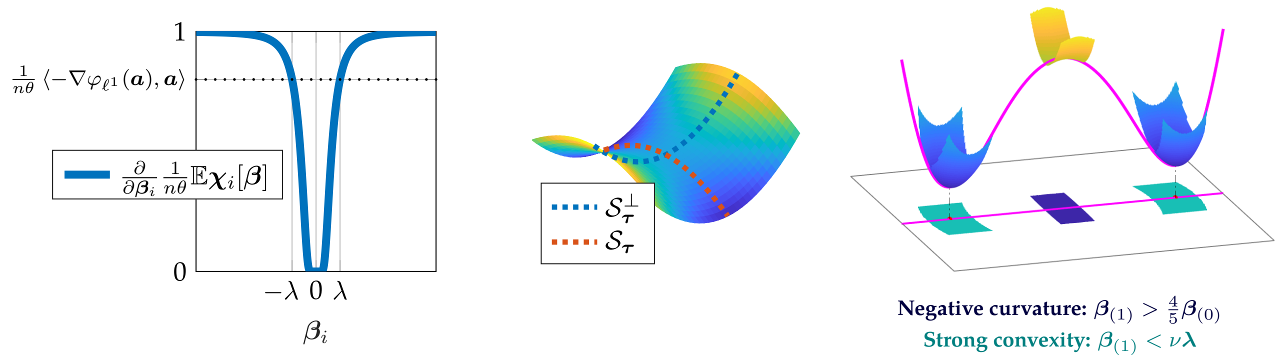


**Figure 11: Hessian Breaks Symmetry.** Left: contribution of $-s_i[\boldsymbol{a}_0]s_i[\boldsymbol{a}_0]^*$ to the Euclidean hessian. If $|\boldsymbol{\beta}_i| \gg \lambda$ the Euclidean hessian exhibits a strong negative component in the $s_i[\boldsymbol{a}_0]$ direction. The Riemannian hessian exhibits negative curvature in directions spanned by $s_i[\boldsymbol{a}_0]$ with corresponding $|\boldsymbol{\beta}_i| \gg \lambda$ and positive curvature in directions spanned by $s_i[\boldsymbol{a}_0]$ with $|\boldsymbol{\beta}_i| \ll \lambda$. Middle: this creates negative curvature along the subspace $\mathcal{S}_{\boldsymbol{\tau}}$ and positive curvature orthogonal to this subspace. Right: our analysis shows that there is always a direction of negative curvature when $\boldsymbol{\beta}_{(1)} > \frac{4}{5}\boldsymbol{\beta}_{(0)}$; conversely when $\boldsymbol{\beta}_{(1)} \ll \lambda$ there is positive curvature in every feasible direction and the function is strongly convex.

measures the curvature of $\varphi_{\ell^1}$ over the sphere. The projection $\boldsymbol{P}_{\boldsymbol{a}^\perp}$ restricts its action to directions $\boldsymbol{v} \perp \boldsymbol{a}$ that are tangent to the sphere. The additional term $\langle -\nabla\varphi_{\ell^1}(\boldsymbol{a}), \boldsymbol{a}\rangle$ accounts for the curvature of the sphere. This term is always positive. The net effect is that directions of strong negative curvature of $\varphi_{\ell^1}$ over $\mathbb{R}^n$ become directions of moderate negative curvature over the sphere. Directions of nearly zero curvature over $\mathbb{R}^n$ become directions of positive curvature over the sphere. This has three implications for the geometry of $\varphi_{\ell^1}$ over the sphere:

**(i) Negative curvature in symmetry breaking directions**: If $|\boldsymbol{\beta}_{(0)}|, |\boldsymbol{\beta}_{(1)}|, \dots, |\boldsymbol{\beta}_{(j)}| \gg \lambda$, $\varphi_{\ell^1}$ will exhibit negative curvature in any tangent direction $\boldsymbol{v} \perp \boldsymbol{a}$ which is in the linear span

$$\operatorname{span}(s_{(0)}[\boldsymbol{a}_0], s_{(1)}[\boldsymbol{a}_0], \dots, s_{(j)}[\boldsymbol{a}_0])$$

of the corresponding shifts of $\boldsymbol{a}_0$.

**(ii) Positive curvature in directions away from $\mathcal{S}_{\boldsymbol{\tau}}$**: The Euclidean Hessian quadratic form $\boldsymbol{v}^*\widetilde{\nabla}^2\varphi_{\ell^1}\boldsymbol{v}$ takes on relatively small values in directions orthogonal to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$. The Riemannian Hessian is positive in these directions, creating positive curvature orthogonal to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$.

**(iii) Strong convexity around minimizers**: Around a minimizer $s_\ell[\boldsymbol{a}_0]$, only a single entry $\boldsymbol{\beta}_\ell$ is large. Any tangent direction $\boldsymbol{v} \perp \boldsymbol{a}$ is nearly orthogonal to the subspace $\operatorname{span}(s_\ell[\boldsymbol{a}_0])$, and hence is a direction of positive (Riemannian) curvature. The objective function $\varphi_\rho$ is strongly convex around the target solutions $\pm s_\ell[\boldsymbol{a}_0]$.

Figure 11 visualizes these regions of negative and positive curvature, and the technical claim of positivity/negativity of curvature in shift space is presented in detail in Appendix D.

## 4.3 Any Local Minimizer is a Near Shift

We close this section by stating a key theorem, which makes the above discussion precise. We will show that a certain neighborhood of any subspace $\mathcal{S}_{\boldsymbol{\tau}}$ can be covered by regions of *negative curvature*, *large gradient*, and regions of *strong convexity* containing target solutions $\pm s_\ell[\boldsymbol{a}_0]$. Furthermore, at the boundary of this neighborhood, the negative gradient points back—*retracts*—toward the subspace $\mathcal{S}_{\boldsymbol{\tau}}$, due to the (directional) convexity of $\varphi_\rho$ away from the subspace.

**Widened subspace region.** To formally state the result, we need a way of measuring how close $\boldsymbol{a}$ is to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$. For technical reasons, it turns out to be convenient to do this in terms of the coefficients $\boldsymbol{\alpha}$ in the representation

$$\boldsymbol{a} = \sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] + \sum_{\ell' \in \boldsymbol{\tau}^c} \boldsymbol{\alpha}_{\ell'} s_{\ell'}[\boldsymbol{a}_0]. \tag{4.14}$$

If $\boldsymbol{a} \in \mathcal{S}_{\boldsymbol{\tau}}$, we can take $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}_{\tau^c} = \mathbf{0}$. We can view the energy $\|\boldsymbol{\alpha}_{\tau^c}\|_2$ as a measure of the distance from $\boldsymbol{a}$ to $\mathcal{S}_{\boldsymbol{\tau}}$. A technical wrinkle arises, because the representation (4.14) is not unique. We resolve this issue by choosing the $\boldsymbol{\alpha}$ that minimizes $\|\boldsymbol{\alpha}_{\tau^c}\|_2$, writing:

$$d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) = \inf \left\{ \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \, : \, \textstyle\sum_\ell \boldsymbol{\alpha}_\ell s_\ell[\boldsymbol{a}_0] = \boldsymbol{a} \right\}. \tag{4.15}$$

The distance $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}})$ is zero for $\boldsymbol{a} \in \mathcal{S}_{\boldsymbol{\tau}}$. Our analysis controls the geometric properties of $\varphi_\rho$ over the set of $\boldsymbol{a}$ for which $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}})$ is not too large. Similar to (3.3), we define an object which contains all points that are close to some $\mathcal{S}_{\boldsymbol{\tau}}$, in the above sense:

$$\Sigma^\gamma_{4\theta p_0} := \bigcup_{|\boldsymbol{\tau}| \le 4\theta p_0} \left\{ \boldsymbol{a} \, : \, d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \le \gamma \right\}. \tag{4.16}$$

The aforementioned geometric properties hold over this set:

**Theorem 4.1** (Three subregions). *Suppose that $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ where $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$ is $\mu$-shift coherent and $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$ satisfying*

$$\theta \in \left[ \frac{c'}{p_0}, \frac{c}{p_0\sqrt{\mu} + \sqrt{p_0}} \right] \cdot \frac{1}{\log^2 p_0} \tag{4.17}$$

*for some constants $c', c > 0$. Set $\lambda = 0.1/\sqrt{p_0 \theta}$ in $\varphi_\rho$ where $\rho(x) = \sqrt{x^2 + \delta^2}$. There exist numerical constants $C, c'', c''', c_1\text{-}c_4 > 0$ such that if $\delta \le \frac{c'' \lambda \theta^8}{p^2 \log^2 n}$ and $n > C p_0^5 \theta^{-2} \log p_0$, then with probability at least $1 - c'''/n$, for every $\boldsymbol{a} \in \Sigma^\gamma_{4\theta p_0}$, we have:*

- *(Negative curvature): If $\left|\boldsymbol{\beta}_{(1)}\right| \ge \nu_1 \left|\boldsymbol{\beta}_{(0)}\right|$, then*

$$\lambda_{\min}\left(\mathrm{Hess}[\varphi_\rho](\boldsymbol{a})\right) \;\le\; -c_1 n \theta \lambda; \tag{4.18}$$

- *(Large gradient): If $\nu_1 \left|\boldsymbol{\beta}_{(0)}\right| \ge \left|\boldsymbol{\beta}_{(1)}\right| \ge \nu_2(\theta)\lambda$, then*

$$\|\mathrm{grad}[\varphi_\rho](\boldsymbol{a})\|_2 \;\ge\; c_2 n \theta \tfrac{\lambda^2}{\log^2 \theta^{-1}}; \tag{4.19}$$

- *(Convex near shifts): If $\nu_2(\theta)\lambda \ge \left|\boldsymbol{\beta}_{(1)}\right|$, then*

$$\mathrm{Hess}[\varphi_\rho](\boldsymbol{a}) \;\succ\; c_3 n \theta \boldsymbol{P}_{\boldsymbol{a}^\perp}; \tag{4.20}$$

- *(Retraction to subspace): If $\frac{\gamma}{2} \le d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \le \gamma$, then for every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}$, there exists $\boldsymbol{\zeta}$ satisfying $\mathrm{grad}[\varphi_\rho](\boldsymbol{a}) = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\zeta}$, such that*

$$\langle \boldsymbol{\zeta}_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \rangle \;\ge\; c_4 \|\boldsymbol{\zeta}_{\boldsymbol{\tau}^c}\|_2 \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2; \tag{4.21}$$

- *(Local minimizers): If $\boldsymbol{a}$ is a local minimizer,*

$$\min_{\substack{\ell \in [\pm p] \\ \sigma \in \{\pm 1\}}} \|\boldsymbol{a} - \sigma\, s_\ell[\boldsymbol{a}_0]\|_2 \;\le\; \tfrac{1}{2} \max\left\{\mu, p_0^{-1}\right\}, \tag{4.22}$$

*where $\nu_1 = \frac{4}{5}$, $\nu_2(\theta) = \frac{1}{4\log^2 \theta^{-1}}$ and $\gamma = \frac{c \cdot \mathrm{poly}(\sqrt{1/\theta}, \sqrt{1/\mu})}{\log^2 \theta^{-1}} \cdot \frac{1}{\sqrt{p_0}}$.*

**Proof.** See Appendix F.5. ■

The retraction property elaborated in (4.21) implies that the negative gradient at $\boldsymbol{a}$ points in a direction that decreases $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}})$. This is a consequence of positive curvature away from $\mathcal{S}_{\boldsymbol{\tau}}$. It essentially implies that the gradient is monotone in $\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}$ space: choose any $\underline{\boldsymbol{a}} \in \mathcal{S}_{\boldsymbol{\tau}} \cap \mathbb{S}^{p-1}$, write $\underline{\boldsymbol{\alpha}}$ to be its coefficient, and let $\underline{\boldsymbol{\zeta}}$ be the coefficient of $\mathrm{grad}[\varphi_\rho](\underline{\boldsymbol{a}})$. Then $\underline{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c} = \mathbf{0}$, $\underline{\boldsymbol{\zeta}}_{\boldsymbol{\tau}^c} \approx \mathbf{0}$ and

$$\langle \boldsymbol{\zeta}_{\boldsymbol{\tau}^c} - \underline{\boldsymbol{\zeta}}_{\boldsymbol{\tau}^c},\, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} - \underline{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c} \rangle \approx \langle \boldsymbol{\zeta}_{\boldsymbol{\tau}^c} - \mathbf{0},\, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} - \mathbf{0} \rangle = \langle \boldsymbol{\zeta}_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \rangle > 0.$$

Our main geometric claim in Theorem 3.1 is a direct consequence of Theorem 4.1. Moreover, it suggests that as long as we can minimize $\varphi_\rho$ within the region $\Sigma^{\gamma}_{4\theta p_0}$, we will solve the SaS deconvolution problem.

# 5 Provable Algorithm

In light of Theorem 4.1, in this section we introduce a two-part algorithm Algorithm 1, which first applies the curvilinear descent method to find a local minimum of $\varphi_\rho$ within $\Sigma^{\gamma}_{4\theta p_0}$, followed by refinement algorithm that uses alternating minimization to exactly recover the ground truth. This algorithm exactly solves SaS deconvolution problem.

## 5.1 Minimization

There are three major issues in finding a local minimizer within $\Sigma^{\gamma}_{4\theta p_0}$. We want …

**(i) Initialization.** the initializer $\boldsymbol{a}^{(0)}$ to reside within $\Sigma^{\gamma}_{4\theta p_0}$,

**(ii) Negative curvature.** the method to avoid stagnating near the saddle points of $\varphi_\rho$,

**(iii) No exit.** the descent method to remain inside $\Sigma^{\gamma}_{4\theta p_0}$.

In the following paragraphs, we describe how our proposed algorithm achieves the above desiderata.

**Initialization within $\Sigma^{\gamma}_{4\theta p_0}$.** Our data-driven initialization scheme produces $\boldsymbol{a}^{(0)}$, where

$$\begin{aligned}
\boldsymbol{a}^{(0)} &= -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla\varphi_\rho \left( \boldsymbol{P}_{\mathbb{S}^{p-1}} \left[ \mathbf{0}^{p_0-1}; \boldsymbol{y}_0; \cdots; \boldsymbol{y}_{p_0-1}; \mathbf{0}^{p_0-1} \right] \right) \\
&= -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla\varphi_\rho \boldsymbol{P}_{\mathbb{S}^{p-1}} \left[ \boldsymbol{P}_{[p_0]}(\boldsymbol{a}_0 * \boldsymbol{x}_0) \right], \\
&\approx -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla\varphi_\rho \left[ \boldsymbol{P}_{[p_0]}(\boldsymbol{a}_0 * \widetilde{\boldsymbol{x}}_0) \right],
\end{aligned}$$

is the normalized gradient vector from a chunk of data $\boldsymbol{a}^{(-1)} := \boldsymbol{P}_{[p_0]}(\boldsymbol{a}_0 * \widetilde{\boldsymbol{x}}_0)$ with $\widetilde{\boldsymbol{x}}_0$ a normalized Bernoulli-Gaussian random vector of length $2p_0 - 1$. Since $\nabla\varphi_\rho \approx \nabla\varphi_{\ell^1}$, expand the gradient $\nabla\varphi_{\ell^1}$ and rewrite the gradient $\nabla_{\ell^1}(\boldsymbol{a}^{(-1)})$ in shift space, we get

$$\begin{aligned}
-\nabla\varphi_{\rho^1}(\boldsymbol{a}^{(-1)}) &\approx \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \breve{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda \left[ \breve{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{C}^*_{\boldsymbol{a}_0} \boldsymbol{P}_{[p_0]}(\boldsymbol{a}_0 * \widetilde{\boldsymbol{x}}_0) \right] \\
&= \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\chi} \left[ \boldsymbol{C}^*_{\boldsymbol{a}_0} \boldsymbol{P}_{[p_0]} \boldsymbol{C}_{\boldsymbol{a}_0} \widetilde{\boldsymbol{x}}_0 \right] \\
&\approx \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\chi} \left[ \widetilde{\boldsymbol{x}}_0 \right] \\
&\approx n\theta \cdot \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \mathcal{S}_\lambda \left[ \widetilde{\boldsymbol{x}}_0 \right],
\end{aligned}$$

where the approximation in the third equation is accurate if the truncated shifts are incoherent

$$\max_{i \neq j} \left| \langle \boldsymbol{\iota}^*_{p_0} s_i[\boldsymbol{a}_0], \boldsymbol{\iota}^*_{p_0} s_j[\boldsymbol{a}_0] \rangle \right| \le \mu \ll 1. \tag{5.1}$$

With this simple approximation, it comes clear that the coefficients (in shift space) of initializer $\boldsymbol{a}^{(0)}$,

$$\boldsymbol{a}^{(0)} \approx \boldsymbol{P}_{\mathbb{S}^{p-1}} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \mathcal{S}_\lambda \left[ \widetilde{\boldsymbol{x}}_0 \right], \tag{5.2}$$

approximate $\mathcal{S}_\lambda \left[ \widetilde{\boldsymbol{x}}_0 \right]$, which resides near the subspace $\mathcal{S}_{\boldsymbol{\tau}}$, in which $\boldsymbol{\tau}$ contains the nonzero entries of $\widetilde{\boldsymbol{x}}_0$ on $\{-p_0+1, \dots, p_0-1\}$. With high probability, the number of non-zero entries is $|\boldsymbol{\tau}| \lessapprox 4\theta p_0$, we therefore conclude that our initializer $\boldsymbol{a}^{(0)}$ satisfies

$$\boldsymbol{a}^{(0)} \in \Sigma^{\gamma}_{4\theta p_0}. \tag{5.3}$$

Furthermore, since $\widetilde{\boldsymbol{x}}_0$ is normalized, the largest magnitude for entries of $|\widetilde{\boldsymbol{x}}_0|$ is likely to be around $1/\sqrt{2p_0\theta}$. To ensure that $\mathcal{S}_\lambda \left[ \widetilde{\boldsymbol{x}}_0 \right]$ does not annihilate all nonzero entries of $\widetilde{\boldsymbol{x}}_0$ (otherwise our initializer $\boldsymbol{a}^{(0)}$ will become $\mathbf{0}$), the ideal $\lambda$ should be slightly less then the largest magnitude of $|\widetilde{\boldsymbol{x}}_0|$. We suggest setting $\lambda$ in $\varphi_\rho$ as

$$\lambda = \frac{c}{\sqrt{p_0\theta}}. \tag{5.4}$$

for some $c \in (0,1)$.

**Minimize $\varphi_\rho$ within $\Sigma^{\gamma}_{4\theta p_0}$.** Many methods have been proposed to optimize functions whose saddle points exhibit strict negative curvature, including the noisy gradient method [GHJY15], trust region methods [AMS09, SQW17] and curvilinear search [WY13]. Any of the above methods can be adapted to minimize $\varphi_\rho$. In this paper, we use *curvilinear method with restricted stepsize* to demonstrate how to analyze an optimization problem using the geometric properties of $\varphi_\rho$ over $\Sigma^{\gamma}_{4\theta p_0}$ – in particular, negative curvature in symmetry-breaking directions and positive curvature away from $\mathcal{S}_{\boldsymbol{\tau}}$.

Curvilinear search uses an update strategy that combines the gradient $\boldsymbol{g}$ and a direction of negative curvature $\boldsymbol{v}$, which here we choose as an eigenvector of the hessian $\boldsymbol{H}$ with smallest eigenvalue, scaled such that $\boldsymbol{v}^*\boldsymbol{g} \ge 0$. In particular, we set

$$\boldsymbol{a}^+ \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}} \left[ \boldsymbol{a} - t\boldsymbol{g} - t^2 \boldsymbol{v} \right] \tag{5.5}$$

For small $t$,

$$\varphi(\boldsymbol{a}^+) \approx \varphi(\boldsymbol{a}) + \langle \boldsymbol{g}, \boldsymbol{\xi} \rangle + \tfrac{1}{2} \boldsymbol{\xi}^* \boldsymbol{H} \boldsymbol{\xi}. \tag{5.6}$$

Since $\boldsymbol{\xi}$ converges to $\mathbf{0}$ only if $\boldsymbol{a}$ converges to the local minimizer (otherwise either gradient $\boldsymbol{g}$ is nonzero or there is a negative curvature direction $\boldsymbol{v}$), this iteration produces a local minimizer for $\varphi_\rho$, whose saddle points near any $\mathcal{S}_{\boldsymbol{\tau}}$ has negative curvature, we just need to ensure all iterates stays near some such subspace. We prove this by showing:

- When $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \le \gamma$, curvilinear steps move a small distance away from the subspace:

$$\left| d_\alpha \left( \boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}} \right) - d_\alpha \left( \boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}} \right) \right| \le \tfrac{\gamma}{2}. \tag{5.7}$$

- When $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \in \left[ \frac{\gamma}{2}, \gamma \right]$, curvilinear steps retract toward subspace:

$$d_\alpha \left( \boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}} \right) \le d_\alpha \left( \boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}} \right). \tag{5.8}$$

Together, we can prove that the iterates $\boldsymbol{a}^{(k)}$ converge to a minimizer, and

$$\forall\, k = 1, 2, \dots, \quad \boldsymbol{a}^{(k)} \in \Sigma^{\gamma}_{4\theta p_0}. \tag{5.9}$$

We conclude this section with the following theorem:

**Theorem 5.1** (Convergence of retractive curvilinear search). *Suppose signals $\boldsymbol{a}_0, \boldsymbol{x}_0$ satisfy the conditions of Theorem 4.1, $\theta > 10^3 c/p_0$ ($c > 1$), and $\boldsymbol{a}_0$ is $\mu$-truncated shift coherent $\max_{i\neq j}\left|\left\langle \boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0], \boldsymbol{\iota}_{p_0}^* s_j[\boldsymbol{a}_0]\right\rangle\right| \le \mu$. Write $\boldsymbol{g} = \mathrm{grad}[\varphi_\rho](\boldsymbol{a})$ and $\boldsymbol{H} = \mathrm{Hess}[\varphi_\rho](\boldsymbol{a})$. When the smallest eigenvalue of $\boldsymbol{H}$ is strictly smaller than $-\eta_v$ let $\boldsymbol{v}$ be the unit eigenvector of smallest eigenvalue, scaled so $\boldsymbol{v}^*\boldsymbol{g} \ge 0$; otherwise let $\boldsymbol{v} = \boldsymbol{0}$. Define a sequence $\left\{\boldsymbol{a}^{(k)}\right\}_{k\in\mathbb{N}}$ where $\boldsymbol{a}^{(0)}$ equals (3.7) and for $k = 1, 2, \ldots, K_1$:*

$$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a}^{(k)} - t\boldsymbol{g}^{(k)} - t^2\boldsymbol{v}^{(k)}\right] \tag{5.10}$$

*with largest $t \in \left(0, \frac{0.1}{n\theta}\right]$ satisfying Armijo steplength:*

$$\varphi_\rho(\boldsymbol{a}^{(k+1)}) < \varphi_\rho(\boldsymbol{a}^{(k)}) - \tfrac{1}{2}\left(t\|\boldsymbol{g}^{(k)}\|_2^2 + \tfrac{1}{2}t^4\eta_v\|\boldsymbol{v}^{(k)}\|_2^2\right), \tag{5.11}$$

*then with probability at least $1 - 1/c$, there exists some signed shift $\bar{\boldsymbol{a}} = \pm s_i[\boldsymbol{a}_0]$ where $i \in [\pm p_0]$ such that $\left\|\boldsymbol{a}^{(k)} - \bar{\boldsymbol{a}}\right\|_2 \le \mu + 1/p$ for all $k \ge K_1 = \mathrm{poly}(n, p)$. Here, $\eta_v = c'n\theta\lambda$ for some $c' < c_1$ in Theorem 4.1.*

**Proof.** See Appendix G.2. ∎

## 5.2 Local Refinement

In this section, we describe and analyze an algorithm which refines an estimate $\bar{\boldsymbol{a}} \approx \boldsymbol{a}_0$ of the kernel to exactly recover $(\boldsymbol{a}_0, \boldsymbol{x}_0)$. Set

$$\boldsymbol{a}^{(0)} \leftarrow \bar{\boldsymbol{a}}, \qquad \lambda^{(0)} \leftarrow C(p\theta + \log n)(\mu + 1/p), \qquad I^{(0)} \leftarrow \mathrm{supp}(\mathcal{S}_\lambda\left[\boldsymbol{C}_{\bar{\boldsymbol{a}}}^*\boldsymbol{y}\right]). \tag{5.12}$$

We alternatively minimize the Lasso objective with respect to $\boldsymbol{a}$ and $\boldsymbol{x}$:

$$\boldsymbol{x}^{(k+1)} \;\leftarrow\; \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)}\sum_{i\notin I^{(k)}} |\boldsymbol{x}_i|, \tag{5.13}$$

$$\boldsymbol{a}^{(k+1)} \;\leftarrow\; \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\operatorname*{argmin}_{\boldsymbol{a}} \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2\right], \tag{5.14}$$

$$\lambda^{(k+1)} \;\leftarrow\; \tfrac{1}{2}\lambda^{(k)}, \qquad I^{(k+1)} \;\leftarrow\; \mathrm{supp}\left(\boldsymbol{x}^{(k+1)}\right). \tag{5.15}$$

One departure from standard alternating minimization procedures is our use of a continuation method, which (i) decreases $\lambda$ and (ii) maintains a running estimate $I^{(k)}$ of the support set. Our analysis will show that $\boldsymbol{a}^{(k)}$ converges to one of the signed shifts of $\boldsymbol{a}_0$ at a linear rate, in the sense that

$$\min_{\sigma\in\pm1,\,\ell\in[\pm p_0]} \left\|\boldsymbol{a}^{(k)} - \sigma\cdot s_\ell[\boldsymbol{a}_0]\right\|_2 \le C'2^{-k}. \tag{5.16}$$

**Modified coherence and support density assumptions** It should be clear that exact recovery is unlikely if $\boldsymbol{x}_0$ contains many consecutive nonzero entries: in fact in this situation, even *non-blind* deconvolution fails. Therefore to obtain exact recovery it is necessary to put an upper bound on signal dimension $n$. Here, we introduce the notation $\kappa_I$ as an upper bound for number of nonzero entries of $\boldsymbol{x}_0$ in a length-$p$ window:

$$\kappa_I := 6\max\left\{\theta p, \log n\right\}, \tag{5.17}$$

where the indexing and addition should be interpreted modulo $n$. We will denote the support sets of true sparse vector $\boldsymbol{x}_0$ and recovered $\boldsymbol{x}^{(k)}$ in the intermediate $k$-th steps as

$$I = \mathrm{supp}(\boldsymbol{x}_0), \qquad I^{(k)} = \mathrm{supp}(\boldsymbol{x}^{(k)}), \tag{5.18}$$

then in the Bernoulli-Gaussian model, with high probability,

$$\max_\ell \left|I \cap ([p] + \ell)\right| \le \kappa_I. \tag{5.19}$$

The $\log n$ term reflects the fact that as $n$ becomes enormous (exponential in $p$) eventually it becomes likely that some length-$p$ window of $\boldsymbol{x}_0$ is densely occupied. In our main theorem statement, we preclude this possibility by putting an upper bound on signal length $n$ with respect to window length $p$ and shift coherence $\mu$. We will assume

$$(\mu + 1/p) \cdot \kappa_I^2 < c \tag{5.20}$$

for some numerical constant $c \in (0, 1)$.

**Alternating minimization produces $\boldsymbol{a}$ that contracts toward $\boldsymbol{a}_0$.** Recall that (B.1) in Theorem 4.1 provides that

$$\left\| \bar{\boldsymbol{a}} - \boldsymbol{a}_0 \right\|_2 \leq (\mu + 1/p)\,, \tag{5.21}$$

which is sufficiently close to $\boldsymbol{a}_0$ as long as (5.19) holds true. Here, we will elaborate this by showing a single iteration of alternating minimization algorithm (5.13)-(5.15) is a contraction mapping for $\boldsymbol{a}$ toward $\boldsymbol{a}_0$.

To this end, at $k$-th iteration, write $T = I^{(k)},\ J = I^{(k+1)}$ and $\boldsymbol{\sigma}^{(k)} = \operatorname{sign}\left(\boldsymbol{x}^{(k)}\right)$, then first observe that the solution to the reweighted Lasso problem (5.13) can be written as

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{\iota}_J \left(\boldsymbol{\iota}_J^* \boldsymbol{C}_{\boldsymbol{a}^{(k)}}^* \boldsymbol{C}_{\boldsymbol{a}^{(k)}} \boldsymbol{\iota}_J\right)^{-1} \boldsymbol{\iota}_J^* \left(\boldsymbol{C}_{\boldsymbol{a}^{(k)}}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 - \lambda^{(k)} \boldsymbol{P}_{J \setminus T} \boldsymbol{\sigma}^{(k+1)}\right), \tag{5.22}$$

and the solution to least squares problem (5.14) will be

$$\boldsymbol{a}^{(k+1)} \;=\; \left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}} \boldsymbol{\iota}\right)^{-1} \left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0\right). \tag{5.23}$$

Here, we are going to illustrate the relationship between $\boldsymbol{a}^{(k+1)} - \boldsymbol{a}_0$ and $\boldsymbol{a}^{(k)} - \boldsymbol{a}_0$ using simple approximations. First, let us assume that $\boldsymbol{a}^{(k)} \approx \boldsymbol{a}_0,\ \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \approx \boldsymbol{I}$, and $I \approx J \approx T$. Then (5.22) gives

$$\boldsymbol{x}^{(k+1)} \;\approx\; \boldsymbol{x}_0, \tag{5.24}$$

$$\begin{aligned} \left(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}_0\right) \;&\approx\; \boldsymbol{P}_I \left(\boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 - \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}^{(k)}} \boldsymbol{x}_0\right) \\ &\approx\; \boldsymbol{P}_I \left[\boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} (\boldsymbol{a}_0 - \boldsymbol{a}^{(k)})\right], \end{aligned} \tag{5.25}$$

which implies, while assuming $\boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \approx n\theta \boldsymbol{I}$, that from (5.23):

$$\begin{aligned} \left(\boldsymbol{a}^{(k+1)} - \boldsymbol{a}_0\right) \;&\approx\; (n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 - \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^* \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}} \boldsymbol{\iota} \boldsymbol{a}_0 \\ &\approx\; (n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} (\boldsymbol{x}_0 - \boldsymbol{x}^{(k+1)}) \\ &\approx\; (n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \,(\boldsymbol{a}^{(k)} - \boldsymbol{a}_0). \end{aligned} \tag{5.26}$$

Now since $\boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0} \approx n\theta\, \boldsymbol{e}_0 \boldsymbol{e}_0^*$, this suggests that $(n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}$ approximates a contraction mapping with fixed point $\boldsymbol{a}_0$, as follows:

$$\begin{aligned} (n\theta)^{-1} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \;&\approx\; \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{e}_0 \boldsymbol{e}_0^* \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \\ &\approx\; \boldsymbol{a}_0 \boldsymbol{a}_0^*. \end{aligned} \tag{5.27}$$

Hence, if we can ensure all above approximation is sufficiently and increasingly accurate as the iterate proceeds, the alternating minimization essentially is a power method which finds the leading eigenvector of matrix $\boldsymbol{a}_0 \boldsymbol{a}_0^*$—and the solution to this algorithm is apparently $\boldsymbol{a}_0$. Indeed, we prove that the iterates produced by this sequence of operations converge to the ground truth at a linear rate, as long as it is initialized sufficiently nearby:

**Theorem 5.2** (Linear rate convergence of alternating minimization). *Suppose $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ where $\boldsymbol{a}_0$ is $\mu$-shift coherent and $\boldsymbol{x}_0 \sim \mathrm{BG}(\theta)$, then there exists some constants $C, c, c_\mu$ such that if $(\mu + 1/p)\,\kappa_I^2 < c_\mu$ and $n > C\theta^{-2}p^2 \log n$, then with probability at least $1 - c/n$, for any starting point $\boldsymbol{a}^{(0)}$ and $\lambda^{(0)}$, $I^{(0)}$ such that*

$$\left\|\boldsymbol{a}^{(0)} - \boldsymbol{a}_0\right\|_2 \le \mu + 1/p, \qquad \lambda^{(0)} = 5\kappa_I(\mu + 1/p), \qquad I^{(0)} = \mathrm{supp}\left(\boldsymbol{C}^*_{\boldsymbol{a}^{(0)}}\boldsymbol{y}\right), \tag{5.28}$$

*and for $k = 1, 2, \ldots$,:*

$$\boldsymbol{x}^{(k+1)} \;\leftarrow\; \mathrm{argmin}_{\boldsymbol{x}}\, \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda^{(k)} \textstyle\sum_{i \notin I^{(k)}} |\boldsymbol{x}_i|\,, \tag{5.29}$$

$$\boldsymbol{a}^{(k+1)} \;\leftarrow\; \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\mathrm{argmin}_{\boldsymbol{a}}\, \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2\right], \tag{5.30}$$

$$\lambda^{(k+1)} \;\leftarrow\; \tfrac{1}{2}\lambda^{(k)}, \qquad I^{(k+1)} \;\leftarrow\; \mathrm{supp}\left(\boldsymbol{x}^{(k+1)}\right) \tag{5.31}$$

*then*

$$\left\|\boldsymbol{a}^{(k+1)} - \boldsymbol{a}_0\right\|_2 \;\le\; (\mu + 1/p)2^{-k} \tag{5.32}$$

*for every $k = 0, 1, 2, \ldots$.*

**Proof.** See Appendix H.3. ■

**Remark 5.3.** *The estimates $\boldsymbol{x}^{(k)}$ also converges to the ground truth $\boldsymbol{x}_0$ at a linear rate.*

# 6 Experiments

We demonstrate that the tradeoffs between the motif length $p_0$ and sparsity rate $\theta$ produce a transition region for successful SaS deconvolution under generic choices of $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$. For fixed values of $\theta \in [10^{-3}, 10^{-2}]$ and $p_0 \in [10^3, 10^4]$, we draw 50 instances of synthetic data by choosing $\boldsymbol{a}_0 \sim \mathrm{Unif}(\mathbb{S}^{p_0-1})$ and $\boldsymbol{x}_0 \in \mathbb{R}^n$ with $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ where $n = 5 \times 10^5$. Note that choosing $\boldsymbol{a}_0$ this way implies $\mu(\boldsymbol{a}_0) \approx \frac{1}{\sqrt{p_0}}$.

For each instance, we recover $\boldsymbol{a}_0$ and $\boldsymbol{x}_0$ from $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ by minimizing problem (2.5). For ease of computation, we modify Algorithm 1 by replacing curvilinear search with *accelerated Riemannian gradient descent* method (Algorithm 2), which is an adaptation of accelerated gradient descent [BT09] to the sphere. In particular, we apply momentum and increment by the Riemannian gradient via the exponential and logarithmic operators

$$\mathrm{Exp}_{\boldsymbol{a}}(\boldsymbol{u}) \;:=\; \cos(\|\boldsymbol{u}\|_2) \cdot \boldsymbol{a} + \sin(\|\boldsymbol{u}\|_2) \cdot \tfrac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2}, \tag{6.1}$$

$$\mathrm{Log}_{\boldsymbol{a}}(\boldsymbol{b}) \;:=\; \arccos(\langle \boldsymbol{a}, \boldsymbol{b}\rangle) \cdot \tfrac{\boldsymbol{P}_{\boldsymbol{a}^\perp}(\boldsymbol{b}-\boldsymbol{a})}{\left\|\boldsymbol{P}_{\boldsymbol{a}^\perp}(\boldsymbol{b}-\boldsymbol{a})\right\|_2}, \tag{6.2}$$

derived from [AMS09]. Here $\mathrm{Exp}_{\boldsymbol{a}} : \boldsymbol{a}^\perp \to \mathbb{S}^{p-1}$ takes a tangent vector of $\boldsymbol{a}$ and produces a new point on the sphere, whereas $\mathrm{Log}_{\boldsymbol{a}} : \mathbb{S}^{p-1} \to \boldsymbol{a}^\perp$ takes a point $\boldsymbol{b} \in \mathbb{S}^{p-1}$ and returns the tangent vector which points from $\boldsymbol{a}$ to $\boldsymbol{b}$.

For each recovery instance, we say the local minimizer $\boldsymbol{a}_{\min}$ generated from Algorithm 2 is sufficiently close to a solution of SaS deconvolution problem, if

$$\mathrm{success}(\boldsymbol{a}_{\min}, ; \boldsymbol{a}_0) \;:=\; \left\{ \max_\ell |\langle s_\ell[\boldsymbol{a}_0], \boldsymbol{a}_{\min}\rangle| > 0.95 \right\}. \tag{6.3}$$

The result is shown in Figure 12. Our source code can be accessed via the following address:

https://github.com/sbdsphere/sbd_experiments.git

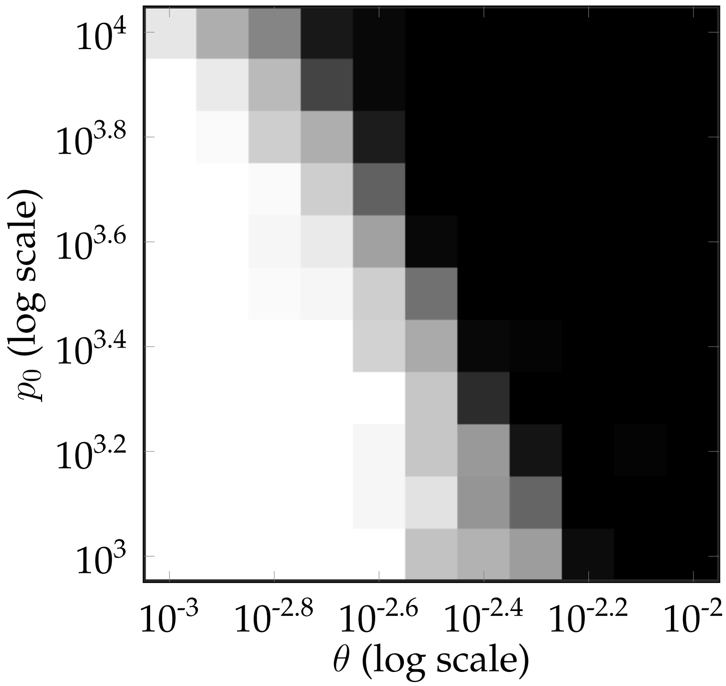


**Figure 12: Success probability of SaS deconvolution under generic $\boldsymbol{a}_0, \boldsymbol{x}_0$ with varying kernel length $p_0$, and sparsity rate $\theta$.** When sparsity rate decreases sufficiently with respect to kernel length, successful recovery becomes very likely (brighter), and vice versa (darker). A transition line is shown with slope $\frac{\log p_0}{\log \theta} \approx -2$, implying Algorithm 2 works with high probability when $\theta \lessapprox \frac{1}{\sqrt{p_0}}$ in generic case.

**Algorithm 2** SaS deconvolution with Accelerated Riemannian gradient descent

**Input:** Observation $\boldsymbol{y}$, sparsity penalty $\lambda = 0.5/\sqrt{p_0\theta}$, momentum parameter $\eta \in [0, 1)$.
Initialize $\boldsymbol{a}^{(0)} \leftarrow -\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla\varphi_\rho \left(\boldsymbol{P}_{\mathbb{S}^{p-1}} \left[\mathbf{0}^{p_0-1}; [\boldsymbol{y}_0, \cdots, \boldsymbol{y}_{p_0-1}]; \mathbf{0}^{p_0-1}\right]\right)$,
**for** $k = 1, 2, \ldots, K$ **do**
Get momentum: $\boldsymbol{w} \leftarrow \mathrm{Exp}_{\boldsymbol{a}^{(k)}}\left(\eta \cdot \mathrm{Log}_{\boldsymbol{a}^{(k-1)}}(\boldsymbol{a}^{(k)})\right)$.
Get negative gradient direction: $\boldsymbol{g} \leftarrow -\,\mathrm{grad}[\varphi_\rho](\boldsymbol{w})$.
Armijo step $\boldsymbol{a}^{(k+1)} \leftarrow \mathrm{Exp}_{\boldsymbol{w}}(t\boldsymbol{g})$, choosing $t \in (0, 1)$ s.t. $\varphi_\rho(\boldsymbol{a}^{(k+1)}) - \varphi_\rho(\boldsymbol{w}) < -t \left\|\boldsymbol{g}\right\|_2^2$.
**end for**
**Output:** Return $\boldsymbol{a}^{(K)}$.

# 7 Discussion

In this section, we close by discussing several of the most important limitations of our results, and highlighting corresponding directions for future work.

**Minimizing $\varphi_\rho$ does not accurately recover coherent kernels.** The main drawback of our proposed method is that it does not succeed when the target motif $\boldsymbol{a}_0$ has shift coherence very close to $1$. For instance, a common scenario in image blind deconvolution involves deblurring an image with a smooth, low-pass point spread function (e.g., Gaussian blur). Both our analysis and numerical experiments show that in this situation minimizing $\varphi_\rho$ does not find the generating signal pairs $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ consistently—the minimizer of $\varphi_\rho$ is often spurious and is not close to any particular shift of $\boldsymbol{a}_0$. We do not suggest minimizing $\varphi_\rho$ in this situation. On the other hand, minimizing the bilinear lasso objective $\varphi_{\mathrm{lasso}}$ over the sphere often succeeds even if the true signal pair $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ is coherent and dense.

**Relation of $\varphi_\rho$ to Bilinear Lasso.** In light of the above observations, we view the analysis of the bilinear lasso as the most important direction for future theoretical work on SaS deconvolution. The drop quadratic formulation studied here has commonalities with the biliinear lasso: both exhibit local minima at signed shifts, and both exhibit negative curvature in symmetry breaking directions. A major difference (and hence, major challenge) is that gradient methods for bilinear lasso do not retract to a union of subspaces – they retract to a more complicated, nonlinear set.

**Suboptimality in the analysis.** Finally, there are several directions in which our analysis could be improved. Our lower bounds on the length $n$ of the random vector $\boldsymbol{x}_0$ required for success are clearly suboptimal. We also suspect our sparsity-coherence tradeoff between $\mu, \theta$ (roughly, $\theta \lessapprox 1/(\sqrt{\mu} p_0)$) is suboptimal, even for the $\varphi_\rho$ objective. Articulating optimal sparsity-coherence tradeoffs for is another interesting direction for future work.

## Acknowledgement

The authors gratefully acknowledge support from NSF 1343282, NSF CCF 1527809, and NSF IIS 1546411.

## References

[AD88] GR Ayers and J Christopher Dainty. Iterative blind deconvolution method and its applications. *Optics letters*, 13(7):547–549, 1988.

[AMS09] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[ARR14] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

[BC11] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 1st edition, 2011.

[BDH+13] David Briers, Donald D Duncan, Evan R Hirst, Sean J Kirkpatrick, Marcus Larsson, Wiendelt Steenbergen, Tomas Stromberg, and Oliver B Thompson. Laser speckle contrast imaging: theoretical and practical limitations. *Journal of biomedical optics*, 18(6):066018, 2013.

[BK02] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002.

[BPSW95] PJ Bones, CR Parker, BL Satherley, and RW Watson. Deconvolution and phase retrieval with use of zero sheets. *JOSA A*, 12(9):1842–1857, 1995.

[BS95] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

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

[Can76] Michael Cannon. Blind deconvolution of spatially invariant image blurs with phase. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(1):58–63, 1976.

[Car01] Alfred S Carasso. Direct blind deconvolution. *SIAM Journal on Applied Mathematics*, 61(6):1980–2007, 2001.

[CE16] Patrizio Campisi and Karen Egiazarian. *Blind image deconvolution: theory and applications*. CRC press, 2016.

[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.

[CM15] Sunav Choudhary and Urbashi Mitra. Fundamental limits of blind deconvolution part ii: Sparsity-ambiguity trade-offs. *arXiv preprint arXiv:1503.03184*, 2015.

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

[CWB08] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[DZSW11] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. *IEEE Transactions on Image Processing*, 20(7):1838–1857, 2011.

[EHJ+04] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[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.

[FSH+06] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM transactions on graphics (TOG)*, volume 25, pages 787–794. ACM, 2006.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[GMWZ17] Donald Goldfarb, Cun Mu, John Wright, and Chaoxu Zhou. Using negative curvature in solving nonlinear programs. *Computational Optimization and Applications*, 68(3):479–502, 2017.

[Gol80] Donald Goldfarb. Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical programming*, 18(1):31–40, 1980.

[HHSS09] Stefan Harmeling, Michael Hirsch, Suvrit Sra, and Berhard Scholkopf. Online blind deconvolution for astronomical imaging. In *2009 IEEE International Conference on Computational Photography (ICCP 2009)*, pages 1–7. IEEE, 2009.

[JSE+98] Richard Johnson, Philip Schniter, Thomas J Endres, James D Behm, Donald R Brown, and Raúl A Casas. Blind equalization using the constant modulus criterion: A review. *Proceedings of the IEEE*, 86(10):1927–1950, 1998.

[JSK08] Neel Joshi, Richard Szeliski, and David J Kriegman. Psf estimation using sharp edge prediction. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[KH96] Deepa Kundur and Dimitrios Hatzinakos. Blind image deconvolution. *IEEE signal processing magazine*, 13(3):43–64, 1996.

[KK17] Michael Kech and Felix Krahmer. Optimal injectivity conditions for bilinear inverse problems with applications to identifiability of deconvolution problems. *SIAM Journal on Applied Algebra and Geometry*, 1(1):20–37, 2017.

[KT98] Kjetil F Kaaresen and Tofinn Taxt. Multichannel blind deconvolution of seismic signals. *Geophysics*, 63(6):2093–2107, 1998.

[KTF11] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 233–240. IEEE, 2011.

[Lan92] Richard G Lane. Blind deconvolution of speckle images. *JOSA A*, 9(9):1508–1514, 1992.

[LB87] RG Lane and RHT Bates. Automatic multidimensional deconvolution. *JOSA A*, 4(1):180–188, 1987.

[LB18] Yanjun Li and Yoram Bresler. Global geometry of multichannel sparse blind deconvolution on the sphere. *arXiv preprint arXiv:1404.4104*, 2018.

[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):R53–R78, 1998.

[LFDF07] Anat Levin, Rob Fergus, Fredo Durand, and William T Freeman. Deconvolution using natural image priors. *Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory*, 3, 2007.

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

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

[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, 2011.

[MC99] Joanne Markham and José-Angel Conchello. Parametric blind deconvolution: a robust method for the simultaneous estimation of image and blur. *JOSA A*, 16(10):2377–2391, 1999.

[MK88] Masato Miyoshi and Yutaka Kaneda. Inverse filtering of room acoustics. *IEEE Transactions on acoustics, speech, and signal processing*, 36(2):145–152, 1988.

[NG10] Patrick A Naylor and Nikolay D Gaubitch. *Speech dereverberation*. Springer Science & Business Media, 2010.

[OPT00] Michael R Osborne, Brett Presnell, and Berwin A Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

[PF14] Daniele Perrone and Paolo Favaro. Total variation blind deconvolution: The devil is in the details. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2916, 2014.

[PSG$^+$16] Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016.

[RV$^+$13] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[Sah07] Swapan K Saha. *Diffraction-limited imaging with large and moderate telescopes*. World Scientific, 2007.

[Sat75] Yoichi Sato. A method of self-recovering equalization for multilevel amplitude-modulation systems. *IEEE Transactions on communications*, 23(6):679–682, 1975.

[SCI75] Thomas G Stockham, Thomas M Cannon, and Robert B Ingebretsen. Blind deconvolution through digital signal processing. *Proceedings of the IEEE*, 63(4):678–692, 1975.

[SGG$^+$09] Gleb Shtengel, James A Galbraith, Catherine G Galbraith, Jennifer Lippincott-Schwartz, Jennifer M Gillette, Suliana Manley, Rachid Sougrat, Clare M Waterman, Pakorn Kanchanawong, Michael W Davidson, et al. Interferometric fluorescent super-resolution microscopy resolves 3d cellular ultrastructure. *Proceedings of the National Academy of Sciences*, 106(9):3125–3130, 2009.

[SJA08] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. In *Acm transactions on graphics (tog)*, volume 27, page 73. ACM, 2008.

[SQW17] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017.

[SW90] Ofir Shalvi and Ehud Weinstein. New criteria for blind deconvolution of nonminimum phase systems (channels). *IEEE Transactions on information theory*, 36(2):312–321, 1990.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

[WJPH17] Philipp Walk, Peter Jung, Götz E Pfander, and Babak Hassibi. Blind deconvolution with additional autocorrelations via convex programs. *arXiv preprint arXiv:1701.04890*, 2017.

[WY13] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434, 2013.

[WZ14] David Wipf and Haichao Zhang. Revisiting bayesian blind deconvolution. *The Journal of Machine Learning Research*, 15(1):3595–3634, 2014.

[XJ10] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *European conference on computer vision*, pages 157–170. Springer, 2010.

[YK96] Yu-Li You and Mostafa Kaveh. Anisotropic blind image restoration. In *Image Processing, 1996. Proceedings., International Conference on*, volume 2, pages 461–464. IEEE, 1996.

[YWHM10] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[ZKW18] Yuqian Zhang, Han-Wen Kuo, and John Wright. Structured local optima in sparse blind deconvolution. *arXiv preprint arXiv:1806.00338*, 2018.

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4902, 2017.

# A Basic bounds for Bernoulli-Gaussian vectors

In this section, we prove several lemmas pertaining to the sparse random vector $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta)$.

**Lemma A.1** (Support of $\boldsymbol{x}_0$)**.** *Let $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta)$ and $I_0 = \text{supp}(\boldsymbol{x}_0) \subseteq [n]$. Suppose $n > 10\theta^{-1}$, then for any $\varepsilon \in \left(0, \frac{1}{10}\right)$, with probability at least $1-\varepsilon$ we have*

$$||I_0| - n\theta| \le 2\sqrt{n\theta}\log\varepsilon^{-1}. \tag{A.1}$$

*And suppose $n \ge C\theta^{-2}\log p$ and $\theta$, then with probability at least $1-2/n$, we have*

$$\forall\, t \in [2p] \setminus \{0\}, \quad \tfrac{1}{2}n\theta^2 \le |I_0 \cap (I_0 + t)| \le 2n\theta^2 \tag{A.2}$$

*where $C$ is a numerical constant.*

**Proof.** Let $\boldsymbol{x}_0 = \boldsymbol{\omega}\cdot\boldsymbol{g} \sim_{\text{i.i.d.}} \text{BG}(\theta)$, notice that the support of the Bernoulli-Gaussian vector $\boldsymbol{x}_0$ is almost surely equal to the support of the Bernoulli vector $\boldsymbol{\omega}$. Applying Bernstein inequality Lemma J.4 with $(\sigma^2, R) = (1,1)$, then if $n\theta > 10$ we have

$$\mathbb{P}\left[\left|\sum_{k\in[n]}\boldsymbol{\omega}_k - n\theta\right| > 2\sqrt{n\theta}\log\varepsilon^{-1}\right] \le 2\exp\left(\frac{-4n\theta\log^2\varepsilon^{-1}}{2n\theta + 4\sqrt{n\theta}\log\varepsilon^{-1}}\right) \le \varepsilon.$$

For (A.2), let $J_t := I_0 \cap (I_0 + t)$. The cardinality of $J_t$ is an inner product between shifts of $\boldsymbol{\omega}$:

$$|J_t| = \sum_{k\in[n]}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k-t}, \tag{A.3}$$

and define two subset $J_{t1} \uplus J_{t2} = J_t$, as follows:

$$\begin{cases} J_{t1} = J_t \cap \mathcal{K}_1, & \mathcal{K}_1 := [n] \cap \{0, \dots, t-1, 2t, \dots, 3t-1, \dots\} \\ J_{t2} = J_t \cap \mathcal{K}_2, & \mathcal{K}_2 := [n] \cap \{t, \dots, 2t-1, 3t, \dots, 4t-1, \dots\} \end{cases}. \tag{A.4}$$

Here, the size of sets $\mathcal{K}_1, \mathcal{K}_2$ has two-side bounds $0.4n \le (n-2p)/2 \le |\mathcal{K}_2| \le |\mathcal{K}_1| \le (n+2p)/2 \le 0.6n$, thus the size of sets $J_{t1}, J_{t2}$ can be derived using Bernstein inequality Lemma J.4 with $n > C\theta^{-2}\log p$ as

$$\begin{aligned}
\mathbb{P}\left[\max_{t\in[2p]\setminus\{0\}}|J_{t_1}| \ge n\theta^2\right] &= \mathbb{P}\left[\max_{t\in[2p]\setminus\{0\}}\sum_{k\in\mathcal{K}_1}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k-t} \ge n\theta^2\right] \le 2p\cdot\mathbb{P}\left[\sum_{k\in\mathcal{K}_1}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k+1} \ge n\theta^2\right] \\
&\le 2p\cdot\mathbb{P}\left[\sum_{k\in\mathcal{K}_1}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k+1} - \mathbb{E}\sum_{k\in\mathcal{K}_1}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k+1} \ge n\theta^2 - 0.6n\theta^2\right] \\
&\le 4p\cdot\exp\left(\frac{-\left(0.4n\theta^2\right)^2}{2\cdot 0.6n\theta^2 + 2\cdot 0.4n\theta^2}\right) = \exp\left(\log(4p) - 0.08n\theta^2\right) \le 1/n,
\end{aligned} \tag{A.5}$$

where the last two inequalities hold with $C > 10^5$. The lower bound can also derived as follows

$$\begin{aligned}
\mathbb{P}\left[\min_{t\in[2p]\setminus\{0\}}|J_{t_1}| \le n\theta^2/4\right] &= \mathbb{P}\left[\min_{t\in[2p]\setminus\{0\}}\sum_{k\in\mathcal{K}_1}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k-t} \le n\theta^2/4\right] \le 2p\cdot\mathbb{P}\left[\sum_{k\in\mathcal{K}_1}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k+1} \le n\theta^2/4\right] \\
&\le 2p\cdot\mathbb{P}\left[\sum_{k\in\mathcal{K}_1}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k+1} - \mathbb{E}\sum_{k\in\mathcal{K}_1}\boldsymbol{\omega}_k\boldsymbol{\omega}_{k+1} \le n\theta^2/4 - 0.4n\theta^2\right] \\
&\le 4p\cdot\exp\left(\frac{-\left(0.15n\theta^2\right)^2}{2\cdot 0.6n\theta^2 + 2\cdot 0.15n\theta^2}\right) = \exp\left(\log(4p) - 0.0015n\theta^2\right) \le 1/n.
\end{aligned} \tag{A.6}$$

The bound for $|J_2|$ can derived similarly to (A.5)-(A.6). ■

**Lemma A.2** (Norms of $\boldsymbol{x}_0$). *Let $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$. If $n \geq 10\theta^{-1}$, then for any $\varepsilon \in \left(0, \frac{1}{10}\right)$, with probability at least $1-\varepsilon$,*

$$\left| \|\boldsymbol{x}_0\|_1 - \sqrt{2/\pi} n\theta \right| \leq 2\sqrt{n\theta} \log \varepsilon^{-1}, \qquad \left| \|\boldsymbol{x}_0\|_2^2 - n\theta \right| \leq 3\sqrt{n\theta} \log \varepsilon^{-1} \tag{A.7}$$

**Proof.** To bound $\|\boldsymbol{x}_0\|_1$, using Bernstein inequality with $(\sigma^2, R) = (\theta, 1)$ and with $n\theta \geq 10$ we have

$$\mathbb{P}\left[ \left| \|\boldsymbol{x}_0\|_1 - \sqrt{\frac{2}{\pi}} n\theta \right| \geq 2\sqrt{n\theta} \log \varepsilon^{-1} \right] \leq 2 \exp\left( \frac{-4n\theta \log^2 \varepsilon^{-1}}{2n\theta + 4\sqrt{n\theta} \log \varepsilon^{-1}} \right) \leq \varepsilon$$

Similarly for $\|\boldsymbol{x}_0\|_2^2$, from Gaussian moments Lemma J.2 , we know the 2-norm $\sum_{i\in[n]} \mathbb{E} |x_{0i}|^4 = 3n\theta$ and $q$-norm $\sum_{i\in[n]} \mathbb{E} |x_{0i}|^{2p} \leq (n\theta)(2q-1)!! \leq \frac{1}{2}(3n\theta)2^{q-2}q!$ for $q \geq 3$. Let $(\sigma^2, R) = (3\theta, 2)$ in Bernstein inequality form Lemma J.4, $n\theta \geq 10$ we have

$$\mathbb{P}\left[ \left| \|\boldsymbol{x}_0\|_2^2 - n\theta \right| \geq 3\sqrt{n\theta} \log \varepsilon^{-1} \right] \leq 2 \exp\left( \frac{-9n\theta \log^2 \varepsilon^{-1}}{2(3n\theta) + 12\sqrt{n\theta} \log \varepsilon^{-1}} \right) \leq \varepsilon,$$

completing the proof. ∎

**Lemma A.3** (Norms of $\boldsymbol{x}_0$ subvectors). *Let $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$ and $n > 10$, then with probability at least $1 - 3/n$, we have*

$$\max_{\substack{U=[2p]+j\\ j\in[n]}} \|\boldsymbol{P}_U \boldsymbol{x}_0\|_2^2 \leq 2p\theta + 6\left( \sqrt{p\theta} + \log n \right) \tag{A.8}$$

*and if $\boldsymbol{a}_0$ is $\mu$-shift coherent and there exists a constance $c_\mu$ such that both $\theta^2 p < c_\mu$ and $\mu p^2 \theta < c_\mu$, then*

$$\max_{\substack{U=[p]+j\\ j\in[n]}} \|\boldsymbol{P}_U [\boldsymbol{a}_0 * \boldsymbol{x}_0]\|_2^2 \leq p\theta + \log n. \tag{A.9}$$

**Proof.** Use Bernstein inequality with $(\sigma^2, R) = (3\theta, 2)$ and $t = \max\left\{\sqrt{p\theta}, \log n\right\}$, with union bound we obtain:

$$\begin{aligned} \mathbb{P}\left[ \max_{\substack{U=[2p]+j\\ j\in[n]}} \|\boldsymbol{P}_U \boldsymbol{x}_0\|_2^2 \;\geq\; 2p\theta + 6\left(\sqrt{p\theta} + \log n\right) \right] \;&\leq\; 2n \exp\left( -\frac{36\left(\sqrt{p\theta} + \log n\right)^2}{6p\theta + 12\left(\sqrt{p\theta} + \log n\right)} \right) \\ &\leq\; 2\exp\left( \log n - \frac{36t^2}{6t^2 + 12t} \right) \leq \frac{2}{n}. \end{aligned} \tag{A.10}$$

For the second inequality, first we know calculate the expectation

$$\begin{aligned} \mathbb{E} \|\boldsymbol{P}_U [\boldsymbol{a}_0 * \boldsymbol{x}_0]\|_2^2 \;&=\; \mathbb{E}\left[ \boldsymbol{x}_0^* \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{P}_U \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 \right] \\ &=\; \theta \cdot \operatorname{tr}\left( \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{P}_U \boldsymbol{C}_{\boldsymbol{a}_0} \right) \|\boldsymbol{a}_0\|_2^2 \;+\; \theta \cdot \sum_{i=1}^{p-1} \|\boldsymbol{\iota}^* s_i[\boldsymbol{a}_0]\|_2^2 \\ &=\; p\theta. \end{aligned} \tag{A.11}$$

Then apply Henson Wright inequality Lemma J.6 with $\left\|\boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{P}_U \boldsymbol{C}_{\boldsymbol{a}_0}\right\|_F^2 = \left\|\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}\right\|_F^2 \leq p\left(1 + \mu p\right)$ and also $\left\|\boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{P}_U \boldsymbol{C}_{\boldsymbol{a}_0}\right\|_2 = \|\boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}\|_2^2 = 1 + \mu p$, we can derive

$$\mathbb{P}\left[ \max_{\substack{U=[p]+j\\ j\in[n]}} \|\boldsymbol{P}_U [\boldsymbol{a}_0 * \boldsymbol{x}_0]\|_2^2 \;\geq\; p\theta + \log n \right] \;\leq\; n \exp\left( -\min\left\{ \frac{\log^2 n}{64\theta^2 p\left(1 + \mu p\right)}, \frac{\log n}{8\sqrt{2}\theta\left(1 + \mu p\right)} \right\} \right)$$

$$\leq \exp\left(\log n - \min\left\{\frac{\log^2 n}{128 c_\mu}, \frac{\log n}{32 c_\mu}\right\}\right) \leq \frac{1}{n} \tag{A.12}$$

when $c_\mu < \frac{1}{300}$. ■

**Lemma A.4** (Inner product between shifted $\boldsymbol{x}_0$)**.** *Let $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta) \in \mathbb{R}^n$. There exists a numerical constant $C$ such that if $n > C\theta^{-2}\log p$ and $p\theta\log^2\theta^{-1} > 1$, with probability at least $1 - 4/n$, the following two statements hold simultaneously:*

$$\max_{i \neq j \in [2p]} \langle s_i[\boldsymbol{x}_0], s_j[\boldsymbol{x}_0] \rangle \leq 6\sqrt{n\theta^2 \log n}; \tag{A.13}$$

*and for $\boldsymbol{x}_i = |\boldsymbol{x}_{0,i}| \in \mathbb{R}^n_+$ the vector of magnitudes of $\boldsymbol{x}_0$,*

$$\max_{i \neq j \in [2p]} \langle s_i[\boldsymbol{x}], s_j[\boldsymbol{x}] \rangle \leq 4n\theta^2. \tag{A.14}$$

**Proof.** We will start from proving (A.14). Write $\boldsymbol{x} = |\boldsymbol{g}| \circ \boldsymbol{\omega}$ where $\boldsymbol{g}$ / $\boldsymbol{\omega}$ are Gaussian/Bernoulli random vectors respectively. Let $I_0$ denote the support of $\boldsymbol{\omega}$ and $t = |j - i|$ with $0 < t < p$. Then (A.14) can be written as summation of Gaussian r.v.s. on intersection of support set between shifts:

$$\langle s_i[\boldsymbol{x}], s_j[\boldsymbol{x}] \rangle = \sum_{k \in I_0 \cap (I_0 + t)} |\boldsymbol{g}_k| \, |\boldsymbol{g}_{k-t}| \tag{A.15}$$

Define $J_t := I_0 \cap (I_0 + t) = J_{t1} \uplus J_{t2}$ same as (A.4). Notice that both $\sum_{k \in J_{t1}} |\boldsymbol{g}_k| \, |\boldsymbol{g}_{k-t}|$ and $\sum_{k \in J_{t2}} |\boldsymbol{g}_k| \, |\boldsymbol{g}_{k-t}|$ are sum of independent r.v.s.. We are left to consider the upper bound of $\sum_{j \in J_{ti}} |\boldsymbol{g}_j| \left|\boldsymbol{g}'_j\right|$ where $\boldsymbol{g}$, $\boldsymbol{g}'$ are independent Gaussian vectors. We condition on the following event

$$\mathcal{E}_J := \left\{\forall t \in [2p] \setminus \{0\}\,,\ n\theta^2/4 \leq |J_{t1}|\,, |J_{t2}| \leq n\theta^2\right\}, \tag{A.16}$$

which holds w.p. at least $1 - 2/n$ from Lemma A.1. Since $\sum_{j \in J_{t1}} |\boldsymbol{g}_j| \left|\boldsymbol{g}'_j\right| \leq \|\boldsymbol{g}_{J_{t1}}\|_2 \left\|\boldsymbol{g}'_{J_{t1}}\right\|_2$, we use Gaussian concentration Lemma J.3 and union bound to obtain

$$\begin{aligned} \mathbb{P}\left[\max_{t \in [2p] \setminus \{0\}} \sum_{j \in J_{t1}} \left|\boldsymbol{g}_j \boldsymbol{g}'_j\right| > 2\,|J_{t1}|\right] &\leq 2p \cdot \mathbb{P}\left[\|\boldsymbol{g}_{J_{t1}}\|_2 \left\|\boldsymbol{g}'_{J_{t1}}\right\|_2 - \mathbb{E}\,\|\boldsymbol{g}_{J_{t1}}\|_2 \left\|\boldsymbol{g}'_{J_{t1}}\right\|_2 > |J_{t1}|\right] \\ &\leq 4p \cdot \mathbb{P}\left[\|\boldsymbol{g}_{J_{t1}}\|_2 - \mathbb{E}\,\|\boldsymbol{g}_{J_{t1}}\|_2 > \sqrt{|J_{t1}|}/3\right] \\ &\leq 4p\exp\left(-(|J_{t1}|/9)/2\right) \leq 4p\exp\left(-n\theta^2/72\right) \leq 1/n \end{aligned} \tag{A.17}$$

where the last inequality is derived simply via assuming $n = C\theta^{-2}\log p$ for some $C > 10^4$, such that

$$\begin{aligned} C > 400 * (4C)^{1/5} &\implies C\log p > 400\log((4C)^{1/5}p) \implies C\log p > 72\log(4Cp^5) > 72\log(4Cp^2\log^3 p) \\ &\implies n\theta^2 > 72\log(p \cdot 4C\theta^{-2}\log p) = 72\log(4np). \end{aligned}$$

Likewise for sum on set $J_{t2}$, we collect all above result and conclude for every $i \neq j \in [2p]$,

$$\langle s_i[\boldsymbol{x}], s_j[\boldsymbol{x}] \rangle = \sum_{k \in J_{t1}} |\boldsymbol{g}_k| \left|\boldsymbol{g}'_{k-t}\right| + \sum_{k \in J_{t2}} |\boldsymbol{g}_k| \left|\boldsymbol{g}'_{k-t}\right| \leq 2\left(|J_{t_1}| + |J_{t_2}|\right) \leq 4n\theta^2. \tag{A.18}$$

For (A.13) similarly condition on event $\mathcal{E}_J$, using Bernstein inequality Lemma J.4 with $(\sigma^2, R) = (1, 1)$:

$$\mathbb{P}\left[\max_{t \in [2p] \setminus \{0\}} \left|\sum_{j \in J_{t1}} \boldsymbol{g}_j \boldsymbol{g}'_j\right| > 3\sqrt{n\theta^2 \log n}\right] \leq p \cdot \exp\left(\frac{-9n\theta^2 \log n}{2\,|J_{t1}| + 6\sqrt{n\theta^2 \log n}}\right) \leq p \cdot \exp\left(\frac{-9n\theta^2 \log n}{3n\theta^2}\right) \leq \frac{1}{n} \tag{A.19}$$

thus for every $i \neq j \in [2p]$,

$$|\langle s_i[\boldsymbol{x}_0], s_j[\boldsymbol{s}_0]\rangle| \leq \left|\sum_{k\in J_{t1}} \boldsymbol{g}_k \boldsymbol{g}'_{k-t}\right| + \left|\sum_{k\in J_{t2}} \boldsymbol{g}_k \boldsymbol{g}'_{k-t}\right| \leq 6\sqrt{n\theta^2 \log n}. \tag{A.20}$$

Finally, both (A.18),(A.20) holds simultaneously with probability at least

$$1 - 2/n - 1/n - 1/n = 1 - 4/n \tag{A.21}$$

■

**Lemma A.5** (Convolution of $\boldsymbol{x}_0$)**.** *Given $\boldsymbol{y} = \boldsymbol{x}_0 * \boldsymbol{a}_0$ where $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \in \mathbb{R}^n$ and $\boldsymbol{a}_0 \in \mathbb{R}^{p_0}$ is $\mu$-shift coherent. Suppose $n \geq C\theta^{-2}\log p$ for some numerical constant $C > 0$, with probability at least $1 - 7/n$, we have the following two statement simultaneously hold:*

$$\|\boldsymbol{C}_{\boldsymbol{y}}\boldsymbol{\iota}\|_2^2 \leq 3(1+\mu p)n\theta \tag{A.22}$$

*and for all $J \subseteq [n]$,*

$$\|\boldsymbol{P}_J \boldsymbol{C}_{\boldsymbol{y}}\boldsymbol{\iota}\|_2^2 \leq 14\,|J|\,(1+\mu p)\,(p\theta + \log n) \tag{A.23}$$

**Proof.** Given any $\boldsymbol{a} \in \mathbb{S}^{p-1}$, write $\boldsymbol{\beta} = \boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{\iota}\boldsymbol{a}$ where $|\boldsymbol{\beta}| \leq 2p$ . Apply $\|\boldsymbol{x}_0\|_2^2 \leq 2n\theta$ from Lemma A.2 by choosing $\varepsilon = 1/n$, also $|\langle s_i[\boldsymbol{x}_0], s_j[\boldsymbol{x}_0]\rangle| \leq 6\sqrt{n\theta^2\log n}$ from Lemma A.4 we get:

$$\begin{aligned}
\|\boldsymbol{C}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\|_2^2 = \|\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\beta}\|_2^2 &\leq \|\boldsymbol{\beta}\|_2^2\,\|\boldsymbol{x}_0\|_2^2 + \sum_{i\neq j\in[\pm p]} |\beta_i\beta_j\,\langle s_i[\boldsymbol{x}_0], s_j[\boldsymbol{x}_0]\rangle| \\
&\leq \|\boldsymbol{\beta}\|_2^2\,\|\boldsymbol{x}_0\|_2^2 + \|\boldsymbol{\beta}\|_1^2 \max_{i\neq j\in[\pm p]} |\langle s_i[\boldsymbol{x}_0], s_j[\boldsymbol{x}_0]\rangle| \\
&\leq \|\boldsymbol{\beta}\|_2^2 \cdot 2n\theta + p\,\|\boldsymbol{\beta}\|_2^2 \cdot 6\sqrt{n\theta^2\log n} \leq 3\,\|\boldsymbol{\beta}\|_2^2\,n\theta
\end{aligned}$$

where $n = C\theta^{-2}\log p$ with $C \geq 10^4$, and the statement holds with probability at least $1 - 5/n$.

For the bound of $\|\boldsymbol{P}_J\boldsymbol{C}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\|_2^2$. Simply apply Lemma A.3 and utilize norm bound of $\|\boldsymbol{\beta}\|_2^2$, with probability at least $1 - 2/n$ we have:

$$\|\boldsymbol{P}_J\boldsymbol{C}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\|_2^2 = \sum_{i\in J} |\langle s_i[\boldsymbol{x}_0], \boldsymbol{\beta}\rangle|^2 \leq |J| \max_{\substack{U=[2p]+j\\ j\in[n]}} \|\boldsymbol{P}_U\boldsymbol{x}_0\|_2^2\,\|\boldsymbol{\beta}\|_2^2 \leq |J| \cdot 14\,(p\theta + \log n) \cdot \|\boldsymbol{\beta}\|_2^2$$

Finally apply Lemma B.4 and Gershgorin disc theorem obtain

$$\|\boldsymbol{\beta}\|_2^2 = \left\|\boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{\iota}\boldsymbol{a}\right\|_2^2 \leq \left\|\boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{\iota}\right\|_2^2 = \sigma_{\max}(\boldsymbol{M}) \leq 1 + \mu p. \tag{A.24}$$

**Remark A.6.** *When $\boldsymbol{a}_0$ is a basis vector $\boldsymbol{e}_0$, the result of Lemma A.5 gives upper bound of $\|\boldsymbol{C}_{\boldsymbol{x}_0}\|_2 < 3n\theta$, whose lower bound can be derived similarly with $\|\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\|_2 \geq \frac{2}{3}n\theta$*

■

## B Vectors in shift space

In this section, we will establish a number of properties of the coefficient vectors $\boldsymbol{\alpha}$ and correlation vector $\boldsymbol{\beta}$. Generally speaking, when $\boldsymbol{a}$ is close to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$, then both vectors $\boldsymbol{\alpha}$,$\boldsymbol{\beta}$ have most of their energy concentrated on the entries $\boldsymbol{\tau}$. In this section, we derive upper bounds on $\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}$ and $\boldsymbol{\beta}_{\boldsymbol{\tau}^c}$ under various assumptions.

In particular, we will introduce a relationship between the sparsity rate $\theta$, coherence $\mu$ and size $|\boldsymbol{\tau}|$, which we term the sparsity-coherence condition. In Lemma B.2 we prove that measuring the distance from $\boldsymbol{a}$ to subspace $\mathcal{S}_{\boldsymbol{\tau}}$ in terms of $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2$ gives a seminorm. We then use this distance to characterize a region $\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ around the subspace $\mathcal{S}_{\boldsymbol{\tau}}$. Later, in Lemma B.4 we illustrate the relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\beta} = \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota}\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}$. Finally in Lemma B.5 and Corollary B.6, controls the magnitude of $\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}$ and $\boldsymbol{\beta}_{\boldsymbol{\tau}^c}$ near $\mathcal{S}_{\boldsymbol{\tau}}$.

**Definition B.1** (Sparsity-coherence condition). *Let $\boldsymbol{a}_0 \in \mathbb{S}^{p_0-1}$ with shift coherence $\mu$. We say that $(\boldsymbol{a}_0, \theta, |\boldsymbol{\tau}|)$ satisfies the sparsity-coherence condition* $\mathrm{SCC}(c_\mu)$ *with constant $c_\mu$, if*

$$\theta \in \left[\frac{1}{p}, \frac{c_\mu}{4\max\left\{|\boldsymbol{\tau}|, \sqrt{p}\right\}}\right] \cdot \frac{1}{\log^2 \theta^{-1}}, \quad \mu \cdot \max\left\{|\boldsymbol{\tau}|^2, p^2\theta^2\right\} \cdot \log^2 \theta^{-1} \leq \frac{c_\mu}{4}, \tag{B.1}$$

*where $p = 3p_0 - 2$.*

**Lemma B.2** ($d_\alpha$ is a seminorm). *For every solution subspace $\mathcal{S}_{\boldsymbol{\tau}}$, the function $d_\alpha(\cdot, \mathcal{S}_{\boldsymbol{\tau}}) : \mathbb{R}^p \to \mathbb{R}_+$ defined as*

$$d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) = \inf\left\{\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \mid \boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}\right\}. \tag{B.2}$$

*is a seminorm, and for all $\boldsymbol{a} \in \mathcal{S}_{\boldsymbol{\tau}}$, $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) = 0$.*

**Proof.** It is immediate from definition that $d(\cdot, \mathcal{S}_{\boldsymbol{\tau}})$ is nonnegative and $\mathcal{S}_{\boldsymbol{\tau}} \subseteq \{\boldsymbol{a} : d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) = 0\}$. Subadditivity can be shown from simple norm inequalities and our definition of $d_\alpha$, for all $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ we have

$$\begin{aligned} d_\alpha(\boldsymbol{a}_1 + \boldsymbol{a}_2, \mathcal{S}_{\boldsymbol{\tau}}) &= \inf\left\{\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \mid \boldsymbol{a}_1 + \boldsymbol{a}_2 = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}\right\} \\ &= \inf\left\{\|\boldsymbol{\alpha}_{1\boldsymbol{\tau}^c} + \boldsymbol{\alpha}_{2\boldsymbol{\tau}^c}\|_2 \mid \boldsymbol{a}_1 = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}_1, \quad \boldsymbol{a}_2 = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}_2\right\} \\ &\leq \inf\left\{\|\boldsymbol{\alpha}_{1\boldsymbol{\tau}^c}\|_2 + \|\boldsymbol{\alpha}_{2\boldsymbol{\tau}^c}\|_2 \mid \boldsymbol{a}_1 = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}_1, \quad \boldsymbol{a}_2 = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}_2\right\} \\ &= \inf\left\{\|\boldsymbol{\alpha}_{1\boldsymbol{\tau}^c}\|_2 \mid \boldsymbol{a}_1 = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}_1\right\} + \inf\left\{\|\boldsymbol{\alpha}_{2\boldsymbol{\tau}^c}\|_2 \mid \boldsymbol{a}_2 = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}_2\right\} \\ &= d_\alpha(\boldsymbol{a}_1, \mathcal{S}_{\boldsymbol{\tau}}) + d_\alpha(\boldsymbol{a}_2, \mathcal{S}_{\boldsymbol{\tau}}). \end{aligned}$$

Similarly the absolute homogeneity, for any $c \in \mathbb{R}$:

$$\begin{aligned} d_\alpha(c \cdot \boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) &= \inf\left\{\|\boldsymbol{\alpha}'_{\boldsymbol{\tau}^c}\|_2 \mid c \cdot \boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}'\right\} = \inf\left\{\|c \cdot \boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \mid \boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}\right\} \\ &= |c| \cdot \inf\left\{\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \mid \boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}\right\} = |c| \cdot d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}), \end{aligned}$$

which completes the proof that $d_\alpha$ is a seminorm. ∎

**Definition B.3** (Widened subspace). *For subspace $\mathcal{S}_{\boldsymbol{\tau}}$ let*

$$\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu)) := \left\{\boldsymbol{a} \in \mathbb{S}^{p-1} \,\middle|\, d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \leq \gamma\right\} \tag{B.3}$$

*denote its widening by $\gamma$, in the seminorm $d_\alpha$.*

Our analysis works with a specific choice of width $\gamma(c_\mu)$, which depends on the problem parameters $\boldsymbol{a}_0, \theta, |\tau|$ and a constant $c_\mu$, via

$$\gamma(c_\mu) = \frac{c_\mu}{4\log^2 \theta^{-1}} \min\left\{\frac{1}{\sqrt{|\boldsymbol{\tau}|}}, \frac{1}{\sqrt{\mu p}}, \frac{1}{\mu p \sqrt{\theta}\,|\boldsymbol{\tau}|}\right\}. \tag{B.4}$$

**Lemma B.4** (Properties of $C^*_{a_0}\iota\iota^*C_{a_0}$)**.** *Let $M = C^*_{a_0}\iota\iota^*C_{a_0}$, with $a_0 \in \mathbb{S}^{p_0-1}$ $\mu$-shift coherent. The diagonal entries of $M$ satisfy*

$$\begin{cases} M_{ii} = 1 & i \in [-p_0+1, p_0-1] = [\pm p_0], \\ 0 \le M_{ii} \le 1 & i \in [-2p_0+2, -p_0] \cup [p_0, 2p_0-2], \\ M_{ii} = 0 & \textit{otherwise}, \end{cases} \tag{B.5}$$

*and the off-diagonal entries satisfy*

$$\begin{cases} |M_{ij}| \le \mu & 0 < |i-j| < p_0, \;\; \{i \in [-p_0+1, p_0-1]\} \cup \{j \in [-p_0+1, p_0-1]\}\,, \\ |M_{ij}| < 1 & \{i,j \in [-2p_0+2, -p_0]\} \cup \{i,j \in [p_0, 2p_0-2]\}\,, \\ 0 & \textit{otherwise}. \end{cases} \tag{B.6}$$

*Furthermore, let $\tau \subset [\pm p_0]$, and $\tau^c = [\pm 2p_0 - 1] \setminus \tau$. The singular values of submatrix $\iota^*_{\tau} M \iota_{\tau}$ can be bounded as:*

$$\begin{cases} 1 - \mu\,|\tau| \le \sigma_{\min}\left(\iota^*_{\tau} M \iota_{\tau}\right) \le \sigma_{\max}\left(\iota^*_{\tau} M \iota_{\tau}\right) \le 1 + \mu\,|\tau|\,, \\ \sigma_{\max}\left(\iota^*_{\tau^c} M \iota_{\tau}\right) \le \mu\sqrt{p\,|\tau|}, \\ \sigma_{\max}\left(\iota^*_{\tau^c} M \iota_{\tau^c}\right) \le 1 + \mu p. \end{cases} \tag{B.7}$$

**Proof.** Recall the definition of $\iota$, which selects the entries $\{-p_0+1, \dots, 2p_0-2\}$. The entrywise properties of $M$ can be derived by carefully counting the entries of the shifted support. The submatrix $M$ on support $\{-2p_0+2, \dots, 2p_0-2\}$ has an upper bound to be characterized as follows:

$$\left|\iota^*_{[\pm 2p_0-1]} M \iota_{[\pm 2p_0-1]}\right| \le \begin{bmatrix} J & \mu\cdot\mathbf{1} & \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} & \mathbf{0} & \mathbf{0} \\ \mu\cdot\mathbf{1} & I + \mu\cdot\mathbf{1}_o & \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} & \mu\cdot\mathbf{1} & \mathbf{0} \\ \begin{bmatrix} 0 \cdots 0 \end{bmatrix} & \begin{bmatrix} \mu \cdots \mu \end{bmatrix} & 1 & \begin{bmatrix} \mu \cdots \mu \end{bmatrix} & \begin{bmatrix} 0 \cdots 0 \end{bmatrix} \\ \mathbf{0} & \mu\cdot\mathbf{1} & \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} & I + \mu\cdot\mathbf{1}_o & \mu\cdot\mathbf{1} \\ \mathbf{0} & \mathbf{0} & \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} & \mu\cdot\mathbf{1} & J \end{bmatrix}. \tag{B.8}$$

Here, the center row/column vector is indexed at $0$, the matrices $J, I, \mathbf{1}$ and $\mathbf{1}_o$ are square and of size $(p_0-1)^2$. Among which, $I$ is the identity matrix, $\mathbf{1}$ is the ones matrix whereas $\mathbf{1}_o$ has all off diagonal entries equal $1$. Also $|J|$ has property $|J_{ij}| < 1$ for all $i, j$.

As for the singular values, notice that the first and second inequalities consider submatrix not containing $J$ since $\tau \subseteq [\pm p_0]$; thus the first inequality can be derived with Gershgorin disc theorem directly, and the second inequality with the upper bound with its Frobenius norm:

$$\sigma_{\max}\left(\iota^*_{\tau^c} M \iota_{\tau}\right) \le \mu\sqrt{(2p_0-1)\,|\tau|} < \mu\sqrt{p\,|\tau|}. \tag{B.9}$$

Finally by recalling $p = 3p_0 - 2 > 2p_0 - 1$. The last inequality is direct from bound of $\iota^* C_{a_0}$:

$$\sigma_{\max}\left(\iota^*_{\tau^c} M \iota_{\tau^c}\right) \le \left\|C^*_{a_0}\iota\iota^* C_{a_0}\right\|_2 = \left\|\iota^* C_{a_0} C^*_{a_0}\iota\right\|_2 = \left\|\iota^* C^*_{a_0} C_{a_0}\iota\right\|_2 \le 1 + \mu p \tag{B.10}$$

where the third equality is derived via commutativity of convolution. ■

**Lemma B.5** (Shift space vectors in widened subspace)**.** *Let* $(\boldsymbol{a}_0, \theta, |\boldsymbol{\tau}|)$ *satisfy the sparsity-coherence condition* $\mathrm{SCC}(c_\mu)$. *Then for every* $\boldsymbol{a} \in \Re(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$, *every* $\boldsymbol{\alpha}$ *satisfying* $\boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}$ *and* $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \leq \gamma(c_\mu)$ *has*

$$
\left| \|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 - 1 \right| \leq c_\mu; \tag{B.11}
$$

*moreover,* $\boldsymbol{\beta} = \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{a}$ *satisfies*

$$
1 - 3c_\mu \leq \|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 \leq 1 + \frac{c_\mu}{|\boldsymbol{\tau}| \log^2 \theta^{-1}}, \quad \|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_\infty \leq \frac{c_\mu}{\sqrt{|\boldsymbol{\tau}|} \log^2 \theta^{-1}}, \quad \|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2 \leq \frac{c_\mu}{|\boldsymbol{\tau}|\, \theta \log \theta^{-1}} \min\left\{ \sqrt{\theta}, \gamma(c_\mu) \right\}. \tag{B.12}
$$

**Proof.** Write $-1/\log\theta = \theta_{\log}$ and $\gamma = \gamma(c_\mu)$ for convenience. First, by using bounds on $\gamma$ in (B.4) and $\mu\,|\boldsymbol{\tau}| < 1$ we obtain:

$$
\begin{cases}
\gamma \cdot \sqrt{1 + \mu p} \;\leq\; \gamma\,(1 + \sqrt{\mu p}) \;\leq\; c_\mu \theta_{\log}^2 / 2 \\
\gamma \cdot \sqrt{1 + \mu^2 p} \;\leq\; \gamma \left(1 + \sqrt{\mu^2 p}\right) \;\leq\; \dfrac{c_\mu \theta_{\log}^2}{4} \left( \dfrac{1}{\sqrt{|\boldsymbol{\tau}|}} + \sqrt{\mu} \right) \;\leq\; \dfrac{c_\mu \theta_{\log}^2}{2\sqrt{|\boldsymbol{\tau}|}} \\
\gamma \cdot \mu \sqrt{p\,|\boldsymbol{\tau}|} \;\leq\; \gamma \cdot \sqrt{\mu p} \cdot \sqrt{\mu\,|\boldsymbol{\tau}|} \;\leq\; c_\mu \theta_{\log}^2 / 4
\end{cases} \tag{B.13}
$$

Let $\boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 < \gamma$. Utilize properties of $\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}$ from Lemma B.4 and $\mu\,|\boldsymbol{\tau}| < c_\mu/4$ and (B.13), we have:

$$
\begin{aligned}
\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 \;&\geq\; \|\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}_{\boldsymbol{\tau}}\|_2^{-1} \left( \|\boldsymbol{a}\|_2 - \|\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \right) \;\geq\; \|\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}_{\boldsymbol{\tau}}\|_2^{-1} \left( 1 - \|\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}\|_2 \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \right) \\
&\geq\; \frac{1}{\sqrt{1 + \mu\,|\boldsymbol{\tau}|}} \left( 1 - \gamma \cdot \sqrt{1 + \mu p} \right) \;\geq\; \frac{1 - c_\mu/2}{\sqrt{1 + c_\mu/4}} \;\geq\; 1 - c_\mu,
\end{aligned} \tag{B.14}
$$

and similarly, the upper bound can be derived as:

$$
\begin{aligned}
\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 \;&\leq\; \sigma_{\min}^{-1} \left( \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}_{\boldsymbol{\tau}} \right) \left( \|\boldsymbol{a}\|_2 + \|\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \right) \;\leq\; \sigma_{\min}^{-1} \left( \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}_{\boldsymbol{\tau}} \right) \left( 1 + \|\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}\|_2 \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \right) \\
&\leq\; \frac{1}{\sqrt{1 - \mu\,|\boldsymbol{\tau}|}} \left( 1 + \gamma \cdot \sqrt{1 + \mu p} \right) \;\leq\; \frac{1 + c_\mu/2}{\sqrt{1 - c_\mu/4}} \;\leq\; 1 + c_\mu.
\end{aligned} \tag{B.15}
$$

The bound of $\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2$ can be simply obtained using $\mu\,|\boldsymbol{\tau}| < c_\mu/4$ and $\gamma$ bound from (B.13) as:

$$
\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 \;\leq\; \sigma_{\max}^2 \left( \boldsymbol{\iota}_{\boldsymbol{\tau}}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota} \right) \;\leq\; 1 + \mu\,|\boldsymbol{\tau}| \;\leq\; 1 + \frac{c_\mu \theta_{\log}^2}{|\boldsymbol{\tau}|} \tag{B.16}
$$

$$
\begin{aligned}
\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 \;&\geq\; \left( \sigma_{\min} \left( \boldsymbol{\iota}_{\boldsymbol{\tau}}^* \boldsymbol{M} \boldsymbol{\iota}_{\boldsymbol{\tau}} \right) \|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 - \sigma_{\max} \left( \boldsymbol{\iota}_{\boldsymbol{\tau}}^* \boldsymbol{M} \boldsymbol{\iota}_{\boldsymbol{\tau}^c} \right) \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \right)^2 \\
&\geq\; \left( (1 - \mu\,|\boldsymbol{\tau}|)(1 - c_\mu) - \mu \sqrt{p\,|\boldsymbol{\tau}|} \cdot \gamma \right)^2 \;\geq\; 1 - 3c_\mu.
\end{aligned} \tag{B.17}
$$

As for the upper bound of and $\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_\infty$, follow from (B.13), we have:

$$
\begin{aligned}
\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_\infty \;&\leq\; \|\boldsymbol{\iota}_{\boldsymbol{\tau}^c}^* \boldsymbol{M} \boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_\infty + \|\boldsymbol{\iota}_{\boldsymbol{\tau}^c}^* \boldsymbol{M} \boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_\infty \;\leq\; \mu \sqrt{|\boldsymbol{\tau}|}\, \|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 + \sqrt{1 + \mu^2 p}\, \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \\
&\leq\; \frac{c_\mu \theta_{\log}^2 (1 + c_\mu)}{4\,|\boldsymbol{\tau}|} + \gamma \cdot \sqrt{1 + \mu^2 p} \;\leq\; \frac{c_\mu \theta_{\log}^2}{\sqrt{|\boldsymbol{\tau}|}};
\end{aligned} \tag{B.18}
$$

the bound for $\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2$ requires two inequalities, we know

$$
\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2 \;\leq\; \|\boldsymbol{\iota}_{\boldsymbol{\tau}^c}^* \boldsymbol{M} \boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 + \|\boldsymbol{\iota}_{\boldsymbol{\tau}^c}^* \boldsymbol{M} \boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \;\leq\; \mu \sqrt{p\,|\boldsymbol{\tau}|}\, \|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 + (1 + \mu p)\, \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2\,, \tag{B.19}
$$

for the first inequality, use $\left(\mu\left|\boldsymbol{\tau}\right|^2\right)^{3/4}\left(\mu p^2\theta^2\right)^{1/4} = \mu\sqrt{p\theta}\left|\boldsymbol{\tau}\right|^{3/2} < c_\mu\theta_{\log}^2/4$ , definition of $\gamma$ and $\theta\left|\boldsymbol{\tau}\right| \le c_\mu\theta_{\log}^2/4$ we have:

$$
\begin{aligned}
(\text{B.19}) &\le \frac{\mu\sqrt{p\theta}\left|\boldsymbol{\tau}\right|^{3/2}}{\sqrt{\theta}\left|\boldsymbol{\tau}\right|}(1+c_\mu) + \frac{\sqrt{\theta\left|\boldsymbol{\tau}\right|}\cdot\sqrt{\left|\boldsymbol{\tau}\right|}\gamma}{\sqrt{\theta}\left|\boldsymbol{\tau}\right|} + \frac{\mu p\sqrt{\theta}\left|\boldsymbol{\tau}\right|\gamma}{\sqrt{\theta}\left|\boldsymbol{\tau}\right|} \\
&\le \frac{2c_\mu\theta_{\log}^2 + c_\mu\theta_{\log}^3 + c_\mu\theta_{\log}^2}{4\sqrt{\theta}\left|\boldsymbol{\tau}\right|} \le \frac{c_\mu\theta_{\log}^2}{\sqrt{\theta}\left|\boldsymbol{\tau}\right|},
\end{aligned} \tag{B.20}
$$

and similarly for the second inequality, use both conditions of $\mu$, we have:

$$
\begin{aligned}
(\text{B.19}) &\le \frac{\gamma}{\theta\left|\boldsymbol{\tau}\right|}\cdot\frac{\mu\sqrt{p}\theta\left|\boldsymbol{\tau}\right|^{3/2}}{\gamma}(1+c_\mu) + \gamma + \mu p\gamma \\
&\le \frac{\gamma}{\theta\left|\boldsymbol{\tau}\right|}\cdot\frac{4\mu\sqrt{p}\theta\left|\boldsymbol{\tau}\right|^{3/2}}{c_\mu\theta_{\log}^2}\cdot\max\left\{\sqrt{\left|\boldsymbol{\tau}\right|},\,\sqrt{\mu p},\,\mu p\sqrt{\theta}\left|\boldsymbol{\tau}\right|\right\} + \frac{\gamma}{\theta\left|\boldsymbol{\tau}\right|}\cdot\theta\left|\boldsymbol{\tau}\right| + \frac{\gamma}{\theta\left|\boldsymbol{\tau}\right|}\cdot\mu p\theta\left|\boldsymbol{\tau}\right| \\
&\le \frac{\gamma}{\theta\left|\boldsymbol{\tau}\right|}\cdot\left(\frac{4}{c_\mu\theta_{\log}^2}\cdot\max\left\{\mu\left|\boldsymbol{\tau}\right|^2\cdot\sqrt{p}\theta,\,\mu(p\theta)\left|\boldsymbol{\tau}\right|\cdot\sqrt{\mu\left|\boldsymbol{\tau}\right|},\,\mu\sqrt{p\theta}\left|\boldsymbol{\tau}\right|^{3/2}\cdot\mu p\theta\left|\boldsymbol{\tau}\right|\right\} + \frac{c_\mu\theta_{\log}^2}{4} + \frac{c_\mu\theta_{\log}^2}{4}\right) \\
&\le \frac{\gamma}{\theta\left|\boldsymbol{\tau}\right|}\left(\frac{c_\mu\theta_{\log}}{4} + \frac{c_\mu\theta_{\log}^2}{4} + \frac{c_\mu\theta_{\log}^2}{4}\right) \le \frac{c_\mu\theta_{\log}\gamma}{\theta\left|\boldsymbol{\tau}\right|},
\end{aligned} \tag{B.21}
$$

which completes the proof. ■

**Corollary B.6** ( $\left|\langle\boldsymbol{\beta}_{\boldsymbol{\tau}^c}, \boldsymbol{x}_{0,\boldsymbol{\tau}^c}\rangle\right|$ is small)**.** *Given $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$ and $\left|\boldsymbol{\tau}\right|, c_\mu$ such that $(\boldsymbol{a}_0, \theta, \left|\boldsymbol{\tau}\right|)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Write $\lambda = c_\lambda/\sqrt{\left|\boldsymbol{\tau}\right|}$ with some $c_\lambda \ge 1/5$, then if $c_\mu \le \frac{c_\lambda}{25}$,*

$$
\mathbb{P}\left[\left|\sum_{i\in\boldsymbol{\tau}^c}\boldsymbol{\beta}_i\boldsymbol{x}_{0i}\right| > \frac{\lambda}{10}\right] \le 2\theta, \qquad \mathbb{P}\left[\left|\sum_{i}\boldsymbol{\beta}_i\boldsymbol{x}_{0i}\right| > \frac{\lambda}{10}\right] \le \theta\left|\boldsymbol{\tau}\right| + 2\theta. \tag{B.22}
$$

**Proof.** We bound tail probability of the first result with Gaussian moments Lemma J.2 and Bernstein inequality Lemma J.4. Via Hölder's inequality, $\sum_{i\in\boldsymbol{\tau}^c}\mathbb{E}(\beta_i x_i)^q = \mathbb{E}x_0^q\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\right\|_q^q \le \theta(q-1)!!\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\right\|_2^2\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\right\|_\infty^{q-2}$, thus

$$
\mathbb{P}\left[\left|\sum_{i\in\boldsymbol{\tau}^c}\boldsymbol{\beta}_i\boldsymbol{x}_{0i}\right| > \lambda/10\right] \le 2\exp\left(\frac{-(\lambda/10)^2}{2\theta\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\right\|_2^2 + 2(\lambda/10)\left\|\boldsymbol{\beta}_{\tau^c}\right\|_\infty}\right) \tag{B.23}
$$

Write $\theta_{\log} = -\frac{1}{\log\theta}$, Lemma B.5 imples when $c_\mu \le \frac{c_\lambda}{25}$, we have $\theta\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\right\|_2^2 \le \frac{c_\mu^2\theta_{\log}^2}{\left|\boldsymbol{\tau}\right|^2} \le \frac{\theta_{\log}\lambda^2}{625}$ and $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\right\|_\infty \le \frac{c_\mu\theta_{\log}}{\sqrt{\left|\boldsymbol{\tau}\right|}} \le \frac{\theta_{\log}\lambda}{25}$, therefore,

$$
(\text{B.23}) \le 2\exp\left(\frac{-\lambda^2/100}{2\theta_{\log}\lambda^2/625 + 2(\theta_{\log}\lambda/25)\cdot(\lambda/10)}\right) \le 2\exp\left(\log\theta\right) \le 2\theta \tag{B.24}
$$

The second tail bound is straight forward from the first tail bound as follows:

$$
\begin{aligned}
\mathbb{P}\left[\left|\sum_{i}\boldsymbol{\beta}_i\boldsymbol{x}_{0i}\right| > \frac{\lambda}{10}\right] &\le \mathbb{P}\left[\left|\boldsymbol{\beta}_{\boldsymbol{\tau}}^*\boldsymbol{x}_{\boldsymbol{\tau}}\right| + \left|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}^*\boldsymbol{x}_{\boldsymbol{\tau}^c}\right| > \lambda/10\right] \\
&\le \mathbb{P}\left[\boldsymbol{x}_{\boldsymbol{\tau}} \ne \mathbf{0}\right] + \mathbb{P}\left[\boldsymbol{x}_{\boldsymbol{\tau}} = \mathbf{0}\right]\cdot\mathbb{P}\left[\left|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}^*\boldsymbol{x}_{\boldsymbol{\tau}^c}\right| > \lambda/10\right] \\
&\le \theta\left|\boldsymbol{\tau}\right| + 2\theta.
\end{aligned} \tag{B.25}
$$

■

**Corollary B.7** ( $\left|\left\langle \boldsymbol{\beta}_{\boldsymbol{\tau}\setminus(0)}, \boldsymbol{x}_{0,\boldsymbol{\tau}\setminus(0)} \right\rangle\right|$ is small near shifts)**.** *Suppose that $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $|\boldsymbol{\tau}|$, $c_\mu$ such that $(\boldsymbol{a}_0, \theta, |\boldsymbol{\tau}|)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$, then if $c_\mu \le \frac{1}{10}$, for any $\boldsymbol{a}$ such that $\left|\boldsymbol{\beta}_{(1)}\right| \le \frac{\lambda}{4\log\theta^{-1}}$, we have*

$$\mathbb{P}\left[\left|\sum_{i\in\boldsymbol{\tau}\setminus(0)} \boldsymbol{\beta}_i \boldsymbol{x}_{0i}\right| > \frac{2\lambda}{5}\right] \le 2\theta \tag{B.26}$$

**Proof.** For the last tail bound, write $\boldsymbol{x} = \boldsymbol{\omega}\circ\boldsymbol{g}$. Wlog define $\boldsymbol{\beta}_0$ be the largest correlation $\boldsymbol{\beta}_{(0)}$, define random variables $s' = \left\langle \boldsymbol{\beta}_{\boldsymbol{\tau}\setminus\{0\}}, \boldsymbol{x}_{\boldsymbol{\tau}\setminus\{0\}} \right\rangle$. Firstly most of the entries of $\boldsymbol{x}_{\boldsymbol{\tau}}$ would be zero since via Bernstein inequality with $\theta|\boldsymbol{\tau}| < 0.1$:

$$\mathbb{P}\left[\sum_{i\in\boldsymbol{\tau}} \boldsymbol{\omega}_i > \log\theta^{-1}\right] \le \mathbb{P}\left[\sum_{i\in\boldsymbol{\tau}} \boldsymbol{\omega}_i > \theta|\boldsymbol{\tau}| + 0.9\log\theta^{-1}\right] \le \exp\left(\frac{-0.9^2\log^2\theta^{-1}}{2\left(\theta|\boldsymbol{\tau}| + 0.9\log\theta^{-1}/3\right)}\right) \le \theta \tag{B.27}$$

thus with probability at least $1-\theta$, we can write $s'$ as a Gaussian r.v. with variation bounded as $\mathbb{E}s'^2 \le \mathbb{E}\left[\sum_{i=1}^{\log\theta^{-1}} \boldsymbol{\beta}_i \boldsymbol{g}_i\right]^2 = \log\theta^{-1}\boldsymbol{\beta}_{(1)}^2$, then via Gaussian tail bound Lemma J.1:

$$\mathbb{P}\left[|s'| > 0.4\lambda\right] \le \mathbb{P}\left[|g| > \frac{0.4\lambda}{\sqrt{\log\theta^{-1}}\left|\boldsymbol{\beta}_{(1)}\right|}\right] + \mathbb{P}\left[\sum_{i\in\boldsymbol{\tau}} \boldsymbol{\omega}_i > \log\theta^{-1}\right] \le \frac{2}{\sqrt{2\pi}}\exp\left(-1.2\log\theta^{-1}\right) + \theta \le 2\theta, \tag{B.28}$$

■

# C Euclidean gradient as soft-thresholding in shift space

In this section, we will study the Euclidean gradient (4.6), by deriving bounds showing that the $\boldsymbol{\chi}$ operator approximates a soft-thresholding function in shift space (Lemma C.2 and Corollary C.4). Furthermore, we will show the operator $\boldsymbol{\chi}[\boldsymbol{\beta}_i]$ is monotone in $|\boldsymbol{\beta}_i|$ from Lemma C.3. A figure of visualized $\boldsymbol{\chi}$ operator is shown in Figure 13.

**Expectation of $\boldsymbol{\chi}$ operator.** To understand the $\boldsymbol{\chi}$ operator, we shall first consider a simple case—when $\boldsymbol{x}_0$ is highly sparse. By definition of $\boldsymbol{\beta}$ from (4.3) we can see that $\boldsymbol{\beta}$ has a short support of size at most $2p-1$, when $\boldsymbol{x}_0$ has support entries separated by at least $2p$, the entries of vector $\boldsymbol{\chi}[\boldsymbol{\beta}]_i$ become sum of independent random variables as:

$$\boldsymbol{\chi}[\boldsymbol{\beta}]_i = \left\langle s_{-i}[\boldsymbol{x}_0], \mathcal{S}_\lambda\left[\boldsymbol{x}_0 * \check{\boldsymbol{\beta}}\right]\right\rangle \underbrace{=}_{\boldsymbol{x}_0 \text{ sep.}} \left\langle s_{-i}[\boldsymbol{x}_0], \mathcal{S}_\lambda\left[\boldsymbol{\beta}_i s_{-i}[\boldsymbol{x}_0]\right]\right\rangle = \sum_{j\in\mathrm{supp}(\boldsymbol{x}_0)} \boldsymbol{g}_j \cdot \mathcal{S}_\lambda\left[\boldsymbol{g}_j \cdot \boldsymbol{\beta}_i\right]$$

where $(\boldsymbol{g}_j)_{j\in[n]}$ are standard Gaussian r.v.s.

The following lemma describes the behavior of the summands in the above expression:

**Lemma C.1** (Gaussian smoothed soft-thresholding)**.** *Let $g \sim \mathcal{N}(0,1)$. Then for every $b, s \in \mathbb{R}$ and $\lambda > 0$,*

$$\mathbb{E}_g\Big[g\mathcal{S}_\lambda\left[b\cdot g + s\right]\Big] = b\left(1 - \mathrm{erf}_b(\lambda, s)\right), \tag{C.1}$$

*where*

$$\mathrm{erf}_b(\lambda, s) = \frac{1}{2}\mathrm{erf}\left(\frac{\lambda+s}{\sqrt{2}\,|b|}\right) + \frac{1}{2}\mathrm{erf}\left(\frac{\lambda-s}{\sqrt{2}\,|b|}\right). \tag{C.2}$$

*Furthermore, for $s=0$, $b\in[-1,1]$ and $\varepsilon\in(0,1/4)$, letting $\sigma = \mathrm{sign}(b)$ we have*

$$\sigma\mathcal{S}_{\nu_2'\lambda}\left[b\right] \le \sigma\mathbb{E}_g\Big[g\mathcal{S}_\lambda\left[b\cdot g\right]\Big] \le \sigma\mathcal{S}_{\nu_1'(\varepsilon)\lambda}\left[b\right] + \varepsilon \tag{C.3}$$

*where $\nu_1'(\varepsilon) = 1/(2\sqrt{-\log\varepsilon})$ and $\nu_2' = \sqrt{2/\pi}$.*

**Proof.** Wlog assume $b>0$. Write $f$ as the pdf of standard Gaussian distribution. With integral by parts:

$$\int_{-\infty}^{t} t' f(t')dt' = -f(t), \qquad \int_{-\infty}^{t} t'^2 f(t')dt' = \frac{1}{2}\mathrm{erf}\left(\frac{t}{\sqrt{2}}\right) - tf(t)$$

Integrating, we obtain

$$\mathbb{E}\Big[g\mathcal{S}_\lambda\left[b\cdot g + s\right]\Big] = \int_{t\ge\frac{\lambda-s}{b}} \left(bt^2 - (\lambda-s)t\right)\, f(t)dt + \int_{t\le-\frac{\lambda+s}{b}} \left(bt^2 + (\lambda+s)t\right)\, f(t)dt,$$

by writing $L = \lambda - s$, the integral of first summand

$$\int_{t\ge\frac{L}{b}} \left(bt^2 - Lt\right) f(t)dt = b\left[\frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{L}{\sqrt{2}b}\right) + \frac{L}{b}f\left(\frac{L}{b}\right)\right] - Lf\left(\frac{L}{b}\right) = \frac{b}{2} - \frac{b}{2}\mathrm{erf}\left(\frac{L}{\sqrt{2}b}\right),$$

and similarly for the second summand, which gives

$$\mathbb{E}\Big[g\mathcal{S}_\lambda\left[b\cdot g + \boldsymbol{s}\right]\Big] = \frac{b}{2} - \frac{b}{2}\mathrm{erf}\left(\frac{\lambda-s}{\sqrt{2}b}\right) + \frac{b}{2} - \frac{b}{2}\mathrm{erf}\left(\frac{\lambda+s}{\sqrt{2}b}\right) = b\left(1 - \mathrm{erf}_b(\lambda, s)\right)$$

For $b < 0$, alternatively we have

$$\mathbb{E}\Big[gS_\lambda[-|b|\cdot g + s]\Big] = -\mathbb{E}[gS_\lambda[|b|\cdot g - s] = -|b|(1-\mathrm{erf}_b(\lambda, -s)) = b(1-\mathrm{erf}_b(\lambda, s)),$$

To show (C.3), via definition of error function, for $x > 0$, we know:

$$\min\left\{1-\varepsilon, \frac{1-\varepsilon}{\sqrt{\log(1/\varepsilon)}}x\right\} \le \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \le \frac{2x}{\sqrt{\pi}} \tag{C.4}$$

where the lower bound is derived by first knowing erf is increasing thus for all $x > \sqrt{\log(1/\varepsilon)}$,

$$\mathrm{erf}(x) \ge 1 - e^{-x^2} \ge 1 - e^{\log\varepsilon} = 1-\varepsilon$$

and from concavity of $\mathrm{erf}$ we have for $0 < x < \sqrt{\log(1/\varepsilon)} = T$,

$$\mathrm{erf}(x) \ge \frac{\mathrm{erf}(T) - \mathrm{erf}(0)}{T-0}x + \mathrm{erf}(0) \ge \frac{1-\varepsilon}{\sqrt{\log(1/\varepsilon)}}x.$$

Lastly plug (C.4) into (C.1) and apply condition $|b| \le 1$ and $\varepsilon < 1/4$ we have

$$|b| - \sqrt{\frac{2}{\pi}}\lambda \le |b| - |b|\,\mathrm{erf}\left(\frac{\lambda}{\sqrt{2}\,|b|}\right) \le \max\left\{|b|\,\varepsilon, |b| - \frac{\lambda(1-\varepsilon)}{\sqrt{2\log(1/\varepsilon)}}\right\} \le \max\left\{\varepsilon, |b| - \frac{\lambda}{2\sqrt{\log(1/\varepsilon)}}\right\},$$

which completes the proof. ■

This lemma establishes when $\boldsymbol{x}_0$ is separated, then $\boldsymbol{\chi}$ is soft thresholding operator on $\boldsymbol{\beta}$ with threshold about $\lambda/2$. This phenomenon extends beyond the separated case, as long as when $\boldsymbol{x}_0$ is sufficiently sparse (when Definition B.1 holds). Recall that $\boldsymbol{\chi} : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$\boldsymbol{\chi}[\boldsymbol{\beta}] = \check{\boldsymbol{C}}_{\boldsymbol{x}_0}\mathcal{S}_\lambda\left[\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{\beta}\right]. \tag{C.5}$$

The following lemma bounds its expectation:

**Lemma C.2** (Expectation of $\boldsymbol{\chi}(\boldsymbol{\beta})$)**.** *Let $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ and $\lambda > 0$, then for every $\boldsymbol{a} \in \mathbb{S}^{p-1}$ and every $i \in [n]$, define the operator $\boldsymbol{\chi}$ as in (C.5), then*

$$n^{-1}\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i = \theta\boldsymbol{\beta}_i\left(1 - \mathbb{E}_{\boldsymbol{s}_i}\mathrm{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right) \tag{C.6}$$

*where $\boldsymbol{s}_i = \sum_{\ell\neq i}\boldsymbol{\beta}_\ell\boldsymbol{x}_{0\ell}$. Suppose $(\boldsymbol{a}_0, \theta, |\boldsymbol{\tau}|)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$ and $\lambda = c_\lambda/\sqrt{|\boldsymbol{\tau}|}$ for some $c_\lambda > 1/5$ and $\sigma_i = \mathrm{sign}(\boldsymbol{\beta}_i)$, then there exists some numerical constant $\bar{c}$ such that if $c_\mu \le \bar{c}$ then for every $\boldsymbol{a} \in \mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ and every $i \in [n]$, (C.6) has upper bound*

$$\sigma_i n^{-1}\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i \le \sigma_i n^{-1}\overline{\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]}_i := \begin{cases} 4\theta^2\,|\boldsymbol{\tau}|\,|\boldsymbol{\beta}_i| & |\boldsymbol{\beta}_i| < \nu_1\lambda \\ \theta\left(|\boldsymbol{\beta}_i| - \nu_1\lambda/2\right) & |\boldsymbol{\beta}_i| \ge \nu_1\lambda \end{cases}, \tag{C.7}$$

*and lower bound*

$$\sigma_i n^{-1}\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i \ge \sigma_i n^{-1}\underline{\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]}_i =: \theta\mathcal{S}_{\nu_2\lambda}\left[|\boldsymbol{\beta}_i|\right], \tag{C.8}$$

*where $\nu_1 = 1/\left(2\sqrt{\log\theta^{-1}}\right)$, $\nu_2 = \sqrt{2/\pi}$.*

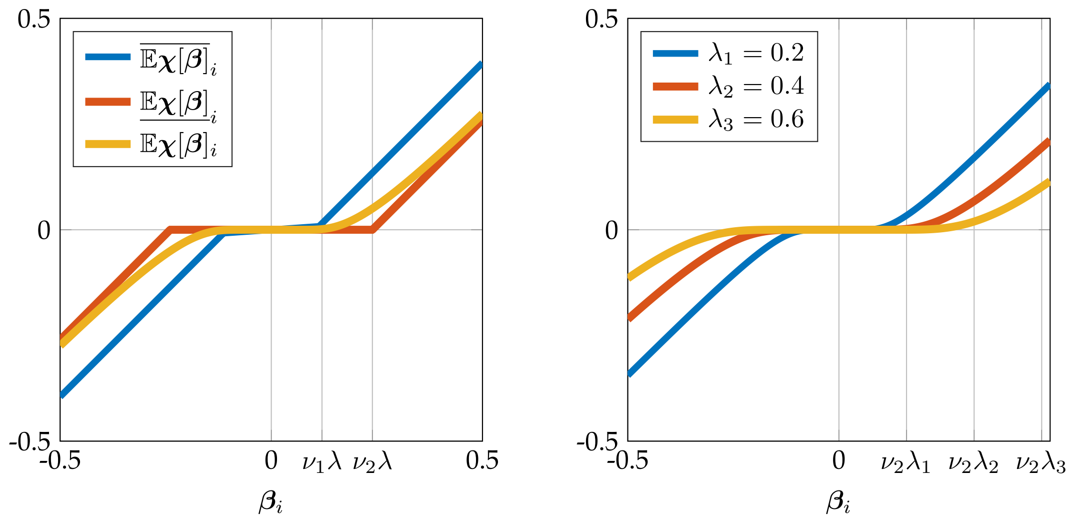


**Figure 13: A numerical example of $\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i$.** We provide figures for the expectation of $\boldsymbol{\chi}$ when entries of $\boldsymbol{x}_0$ are $2p$-separated. Left: the yellow line is the function $\boldsymbol{\beta}_i \to \boldsymbol{\beta}_i\left(1-\operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda,0)\right)$ derived from (C.1), and the blue/red lines are its upper/lower bound (C.3) utilized in the analysis respectively. Right: functions of $\boldsymbol{\beta}_i \to \boldsymbol{\beta}_i\left(1-\operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda,0)\right)$ with different $\lambda$, the section of function of $\boldsymbol{\beta}_i > \nu_2\lambda$ are close to linear.

This lemma shows the expectation of $\boldsymbol{\chi}[\boldsymbol{\beta}]_i$ acts like a shrinkage operation on $|\boldsymbol{\beta}_i|$: for large $|\boldsymbol{\beta}_i|$, it acts like a soft thresholding operation, and for small $|\boldsymbol{\beta}_i|$, it reduces $|\boldsymbol{\beta}_i|$ by multiplying a very small number $4\theta\,|\boldsymbol{\tau}| \ll 1$. We rigorously prove this segmentation of $\boldsymbol{\chi}$ operator as follows:

**Proof.** First, since $s_i[\boldsymbol{x}_0] \equiv_d s_j[\boldsymbol{x}_0]$,

$$\boldsymbol{\chi}[\boldsymbol{\beta}]_i = \boldsymbol{e}_i^*\breve{\boldsymbol{C}}_{\boldsymbol{x}_0}\mathcal{S}_\lambda\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{\beta}\right] = \left\langle s_{-i}[\boldsymbol{x}_0], \mathcal{S}_\lambda\left[\boldsymbol{x}_0 * \breve{\boldsymbol{\beta}}\right]\right\rangle \equiv_d \left\langle s_{-j}[\boldsymbol{x}_0], \mathcal{S}_\lambda\left[s_{i-j}[\boldsymbol{x}_0] * \breve{\boldsymbol{\beta}}\right]\right\rangle = \boldsymbol{\chi}[s_{j-i}[\boldsymbol{\beta}]]_j$$

Thus wlog let us consider $i=0$ and write $\boldsymbol{x}$ as $\boldsymbol{x}_0$. The random variable $\boldsymbol{\chi}[\boldsymbol{\beta}]_0$ can be written sum of random variables as:

$$\boldsymbol{\chi}\left[\boldsymbol{\beta}\right]_0 = \left\langle \boldsymbol{x}, \mathcal{S}_\lambda\left[\boldsymbol{\beta}_0\boldsymbol{x}_0 + \sum_{\ell\neq 0}\boldsymbol{\beta}_\ell s_{-\ell}[\boldsymbol{x}]\right]\right\rangle = \sum_{j\in[n]}\boldsymbol{x}_j\mathcal{S}_\lambda\left[\boldsymbol{\beta}_0\boldsymbol{x}_j + \sum_{\ell\neq 0}\boldsymbol{\beta}_\ell\boldsymbol{x}_{j+\ell}\right],$$

and a random variable $Z_j(\boldsymbol{\beta})$ is defined as

$$Z_j(\boldsymbol{\beta}) = \boldsymbol{x}_j\mathcal{S}_\lambda\left[\boldsymbol{\beta}_0\boldsymbol{x}_j + \sum_{\ell\in[\pm p]\setminus 0}\boldsymbol{\beta}_\ell\boldsymbol{x}_{j+\ell}\right], \tag{C.9}$$

gives $\boldsymbol{\chi}[\boldsymbol{\beta}]_0 = \sum_{j\in[n]} Z_j(\boldsymbol{\beta})$ as sum of r.v.s. of same distribution and thus $n^{-1}\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_0 = \mathbb{E}Z_0(\boldsymbol{\beta})$. Define a random variable $\boldsymbol{s}_0 = \sum_{\ell\neq 0}\boldsymbol{\beta}_\ell\boldsymbol{x}_\ell$, which is independent of $\boldsymbol{x}_0$. From Lemma C.1, we can conclude

$$n^{-1}\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_0 = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{s}_0}\boldsymbol{x}_0\mathcal{S}_\lambda\left[\boldsymbol{\beta}_0\boldsymbol{x}_0 + \boldsymbol{s}_0\right] = \theta\boldsymbol{\beta}_0\left(1 - \mathbb{E}_{\boldsymbol{s}_0}\operatorname{erf}_{\boldsymbol{\beta}_0}(\lambda, \boldsymbol{s}_0)\right) \tag{C.10}$$

so that (C.6) holds for $i=0$, and hence for all $i$.

1. (Upper bound of $\mathbb{E}Z$) Wlog assume $\boldsymbol{\beta}_0 \geq 0$ and write $Z = Z_0$. We derive the upper bound on $\mathbb{E}Z$ in two pieces.

(1). First, since $\mathbb{E}\boldsymbol{x}_0\mathcal{S}_\lambda\left[0\cdot\boldsymbol{x}_0+\boldsymbol{s}_0\right]=0$, we have

$$
\begin{aligned}
\mathbb{E}Z(\boldsymbol{\beta}) &\le \boldsymbol{\beta}_0 \sup_{\beta\in[0,\boldsymbol{\beta}_0]} \frac{d}{d\beta}\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{s}_0}\boldsymbol{x}_0\mathcal{S}_\lambda\left[\beta\boldsymbol{x}_0+\boldsymbol{s}_0\right] = \theta\boldsymbol{\beta}_0 \sup_{\beta\in[0,\boldsymbol{\beta}_0]} \frac{d}{d\beta}\int_{|\beta g+\boldsymbol{s}_0|>\lambda} g\left(\beta g+\boldsymbol{s}_0-\operatorname{sign}(\beta g+\boldsymbol{s}_0)\cdot\lambda\right)d\mu(g)d\mu(\boldsymbol{s}_0)\\
&= \theta\boldsymbol{\beta}_0 \sup_{\beta\in[0,\boldsymbol{\beta}_0]} \mathbb{E}_{g,\boldsymbol{s}_0}\left[g^2\mathbf{1}_{\{|\beta g+\boldsymbol{s}_0|>\lambda\}}\right] \le \theta\boldsymbol{\beta}_0 \sup_{\beta\in[0,\boldsymbol{\beta}_0]} \mathbb{E}_{g,\boldsymbol{s}_0}\left[g^2\left(\mathbf{1}_{\{|\beta g|>\frac{9\lambda}{10}\}}+\mathbf{1}_{\{|\boldsymbol{s}_0|>\frac{\lambda}{10}\}}\right)\right]\\
&\le \theta\boldsymbol{\beta}_0\left(\left(\mathbb{E}g^6\right)^{1/3}\mathbb{P}\left[|\boldsymbol{\beta}_0 g|>(9\lambda/10)\right]^{2/3}+\mathbb{P}\left[|\boldsymbol{s}_0|>\lambda/10\right]\right) \qquad (C.11)
\end{aligned}
$$

We bound the tail probability of $\boldsymbol{s}_0$ using Corollary B.6 where

$$
\mathbb{P}\left[|\boldsymbol{s}_0|>\lambda/10\right]\le\mathbb{P}\left[|\textstyle\sum_i\boldsymbol{\beta}_i\boldsymbol{x}_i|>\lambda/10\right]\le\theta\,|\boldsymbol{\tau}|+2\theta\le3\theta\,|\boldsymbol{\tau}|\,. \qquad (C.12)
$$

On the other hand, the first term in (C.11) can be derived by pdf of Gaussian r.v. Lemma J.1 as:

$$
\left(\mathbb{E}g^6\right)^{1/3}\mathbb{P}\left[|\boldsymbol{\beta}_0 g|>(9\lambda/10)\right]^{2/3}\le\sqrt[3]{15}\left(\frac{10\boldsymbol{\beta}_0}{9\lambda\sqrt{2\pi}}\right)^{2/3}\exp\left(-\frac{\lambda^2}{4\boldsymbol{\beta}_0^2}\right)\le\frac{3}{2}\left(\frac{\boldsymbol{\beta}_0}{\lambda}\right)^{2/3}\exp\left(-\frac{\lambda^2}{4\boldsymbol{\beta}_0^2}\right). \qquad (C.13)
$$

Combine (B.24), (C.13), when $\boldsymbol{\beta}_0<\nu_1\lambda$, we know $e^{-\frac{\lambda^2}{4\boldsymbol{\beta}_0^2}}\le e^{\log\theta}\le\theta\,|\boldsymbol{\tau}|$. The first type of upper bound $\mathbb{E}Z$ is derived as

$$
\forall\,\boldsymbol{\beta}_0\in[0,\nu_1\lambda]\,,\quad \mathbb{E}Z(\boldsymbol{\beta})\le\theta\boldsymbol{\beta}_0\left(\frac{3}{2}\nu_1^{2/3}\exp\left(-\frac{\lambda^2}{4\boldsymbol{\beta}_0^2}\right)+3\theta\,|\boldsymbol{\tau}|\right)\le4\theta^2\,|\boldsymbol{\tau}|\,\boldsymbol{\beta}_0. \qquad (C.14)
$$

(2). The second type of upper bound can be derived directly from Lemma C.1:

$$
\begin{aligned}
\mathbb{E}Z(\boldsymbol{\beta}) &\le \mathbb{E}_{\boldsymbol{x}_0}\mathbb{E}_{\boldsymbol{s}_0}\boldsymbol{x}_0\mathcal{S}_\lambda\left[\boldsymbol{\beta}_0\boldsymbol{x}_0+\boldsymbol{s}_0\right]\le\mathbb{E}_{\boldsymbol{x}_0}\boldsymbol{x}_0\mathcal{S}_\lambda\left[\boldsymbol{\beta}_0\boldsymbol{x}_0\right]+\mathbb{E}_{\boldsymbol{x}_0}|\boldsymbol{x}_0|\,\mathbb{E}_{\boldsymbol{s}_0}|\boldsymbol{s}_0|\\
&\le\theta\cdot\left(\mathcal{S}_{\nu_1'\lambda}\left[\boldsymbol{\beta}_0\right]+\varepsilon+\sqrt{2/\pi}\cdot\mathbb{E}\,|\boldsymbol{s}_0|\right), \qquad (C.15)
\end{aligned}
$$

where $\mathbb{E}\,|s|$ can be bounded with $\|\boldsymbol{\beta}\|_2$ and $\theta\,|\boldsymbol{\tau}|<c_\mu\theta_{\log}$ from Lemma B.5. When $c_\mu<\frac{1}{10}$, observe that

$$
\mathbb{E}\,|s|\le\sqrt{\sum_\ell\mathbb{E}\boldsymbol{x}_\ell^2\boldsymbol{\beta}_\ell^2}\le\sqrt{\theta}\left(\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2+\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2\right)\le\sqrt{\theta}\,(1+c_\mu)+\frac{c_\mu\theta_{\log}}{|\boldsymbol{\tau}|}\le\frac{2c_\mu\theta_{\log}}{\sqrt{|\boldsymbol{\tau}|}}. \qquad (C.16)
$$

Now choose $\varepsilon=\theta\le\frac{c_\mu\theta_{\log}}{|\boldsymbol{\tau}|}$, so that $\nu_1'=\nu_1=\frac{\sqrt{\theta_{\log}}}{2}$ in (C.15). Since $c_\mu<\frac{c_\lambda}{25}$ we gain

$$
\begin{aligned}
\mathbb{E}Z(\boldsymbol{\beta}) &\le \theta\left(\mathcal{S}_{\nu_1\lambda}\left[\boldsymbol{\beta}_0\right]+\frac{c_\mu\theta_{\log}}{|\boldsymbol{\tau}|}+\sqrt{\frac{2}{\pi}}\cdot\frac{2c_\mu\theta_{\log}}{\sqrt{|\boldsymbol{\tau}|}}\right)\le\theta\left(\mathcal{S}_{\nu_1\lambda}\left[\boldsymbol{\beta}_0\right]+\frac{3c_\mu\theta_{\log}}{\sqrt{|\boldsymbol{\tau}|}}\right)\\
&\le\theta\left(\mathcal{S}_{\nu_1\lambda}\left[\boldsymbol{\beta}_0\right]+\frac{\sqrt{\theta_{\log}}}{5}\lambda\right)\le\theta\left(\mathcal{S}_{\nu_1\lambda}\left[\boldsymbol{\beta}_0\right]+\frac{1}{2}\nu_1\lambda\right) \qquad (C.17)
\end{aligned}
$$

(3). Combine both (C.14) and (C.17), we can thus conclude that

$$
\mathbb{E}Z(\boldsymbol{\beta}):=\overline{\mathbb{E}Z(\boldsymbol{\beta})}\le\begin{cases}4\theta^2\,|\boldsymbol{\tau}|\,\boldsymbol{\beta}_0 & \boldsymbol{\beta}_0\le\nu_1\lambda\\ \theta\left(\boldsymbol{\beta}_0-\frac{\nu_1}{2}\lambda\right) & \boldsymbol{\beta}_0>\nu_1\lambda\end{cases}. \qquad (C.18)
$$

2. (<u>Lower bound of $\mathbb{E}Z$</u>) On the other hand, for the lower bound for $\mathbb{E}Z$, use the fact that $\operatorname{erf}_{\boldsymbol{\beta}}(\lambda,s)$ is concave in $\boldsymbol{s}_0$, we have

$$
\mathbb{E}Z(\boldsymbol{\beta})=\mathbb{E}_{\boldsymbol{s}_0}\mathbb{E}_{\boldsymbol{x}_0}\boldsymbol{x}_0\mathcal{S}_\lambda\left[\boldsymbol{\beta}_0\boldsymbol{x}_0+\boldsymbol{s}_0\right]=\theta\cdot\mathbb{E}_{\boldsymbol{s}_0}\left[\boldsymbol{\beta}_0-\frac{\boldsymbol{\beta}_0}{2}\cdot\operatorname{erf}\left(\frac{\lambda-\boldsymbol{s}_0}{\sqrt{2}\,|\boldsymbol{\beta}_0|}\right)-\frac{\boldsymbol{\beta}_0}{2}\cdot\operatorname{erf}\left(\frac{\lambda+\boldsymbol{s}_0}{\sqrt{2}\,|\boldsymbol{\beta}_0|}\right)\right]
$$

$$\geq \theta\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0 \cdot \operatorname{erf}\left(\frac{\lambda}{\sqrt{2}\,|\boldsymbol{\beta}_0|}\right)\right) \geq \theta \cdot \mathcal{S}_{\nu_2'\lambda}\left[\boldsymbol{\beta}_0\right] =: \underline{\mathbb{E}Z(\boldsymbol{\beta})}. \tag{C.19}$$

The proof of $\boldsymbol{\beta}_0 < 0$ is in the same vein. For cases of $i \neq 0$, since $\boldsymbol{\chi}[\boldsymbol{\beta}]_i \equiv_d \boldsymbol{\chi}[s_{-i}[\boldsymbol{\beta}]]_0$, replace $\boldsymbol{\beta}_0$ with $\boldsymbol{\beta}_i$ we obtain the desired result. ■

**Monotonicity of $\boldsymbol{\chi}$.** Another convenient fact of $\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i$ is that it is monotone increasing w.r.t. $|\boldsymbol{\beta}_i|$. The monotonicity is clear in Figure 13; it is demonstrated rigorously with the following lemma:

**Lemma C.3** (Monotonicity of $\mathbb{E}\boldsymbol{\chi}(\boldsymbol{\beta})$)**.** *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $|\boldsymbol{\tau}|, c_\mu$ such that $(\boldsymbol{a}_0, \theta, |\boldsymbol{\tau}|)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{|\boldsymbol{\tau}|}$ in $\varphi_{\ell^1}$ where $c_\lambda \in \left[0, \frac{1}{4}\right]$, then there exists some numerical constant $\bar{c} > 0$, such that if $c_\mu < \bar{c}$, the expectation $|\mathbb{E}[\boldsymbol{\chi}[\boldsymbol{\beta}]]_i|$ is monotone increasing in $|\boldsymbol{\beta}_i|$. In other words, if $|\boldsymbol{\beta}_i| > |\boldsymbol{\beta}_j|$ then*

$$\sigma_i \mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i \geq \sigma_j \mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_j \tag{C.20}$$

*where $\sigma_i = \operatorname{sign}(\boldsymbol{\beta}_i)$.*

The proof first operate simple calculus and then followed by studying cases of $|\boldsymbol{\beta}_i| - |\boldsymbol{\beta}_j|$ when either it is smaller are larger then $\lambda$.

**Proof.** 1. (Monotonicity by gradient negativity) Wlog assume $\boldsymbol{\beta}_i > \boldsymbol{\beta}_j > 0$, and from Lemma C.2 we can write $(n\theta)^{-1}\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i = \boldsymbol{\beta}_i\left(1 - \mathbb{E}_{\boldsymbol{s}_i}\operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right)$. Consider $t \in [0,1]$ and define $\ell(t) = t\boldsymbol{\beta}_i - t\boldsymbol{\beta}_j$. Write the random variable $\boldsymbol{s}_{ij} = \sum_{\ell \neq i,j} \boldsymbol{\beta}_\ell \boldsymbol{x}_\ell$. Define $h$ as a function of $t$ such that

$$\begin{aligned} h(t) &= \mathbb{E}_{x,\boldsymbol{s}_{ij}}\left[\left((1-t)\boldsymbol{\beta}_i + t\boldsymbol{\beta}_j\right)\left(1 - \operatorname{erf}_{(1-t)\boldsymbol{\beta}_i + t\boldsymbol{\beta}_j}(\lambda, ((1-t)\boldsymbol{\beta}_j + t\boldsymbol{\beta}_i)x + \boldsymbol{s}_{ij})\right)\right] \\ &= \mathbb{E}_{x,\boldsymbol{s}_{ij}}\left[\left(\boldsymbol{\beta}_i - \ell(t)\right)\left(1 - \operatorname{erf}_{\boldsymbol{\beta}_i - \ell(t)}(\lambda, x \cdot (\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij})\right)\right]. \end{aligned} \tag{C.21}$$

Notice that $\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i = h(0)$ and $\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_j = h(1)$ respectively, thus it suffices to prove $h'(t) < 0$ for all $t \in [0,1]$. Write $f$ as pdf of standard Gaussian r.v. where

$$\operatorname{erf}_\beta(\lambda, \boldsymbol{s}_{ij}) = \int_0^{\frac{\lambda + \boldsymbol{s}_{ij}}{\beta}} f(z)\,dz + \int_0^{\frac{\lambda - \boldsymbol{s}_{ij}}{\beta}} f(z)\,dz,$$

and use chain rule:

$$\begin{aligned} h'(t) = \mathbb{E}_{x,\boldsymbol{s}_{ij}}\Big[&(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\left(1 - \operatorname{erf}_{\boldsymbol{\beta}_i - \ell(t)}(\lambda, x \cdot (\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij})\right) \\ &- (\boldsymbol{\beta}_i - \ell(t)) \cdot \frac{d}{dt}\left(\frac{\lambda + x \cdot (\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij}}{\boldsymbol{\beta}_i - \ell(t)}\right) \cdot f\left(\frac{\lambda + x \cdot (\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij}}{\boldsymbol{\beta}_i - \ell(t)}\right) \\ &- (\boldsymbol{\beta}_i - \ell(t)) \cdot \frac{d}{dt}\left(\frac{\lambda - x \cdot (\boldsymbol{\beta}_j + \ell(t)) - \boldsymbol{s}_{ij}}{\boldsymbol{\beta}_i - \ell(t)}\right) \cdot f\left(\frac{\lambda - x \cdot (\boldsymbol{\beta}_j + \ell(t)) - \boldsymbol{s}_{ij}}{\boldsymbol{\beta}_i - \ell(t)}\right)\Big] \\ = (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\mathbb{E}_{x,\boldsymbol{s}_{ij}}\Big[&1 - \operatorname{erf}_{\boldsymbol{\beta}_i - \ell(t)}(\lambda, x \cdot (\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij}) \\ &+ \left(\frac{\lambda + x(\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij}}{\boldsymbol{\beta}_i - \ell(t)} + x\right) \cdot f\underbrace{\left(\frac{\lambda + x(\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij}}{\boldsymbol{\beta}_i - \ell(t)}\right)}_{z_{\lambda_+}} \\ &+ \left(\frac{\lambda - x(\boldsymbol{\beta}_j + \ell(t)) - \boldsymbol{s}_{ij}}{\boldsymbol{\beta}_i - \ell(t)} - x\right) \cdot f\underbrace{\left(\frac{\lambda - x(\boldsymbol{\beta}_j + \ell(t)) - \boldsymbol{s}_{ij}}{\boldsymbol{\beta}_i - \ell(t)}\right)}_{z_{\lambda_-}}\Big] \\ = (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\mathbb{E}_{x,\boldsymbol{s}_{ij}}&\left[1 - \int_0^{z_{\lambda_+}} f(z)\,dz - \int_0^{z_{\lambda_-}} f(z)\,dz + (z_{\lambda_+} + x)f(z_{\lambda_+}) + (z_{\lambda_-} - x)f(z_{\lambda_-})\right]. \end{aligned} \tag{C.22}$$

Consider the term only related to $z_{\lambda_+}$, condition on cases that it is either positive or negative, observe that

$$\begin{cases} \mu_{+-} := \mathbb{E}_{x,\boldsymbol{s}_{ij}|z_{\lambda_+}\le 0}\left[\int_0^{z_{\lambda_+}} f(z)\,dz - z_{\lambda_+} f(z_{\lambda_+})\right] = \mathbb{E}_{x,s|z_{\lambda_+}\le 0}\left[-\int_0^{-z_{\lambda_+}} f(z)\,dz - z_{\lambda_+} f(z_{\lambda_+})\right] \le 0 \\ \mu_{++} := \mathbb{E}_{x,\boldsymbol{s}_{ij}|z_{\lambda_+}>0}\left[\int_0^{z_{\lambda_+}} f(z)\,dz - z_{\lambda_+} f(z_{\lambda_+})\right] \le \min\left\{\frac{1}{2}, \frac{1}{\sqrt{2\pi}}\mathbb{E}_{x,\boldsymbol{s}_{ij}|z_{\lambda_+}>0}\, z_{\lambda_+}\right\} \end{cases},$$

where the negativity of the first equation can be observed by writing $v = -z_{\lambda_+}$ and take derivative:

$$\begin{cases} -\int_0^v f(z)dz + v\cdot f(v) = 0 & v = 0 \\ \frac{d}{dv}\left\{-\int_0^v f(z)dz + v\cdot f(v)\right\} = -f(v) + f(v) + v\cdot f'(v) < 0 & v > 0 \end{cases};$$

and similarly for $z_{\lambda_-}$:

$$\begin{cases} \mu_{--} := \mathbb{E}_{x,\boldsymbol{s}_{ij}|z_{\lambda_-}\le 0}\left[\int_0^{z_{\lambda_-}} f(z)\,dz - z_{\lambda_-} f(z_{\lambda_-})\right] \le 0 \\ \mu_{-+} := \mathbb{E}_{x,\boldsymbol{s}_{ij}|z_{\lambda_-}>0}\left[\int_0^{z_{\lambda_-}} f(z)\,dz - z_{\lambda_-} f(z_{\lambda_-})\right] \le \min\left\{\frac{1}{2}, \frac{1}{\sqrt{2\pi}}\mathbb{E}_{x,\boldsymbol{s}_{ij}|z_{\lambda_-}>0} z_{\lambda_-}\right\} \end{cases},$$

then combine every term to (C.22) using tower property and from assumption $\boldsymbol{\beta}_j - \boldsymbol{\beta}_i < 0$ we obtain

$$\begin{aligned} \text{(C.22)} &\le (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\left(1 - \mathbb{P}\left[z_{\lambda_+} > 0\right]\cdot\mu_{++} - \mathbb{P}\left[z_{\lambda_-} > 0\right]\cdot\mu_{-+} + \mathbb{E}_{x,\boldsymbol{s}_{ij}}\left[x(f(z_{\lambda_+}) - f(z_{\lambda_-}))\right]\right) \\ &\le (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\left(1 - \min\left\{\frac{\mathbb{P}\left[z_{\lambda_+} > 0\right]}{2}, \frac{\mathbb{E}\left|z_{\lambda_+}\right|}{\sqrt{2\pi}}\right\} - \min\left\{\frac{\mathbb{P}\left[z_{\lambda_-} > 0\right]}{2}, \frac{\mathbb{E}\left|z_{\lambda_-}\right|}{\sqrt{2\pi}}\right\} - \frac{\theta}{\sqrt{2\pi}}\cdot\mathbb{E}\left|g\right|\right), \end{aligned} \tag{C.23}$$

where $g$ is standard Gaussian r.v..

2. (<u>Cases of varying $\boldsymbol{\beta}_i, \boldsymbol{\beta}_j$</u>) Let $c_\lambda < \frac{1}{4}$. Suppose $\boldsymbol{\beta}_i - \ell(t) \le \frac{1}{4\sqrt{|\boldsymbol{\tau}|}}$. Recall that $\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 \ge 1 - 3c_\mu$. We are going to show there is at least one of the entry $\boldsymbol{\beta}_* \in \{\boldsymbol{\beta}_r\}_{r\in\boldsymbol{\tau}\neq i,j} \uplus \{\boldsymbol{\beta}_j + \ell(t)\}$ is greater than $\frac{0.85}{\sqrt{|\boldsymbol{\tau}|}}$. First, if both $i, j \notin \boldsymbol{\tau}$, the lower bound is immediate since $\boldsymbol{\beta}_*^2 = \|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_\infty^2 > \frac{1-3c_\mu}{|\boldsymbol{\tau}|}$. On the other hand if at least one of $i, j$ is in $\boldsymbol{\tau}$ and all other $\boldsymbol{\beta}_{\boldsymbol{\tau}}$ entries are small where $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}\setminus\{i,j\}}\right\|_\infty^2 < \frac{1-3c_\mu}{|\boldsymbol{\tau}|}$, then we know via norm inequalities,

$$(\boldsymbol{\beta}_i + \boldsymbol{\beta}_j)^2 > \boldsymbol{\beta}_i^2 + \boldsymbol{\beta}_j^2 > \|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 - (|\boldsymbol{\tau}| - 1)\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}\setminus\{i,j\}}\right\|_\infty^2 \ge \frac{1-3c_\mu}{|\boldsymbol{\tau}|}, \tag{C.24}$$

which implies if $c_\mu < \frac{1}{100}$,

$$\boldsymbol{\beta}_* = \boldsymbol{\beta}_j + \ell(t) = (\boldsymbol{\beta}_i + \boldsymbol{\beta}_j) - (\boldsymbol{\beta}_i - \ell(t)) \ge \frac{\sqrt{1-3c_\mu}}{\sqrt{|\boldsymbol{\tau}|}} - \frac{1}{4\sqrt{|\boldsymbol{\tau}|}} \ge \frac{0.72}{\sqrt{|\boldsymbol{\tau}|}}. \tag{C.25}$$

In this case, adopt result from Corollary B.6 such that $\mathbb{P}\left[\left|\sum \boldsymbol{\beta}_\ell x_\ell\right| > \lambda/10\right] \le 3\theta\,|\boldsymbol{\tau}| \le .01$, we have

$$\begin{aligned} \mathbb{P}\left[z_{\lambda_-} > 0\right] = \mathbb{P}\left[z_{\lambda_+} > 0\right] &= 1 - \mathbb{P}\left[x(\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij} < -\lambda\right] \\ &\le 1 - \mathbb{P}\left[\boldsymbol{x}_*\boldsymbol{\beta}_* < -11\lambda/10\right]\cdot\mathbb{P}\left[x(\boldsymbol{\beta}_j + \ell(t)) + \boldsymbol{s}_{ij} - \boldsymbol{x}_*\boldsymbol{\beta}_* < \lambda/10\right] \\ &\le 1 - \theta\cdot\mathbb{P}\left[\boldsymbol{g}_*\cdot\frac{0.72}{\sqrt{|\boldsymbol{\tau}|}} < \frac{-11c_\lambda}{10\sqrt{|\boldsymbol{\tau}|}}\right]\cdot\left(1 - \mathbb{P}\left[\sum \boldsymbol{\beta}_\ell \boldsymbol{x}_\ell > \frac{\lambda}{10}\right]\right) \\ &\le 1 - \theta\cdot\mathbb{P}\left[0.72\cdot\boldsymbol{g}_* \le -1.1\cdot 0.25\right]\cdot(1 - 3c_\mu) \\ &\le 1 - 0.35\theta. \end{aligned} \tag{C.26}$$

On the other hand, when $\boldsymbol{\beta}_i - \ell(t) \ge \frac{1}{4\sqrt{|\boldsymbol{\tau}|}}$, both $z_{\lambda_+}$, $z_{\lambda_-}$ are upper bounded via $|\boldsymbol{\tau}|\,\theta \le \frac{1}{800}$ such as:

$$\mathbb{E}_{x,\boldsymbol{s}_{ij}}\left|z_{\lambda_-}\right| = \mathbb{E}_{x,\boldsymbol{s}_{ij}}\left|z_{\lambda_+}\right| \le \mathbb{E}_{x,\boldsymbol{s}_{ij}}\frac{\lambda + \left|x(\boldsymbol{\beta}_j + \ell(t)) - \boldsymbol{s}_{ij}\right|}{\boldsymbol{\beta}_i - \ell(t)} \le 1 + 4\sqrt{|\boldsymbol{\tau}|}\cdot\left(\mathbb{E}_{x,\boldsymbol{s}_{ij}}\left|x(\boldsymbol{\beta}_j + \ell(t)) - \boldsymbol{s}_{ij}\right|^2\right)^{1/2}$$

$$\leq 1 + 4\sqrt{|\boldsymbol{\tau}|\,\theta}\,\|\boldsymbol{\beta}\|_2 \leq 1 + 4\sqrt{|\boldsymbol{\tau}|\,\theta}\left(1 + c_\mu + \frac{c_\mu}{\sqrt{\theta}\,|\boldsymbol{\tau}|}\right) \leq 1.2. \tag{C.27}$$

Combine (C.23), (C.26) we have

$$h'(t) \leq (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\left(1 - 2 \cdot \frac{(1-0.35\theta)}{2} - \frac{\theta}{\sqrt{2\pi}} \cdot \sqrt{\frac{2}{\pi}}\right) \leq 0.03\theta(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i) < 0, \tag{C.28}$$

and combine (C.23), (C.27) and $\theta < c_\mu$ we have

$$h'(t) \leq (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\left(1 - 2 \cdot \frac{1.2}{\sqrt{2\pi}} - \frac{\theta}{\sqrt{2\pi}} \cdot \sqrt{\frac{2}{\pi}}\right) \leq 0.03(\boldsymbol{\beta}_j - \boldsymbol{\beta}_i) < 0, \tag{C.29}$$

which proves the monotonicity. ■

**Finite sample deviation of $\boldsymbol{\chi}$.** When the signal length of $\boldsymbol{y}$ is sufficiently large, operator $\boldsymbol{\chi}$ will be enough close to its expected value.

**Corollary C.4** (Finite sample deviation of $\boldsymbol{\chi}(\boldsymbol{\beta})$)**.** *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_{\ell^1}$ for some $c_\lambda > 1/5$, then there exists some numerical constants $C, c, \bar{c} > 0$, such that if $n \geq Cp^5\theta^{-2}\log p$ and $c_\mu \leq \bar{c}$, then with probability at least $1 - 3/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ and every $i \in [n]$, we have:*

$$\left|n^{-1}\boldsymbol{\chi}[\boldsymbol{\beta}]_i - n^{-1}\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i\right| \leq c\theta/p^{3/2}, \tag{C.30}$$

**Proof.** See Appendix I.1 ■

## D Euclidean Hessian as logic function in shift space

We can express the (pseudo) curvature (4.10) in direction $\boldsymbol{v} \in \mathbb{S}^{p-1}$ in terms of the correlation $\boldsymbol{\gamma} = \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{v}$ between $\boldsymbol{v}$ and $\boldsymbol{a}_0$, giving

$$\boldsymbol{v}^* \widetilde{\nabla}^2 \varphi_{\ell^1}(\boldsymbol{a}) \boldsymbol{v} = -\boldsymbol{\gamma}^* \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{P}_I \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{\gamma},$$

where

$$I(\boldsymbol{a}) = \operatorname{supp}\left( \mathcal{S}_\lambda \left[ \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{\iota} \boldsymbol{a} \right] \right) = \left\{ i \in [n] \middle| \left| \boldsymbol{x}_0 * \check{\boldsymbol{\beta}} \right|_i > \lambda \right\}. \tag{D.1}$$

The $i$-th diagonal entry of $\check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{P}_{I(\boldsymbol{a})} \check{\boldsymbol{C}}_{\boldsymbol{x}_0}$ is

$$-\boldsymbol{e}_i^* \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{P}_{I(\boldsymbol{a})} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{e}_i = -\left\| \boldsymbol{P}_{I(\boldsymbol{a})} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{e}_i \right\|_2^2 = -\left\| \boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0] \right\|_2^2, \tag{D.2}$$

which is the core component for us to study the curvature of objective $\varphi_{\ell^1}$. We illustrate the expectation of diagonal term of Hessian in Lemma D.2 and Corollary D.3, whose figure of visualized $\left\| \boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0] \right\|_2$ is shown in Figure 13. Lastly, we also prove the off-diagonal terms $\boldsymbol{e}_i^* \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{P}_{I(\boldsymbol{a})} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{e}_j$ of Hessian is likely inconsequential in calculation of curvature in Lemma D.4.

**Expectation of Hessian diagonals.** We expect the Hessian to have stronger negative component in the $s_i[\boldsymbol{a}_0]$ direction as $\left\| \boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0] \right\|_2^2$ becomes larger. This term can by tremendously simplified when $\boldsymbol{x}_0$ is very sparse: suppose all entries of its support $I_0$ are separated by at least $2p-1$ samples, then by implementing the definition of support from (D.1), we can derive

$$-\left\| \boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0] \right\|_2^2 = -\sum_{j \in I_0} \boldsymbol{x}_{0j}^2 \mathbf{1}_{\left\{ \left| \sum_\ell \boldsymbol{\beta}_\ell \boldsymbol{x}_{0(\ell+j-i)} \right| > \lambda \right\}} \underbrace{=}_{\text{sep.}} -\sum_{j \in I_0} \boldsymbol{g}_j^2 \mathbf{1}_{\{ |\boldsymbol{\beta}_i \boldsymbol{g}_j| > \lambda \}}, \tag{D.3}$$

where $\mathbf{1}$ is the indicator function and $\boldsymbol{g}_j$ are independent standard Gaussian r.v.s.. In expectation, the summands in (D.3) acts like a smoothed logic function on entry $\boldsymbol{\beta}_i$:

**Lemma D.1** (Gaussian smoothed indicator)**.** *Let $g \sim \mathcal{N}(0,1)$, then for any $b, s \in \mathbb{R}$ and $\lambda > 0$.*

$$\mathbb{E}_g \left[ g^2 \mathbf{1}_{\{|b \cdot g + s| > \lambda\}} \right] = 1 - \operatorname{erf}_b(\lambda, s) + f_b(\lambda, s), \tag{D.4}$$

*where*

$$f_b(\lambda, s) = \frac{1}{\sqrt{2\pi}} \left[ \left( \frac{\lambda + s}{|b|} \right) e^{-\frac{(\lambda+s)^2}{2b^2}} + \left( \frac{\lambda - s}{|b|} \right) e^{-\frac{(\lambda-s)^2}{2b^2}} \right]. \tag{D.5}$$

**Proof.** The proof can be derived via same calculation of integrals in Lemma C.1. ■

Although the definition (D.4) seems incomprehensible at first glance, we can actually interpret it as a smoothed indicator function which compares $|b|$ to the threshold $\sqrt{2/\pi}\lambda$. Once we assign $s = 0$, then we can see that $\mathbb{E} g^2 \mathbf{1}_{\{|b \cdot g| > \lambda\}}$ is be an increasing function of $|b|$. Moreover by assigning different values for $|b|$ we obtain:

$$\mathbb{E} g^2 \mathbf{1}_{\{|b \cdot g| > \lambda\}} \approx \begin{cases} 1, & |b| \approx 1 \\ 1/2, & |b| \approx \sqrt{2/\pi}\lambda \\ 0, & |b| \approx 0 \end{cases}. \tag{D.6}$$

Relate (D.6) to (D.3), when $|\boldsymbol{\beta}_i|$ is close to 1 then we expect $-\frac{1}{n\theta} \left\| \boldsymbol{P}_I s_{-i}[\boldsymbol{x}_0] \right\|_2^2$ to be close to $-1$, and it increases to 0 as $|\boldsymbol{\beta}_i|$ decreases, suggests that the Euclidean Hessian at point $\boldsymbol{a}$ has stronger negative component at $s_i[\boldsymbol{a}_0]$ direction if $|\langle \boldsymbol{a}, s_i[\boldsymbol{a}_0] \rangle|$ is larger. See Figure 14 for a numerical example. This phenomenon can be extend beyond the idealistic separating case as follows:

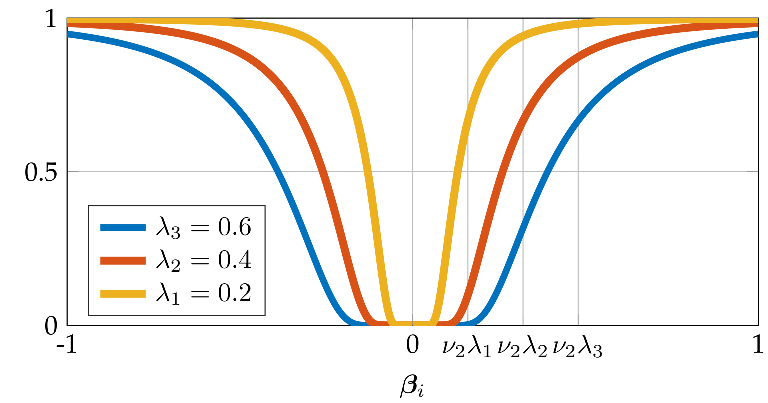


**Figure 14: A numerical example for $\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_i[\boldsymbol{x}_0]\right\|_2^2$.** We provide a figure to illustrate the expectation of $-\frac{1}{n\theta}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_i[\boldsymbol{x}_0]\right\|_2^2$ when entries of $\boldsymbol{x}_0$ are $2p$-separated, as a function plot of $\boldsymbol{\beta}_i \to 1 - \mathrm{erf}_{\boldsymbol{\beta}_i}(\lambda, 0) + f_{\boldsymbol{\beta}_i}(\lambda, 0)$ from (D.4) with different $\lambda$. When $|\boldsymbol{\beta}_i| \approx \nu_2\lambda$ where $\nu_2 = \sqrt{2/\pi}$, then the its function value is close to $0.5$. If $|\boldsymbol{\beta}_i|$ is much larger then $\lambda$ its value grow to 1, implies there is a negative curvature at $s_i[\boldsymbol{a}_0]$ direction. Similarly if $|\boldsymbol{\beta}_i|$ is much smaller then $\lambda$ the function value is 0 thus the curvature is positive in $s_i[\boldsymbol{a}_0]$ direction.

**Lemma D.2** (Expected Hessian diagonals)**.** *Let $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ and $\lambda > 0$, define the set $I(\boldsymbol{a})$ in (D.1), write $\boldsymbol{s}_i = \sum_{\ell\neq i}\boldsymbol{\beta}_\ell \boldsymbol{x}_{0\ell}$, then for every $\boldsymbol{a} \in \mathbb{S}^{p-1}$ and $i \in [n]$:*

$$n^{-1}\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0]\right\|_2^2 = \theta\left[1 - \mathbb{E}_{\boldsymbol{s}_i}\mathrm{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) + \mathbb{E}_{\boldsymbol{s}_i} f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right] \tag{D.7}$$

**Proof.** Write $\boldsymbol{x}_0$ as $\boldsymbol{x}$. Observe that $\boldsymbol{y} * \check{\boldsymbol{a}} = \boldsymbol{x}_0 * \check{\boldsymbol{\beta}} = \sum_\ell \boldsymbol{\beta}_\ell s_{-\ell}[\boldsymbol{x}_0]$. Thus for any $j \in [n]$ and $i \in [\pm p]$:

$$(\boldsymbol{y} * \check{\boldsymbol{a}})_{j-i} = \Big(\boldsymbol{\beta}_i s_{-i}[\boldsymbol{x}] + \sum_{\ell\neq i}\boldsymbol{\beta}_\ell s_{-\ell}[\boldsymbol{x}]\Big)_{j-i} = \boldsymbol{\beta}_i \boldsymbol{x}_j + \sum_{\ell\neq i}\boldsymbol{\beta}_\ell \boldsymbol{x}_{j+\ell-i} =: \boldsymbol{\beta}_i\boldsymbol{x}_j + \boldsymbol{s}_j, \tag{D.8}$$

where $\boldsymbol{x}_j$ is independent of $\boldsymbol{s}_j$, and both $\boldsymbol{x}_j, \boldsymbol{s}_j$ are symmetric and identically distributed for all $j \in [n]$. Rewrite the random variable using (D.1) as

$$\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0]\right\|_2^2 = \left\|\boldsymbol{P}_{I(\boldsymbol{a})}\textstyle\sum_{j\in[n]}(\boldsymbol{x}_{0j}\boldsymbol{e}_{j-i})\right\|_2^2 = \sum_{j\in[n]}\boldsymbol{x}_{0j}^2\mathbf{1}_{\left\{|\boldsymbol{y}*\check{\boldsymbol{a}}|_{j-i}>\lambda\right\}} = \sum_{j\in[n]}\boldsymbol{x}_{0j}^2\mathbf{1}_{\{|\boldsymbol{\beta}_i\boldsymbol{x}_{0j}+\boldsymbol{s}_j|>\lambda\}}$$

Write $\boldsymbol{x} = \boldsymbol{g}\circ\boldsymbol{\omega}$ as composition of Gaussian/Bernoulli r.v.s., the expectation has a simple form:

$$\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0]\right\|_2^2 = n\theta\cdot\mathbb{E}\boldsymbol{g}_0^2\mathbf{1}_{\{|\boldsymbol{\beta}_i\boldsymbol{g}_0+\boldsymbol{s}_0|>\lambda\}} = n\theta\cdot\mathbb{E}\left(1 - \mathrm{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) + f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right)$$

where $\boldsymbol{s}_i = \sum_{\ell\neq i}\boldsymbol{x}_{0i}\boldsymbol{\beta}_i$ with $\boldsymbol{x}_{0i} \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$, yielding the claimed expression. ■

**Finite sample deviation of Hessian diagonals.** When the signal length of $\boldsymbol{y}$ is sufficiently large, then $i$-th diagonal term for Hessian $\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0]\right\|_2^2$ will be close enough to its expected value.

**Corollary D.3** (Large sample deviation of curvature)**.** *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_{\ell^1}$ for some $c_\lambda > 1/5$, then there exists some numerical constant $C, c, \bar{c} > 0$, such that if $n \geq Cp^4\theta^{-1}\log p$ and $c_\mu \leq \bar{c}$, then with probability at least $1 - 3/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\leq k}\Re(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ and every $i \in [n]$, we have:*

$$\left|n^{-1}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0]\right\|_2^2 - n^{-1}\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}_0]\right\|_2^2\right| \leq c\theta/p \tag{D.9}$$

**Proof.** See Appendix I.2. ■

**Hessian off-diagonal terms near solution.** The off-diagonal entries of Hessian in general are much smaller then the diagonal entries; however, it affects the region near sign shifts of $\boldsymbol{a}_0$ the most where we need to show strong convexity in the region. We provide an upper bound for off-diagonal entries in the vicinity of signed shifts. In these regions, only one entry of the correlations $\left|\boldsymbol{\beta}_{(0)}\right|$ is large and the rest is small.

**Lemma D.4** (Hessian off-diagonal term near solution). *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Let $\lambda = c_\lambda/\sqrt{k}$ with $c_\lambda > 1/5$, then there exists some numerical constant $C, \bar{c} > 0$ such that if $n \geq C\theta^{-4}\log p$ and $c_\mu \leq \bar{c}$, then with probability at least $1 - 4/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\leq k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$, where $\left|\boldsymbol{\beta}_{(1)}\right| \leq \frac{1}{4\log\theta^{-1}}\lambda$ and every $i \neq j \in [\pm p] \setminus \{(0)\}$, we have*

$$\left|s_i[\boldsymbol{x}_0]^*\right| \boldsymbol{P}_{I(\boldsymbol{a})} \left|s_j[\boldsymbol{x}_0]\right| < 8n\theta^3 \tag{D.10}$$

**Proof.** Write $\theta_{\log} = -1/\log\theta$ and $\boldsymbol{x}_0$ as $\boldsymbol{x} = \boldsymbol{\omega} \circ \boldsymbol{g}$. Wlog let $\boldsymbol{\beta}_0$ be the largest correlation $\boldsymbol{\beta}_{(0)}$. Define random variables $s' = \left\langle \boldsymbol{\beta}_{\boldsymbol{\tau}\setminus\{0,i,j\}}, \boldsymbol{x}_{\boldsymbol{\tau}\setminus\{0,i,j\}} \right\rangle$. Firstly via Corollary B.7 we have $\mathbb{P}\left[|s'| > 0.4\lambda\right] \leq 2\theta$; also define $s = \left\langle \boldsymbol{\beta}_{\boldsymbol{\tau}^c\setminus\{0,i,j\}}, \boldsymbol{x}_{\boldsymbol{\tau}^c\setminus\{0,i,j\}} \right\rangle$, and base on Corollary B.6 we have $\mathbb{P}\left[|s| > \lambda/10\right] \leq 2\theta$. Expand the $(-i,-j)$-th cross term with $\theta < 0.1$ we have:

$$\begin{aligned}
\mathbb{E}\left|s_{-i}[\boldsymbol{x}]^*\right| \boldsymbol{P}_{I(\boldsymbol{a})} \left|s_{-j}[\boldsymbol{x}]\right| &= \mathbb{E}\textstyle\sum_{k\in[n]} \left|\boldsymbol{x}_{k+i}\boldsymbol{x}_{k+j}\right| \mathbf{1}_{\{|\boldsymbol{\beta}_0\boldsymbol{x}_k + \boldsymbol{\beta}_i\boldsymbol{x}_{k+i} + \boldsymbol{\beta}_j\boldsymbol{x}_{k+j} + s + s'| > \lambda\}} \\
&= n\theta^2 \cdot \mathbb{E}\left|\boldsymbol{g}_i\boldsymbol{g}_j\right| \mathbf{1}_{\{|\boldsymbol{\beta}_0\boldsymbol{x}_0 + \boldsymbol{\beta}_i\boldsymbol{g}_i + \boldsymbol{\beta}_j\boldsymbol{g}_j + s + s'| > \lambda\}} \\
&\leq n\theta^2 \cdot \mathbb{E}\left[\left|\boldsymbol{g}_i\boldsymbol{g}_j\right| \left(2\mathbf{1}_{\{|\boldsymbol{\beta}_i\boldsymbol{g}_i| > \lambda/4\}} + \mathbb{P}\left[\boldsymbol{x}_0 \neq 0\right] + \mathbb{P}\left[|s| > 0.1\lambda\right] + \mathbb{P}\left[|s'| > 0.4\lambda\right]\right)\right] \\
&\leq n\theta^2 \cdot \left(\exp\left(-\log^2\theta^{-1}\right) + \theta + 2\theta + 2\theta\right) \\
&\leq 6n\theta^3.
\end{aligned} \tag{D.11}$$

Write (D.10) as two summation of independent random variables with $t = j - i$ by separating sum into two sets $J_{t1}, J_{t2}$ defined in (A.4) with both $\left|J_{t1}\right|, \left|J_{t2}\right| < n\theta^2$ with probability at least $1 - 2/n$ from Lemma A.1

$$\mathbb{E}\left|s_{-i}[\boldsymbol{x}]^*\right| \boldsymbol{P}_{I(\boldsymbol{a})} \left|s_{-j}[\boldsymbol{x}]\right| = \sum_{(k-i)\in I(\boldsymbol{a})} \left|\boldsymbol{x}_k\right| \left|\boldsymbol{x}_{k+t}\right| = \sum_{(k-i)\in I(\boldsymbol{a})\cap J_{t1}} \left|\boldsymbol{g}_k\right| \left|\boldsymbol{g}_{k+t}\right| + \sum_{(k-i)\in I(\boldsymbol{a})\cap J_{t2}} \left|\boldsymbol{g}_k\right| \left|\boldsymbol{g}_{k+t}\right|,$$

whose first summands can be upper bounded w.h.p. via Bernstein inequality Lemma J.4 with $(\sigma^2, R) = (1, 1)$ and writes $\mathcal{C} := \cup_{|\boldsymbol{\tau}|\leq k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu)) \cap \left\{\boldsymbol{a} \;\middle|\; \left|\boldsymbol{\beta}_{(1)}\right| \leq \frac{1}{4\log\theta^{-1}}\lambda\right\}$, then we have

$$\begin{aligned}
&\mathbb{P}\left[\max_{\substack{i\neq j\in[\pm p]\setminus\{0\} \\ \boldsymbol{a}\in\mathcal{C}}} \left(\sum_{(k-i)\in I(\boldsymbol{a})\cap J_{t1}} \left|\boldsymbol{g}_k\right| \left|\boldsymbol{g}_{k+t}\right| - \mathbb{E}\sum_{(k-i)\in I(\boldsymbol{a})\cap J_{t1}} \left|\boldsymbol{g}_k\right| \left|\boldsymbol{g}_{k+t}\right|\right) \geq n\theta^3\right] \\
&\mathbb{P}\left[\max_{i\neq j\in[\pm p]\setminus\{0\}} \left(\sum_{(k-i)\in\cap J_{t1}} \left|\boldsymbol{g}_k\right| \left|\boldsymbol{g}_{k+t}\right| - \mathbb{E}\sum_{(k-i)\cap J_{t1}} \left|\boldsymbol{g}_k\right| \left|\boldsymbol{g}_{k+t}\right|\right) \geq n\theta^3\right] \\
&\leq 4p^2 \cdot \exp\left(\frac{-n^2\theta^6}{2\left|J_{t1}\right| + 2n\theta^3}\right) \leq \exp\left(8\log p - \frac{-n^2\theta^6}{3n\theta^2}\right) \leq \exp\left(-\frac{n\theta^4}{10}\right) \leq \frac{1}{n}
\end{aligned} \tag{D.12}$$

when $n = C\theta^{-4}\log p$ with $C > 10^4$ and $\theta\log^2\theta^{-1} \geq 1/p$. Thus for all $i \neq j \in [\pm p] \setminus \{0\}$ and $\boldsymbol{a}$ satisfies our condition of lemma, from (D.11) and (D.12) we can conclude :

$$\left|s_{-i}[\boldsymbol{x}]^*\right| \boldsymbol{P}_{I(\boldsymbol{a})} \left|s_{-j}[\boldsymbol{x}]\right| \leq \sum_{I(\boldsymbol{a})\cap J_{t1}} \mathbb{E}\left|\boldsymbol{g}_k\right| \left|\boldsymbol{g}_{k+t}\right| + \sum_{I(\boldsymbol{a})\cap J_{t2}} \mathbb{E}\left|\boldsymbol{g}_k\right| \left|\boldsymbol{g}_{k+t}\right| + 2n\theta^3 \leq 8n\theta^3$$

which holds with probability at least $1 - 2/n - 2 \cdot 1/n = 1 - 4/n$ base on Lemma A.1 and (D.12). ■

# E Geometric relation between $\rho$ and $\ell^1$-norm

In this section, we discuss how to ensure that the smooth sparsity surrogate $\rho$ approximates $\|\cdot\|_1$ accurately enough that guarantees $\varphi_\rho$ inherits the good properties of $\varphi_{\ell^1}$. We prove several lemmas which allow us to transfer properties of $\varphi_{\ell^1}$ to $\varphi_\rho$. Our result does not pertain to the suggested pseudo-Huber surrogate $\rho(x)_i = \sqrt{x_i^2 + \delta^2}$ in the main script, and is general for a class of function class defined in Definition E.2 that is smooth and well approximates $\ell^1$ when the proper smoothing parameter $\delta$ is chosen from the result of Lemma E.6. In particular we ask the regularizer $\rho_\delta(x)$ to be uniformly bounded to $|x|$ by $\delta/2$:

$$\forall\, x \in \mathbb{R}, \qquad |\rho_\delta(x) - |x|| \le \delta/2 \tag{E.1}$$

then if $\delta \to 0$ we have for every $\boldsymbol{a}$ near subspace,

$$\left\|\operatorname{prox}_{\lambda\ell^1}[\check{\boldsymbol{a}} * \boldsymbol{y}] - \operatorname{prox}_{\lambda\rho_\delta}[\check{\boldsymbol{a}} * \boldsymbol{y}]\right\|_2 \to 0, \tag{E.2}$$

$$\|\nabla\varphi_{\ell^1}(\boldsymbol{a}) - \nabla\varphi_{\rho_\delta}(\boldsymbol{a})\|_2 \to 0, \tag{E.3}$$

$$\|\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a}) - \nabla^2\varphi_{\rho_\delta}(\boldsymbol{a})\|_2 \to 0. \tag{E.4}$$

An example choices of eligible smooth sparse surrogate is demonstrated in Table 1.

**Calculus of $\varphi_\rho$.** The marginal minimizer over $\boldsymbol{x}$ in (2.7) can be expressed in terms of the proximal operator [BC11] of $\rho$ at point $\check{\boldsymbol{a}} * \boldsymbol{y}$:

$$\operatorname{prox}_{\lambda\rho}[\check{\boldsymbol{a}} * \boldsymbol{y}] = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\lambda\rho(\boldsymbol{x}) + \tfrac{1}{2}\|\boldsymbol{x}\|_2^2 - \langle \boldsymbol{a} * \boldsymbol{x}, \boldsymbol{y}\rangle\right\}.$$

Plugging in, we obtain

$$\varphi_\rho(\boldsymbol{a}) = \lambda\rho\big(\operatorname{prox}_{\lambda\rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]\big) + \tfrac{1}{2}\left\|\check{\boldsymbol{a}} * \boldsymbol{y} - \operatorname{prox}_{\lambda\rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]\right\|_2^2 - \tfrac{1}{2}\|\check{\boldsymbol{a}} * \boldsymbol{y}\|_2^2 + \tfrac{1}{2}\|\boldsymbol{y}\|_2^2 \tag{E.5}$$

The objective function $\varphi_\rho(\boldsymbol{a})$ is a differentiable function of $\boldsymbol{a}$. This can be seen, e.g., by noting that

$$\varphi_\rho(\boldsymbol{a}) = \epsilon(\lambda\rho)(\check{\boldsymbol{a}} * \boldsymbol{y}) - \tfrac{1}{2}\|\check{\boldsymbol{a}} * \boldsymbol{y}\|_2^2 + \tfrac{1}{2}\|\boldsymbol{y}\|_2^2, \tag{E.6}$$

where $\epsilon(g)(\boldsymbol{z}) = g\left(\operatorname{prox}_g(\boldsymbol{z})\right) + \frac{1}{2}\left\|\boldsymbol{z} - \operatorname{prox}_g(\boldsymbol{z})\right\|_2^2$ is the *Moreau envelope* of a function $g$. The Moreau envelope is differentiable:

**Fact E.1** (Derivative of Moreau envelope, [BC11], Prop.12.29)**.** *Let $f$ be a proper lower semicontinuous convex function and $\lambda > 0$ then the Moreau envelope $\epsilon(\lambda f)(\boldsymbol{z}) = \lambda f(\operatorname{prox}_{\lambda f}[\boldsymbol{z}]) + \frac{1}{2}\left\|\boldsymbol{z} - \operatorname{prox}_{\lambda f}[\boldsymbol{z}]\right\|_2^2$ is Fréchet differentiable with $\nabla\epsilon(\lambda f)(\boldsymbol{z}) = \boldsymbol{z} - \operatorname{prox}_{\lambda\rho}[\boldsymbol{z}]$.*

Furthermore, $\varphi_\rho$ is twice differentiable whenever $\operatorname{prox}_{\lambda\rho}$ is differentiable. In this case, the (Euclidean) gradient and hessian of $\varphi_\rho$ are given by

$$\nabla\varphi_\rho(\boldsymbol{a}) = -\boldsymbol{\iota}^*\check{\boldsymbol{C}}_{\boldsymbol{y}} \operatorname{prox}_{\lambda\rho}\left[\check{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\right], \tag{E.7}$$

$$\nabla^2\varphi_\rho(\boldsymbol{a}) = -\boldsymbol{\iota}^*\check{\boldsymbol{C}}_{\boldsymbol{y}}\nabla \operatorname{prox}_{\lambda\rho}\left[\check{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\right]\check{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}. \tag{E.8}$$

The Riemannian gradient and hessian over $\mathbb{S}^{p-1}$ are

$$\operatorname{grad}[\varphi_\rho](\boldsymbol{a}) = -\boldsymbol{P}_{\boldsymbol{a}^\perp}\boldsymbol{\iota}^*\check{\boldsymbol{C}}_{\boldsymbol{y}} \operatorname{prox}_{\lambda\rho}\left[\check{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\right], \tag{E.9}$$

$$\operatorname{Hess}[\varphi_\rho](\boldsymbol{a}) = -\boldsymbol{P}_{\boldsymbol{a}^\perp}\left(\boldsymbol{\iota}^*\check{\boldsymbol{C}}_{\boldsymbol{y}}\nabla \operatorname{prox}_{\lambda\rho}\left[\check{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\right]\check{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota} - \langle\nabla\varphi_\rho(\boldsymbol{a}), \boldsymbol{a}\rangle\,\boldsymbol{I}\right)\boldsymbol{P}_{\boldsymbol{a}^\perp}. \tag{E.10}$$

| Surrogate class | $\rho_i(x)$ | $\nabla\rho_i(x)$ | $\nabla^2\rho_i(x)$ |
|---|---|---|---|
| Log hyperbolic cosine | $\frac{\delta}{2}\log\left(e^{2x/\delta}+e^{-2x/\delta}\right)$ | $\frac{e^{4x/\delta}-1}{e^{4x/\delta}+1}$ | $\frac{4e^{4x/\delta}}{\delta(e^{4x/\delta}+1)^2}$ |
| Pseudo Huber | $\sqrt{x^2+\delta^2}$ | $\frac{x}{\sqrt{x^2+\delta^2}}$ | $\frac{\delta^2}{(x^2+\delta^2)^{3/2}}$ |
| Gaussian convolution | $\int \lvert x-t\rvert\, f_\delta(t)dt$ | $\operatorname{erf}(x/\sqrt{2}\delta)$ | $2f_\delta(x)$ |

**Table 1: Classes of smooth sparse surrogate $\rho$ and how to set its parameter.** Three common classes are listed with parameter $\delta$ to tune the smoothness. All the listed functions are greater then $|x|$ pointwise and has largest distance to $|x|$ at origin where $\rho(0)-|x|\le\delta$, satisfies the condition (E.11). Also its second order derivatives $\nabla^2\rho_i(x)$ are monotone decreasing w.r.t. $|x|$, hence are certified to be eligible $\delta$-smoothed $\ell^1$ surrogates.

**Sparse regularizer $\rho$ as smoothed $\ell^1$ function.** Our analysis accommodates any sufficiently accurate smooth approximation $\rho$ to the $\ell^1$ function. The requisite sense of approximation is captured in the following definition:

**Definition E.2** ($\delta$-smoothed $\ell^1$ function)**.** *We call an additively separable function $\rho(\boldsymbol{x})=\sum_{i=1}^n\rho_i(\boldsymbol{x}_i):\mathbb{R}^n\to\mathbb{R}$, a $\delta$-smoothed $\ell^1$ function with $\delta>0$ if for each $i\in[n]$, $\rho_i$ is even, convex, twice differentiable and $\nabla^2\rho_i(x)$ being monotone decreasing w.r.t. $|x|$, where, there exists some constant $c$, such that for all $x\in\mathbb{R}$:*

$$|\rho_i(x)-|x|+c|\le\delta/2 \tag{E.11}$$

The proximal operator of the $\ell^1$ norm is the entrywise soft thresholding function $\mathcal{S}_\lambda$; the proximal operator associated to a smoothed $\ell^1$ function turns out to be a differentiable approximation to $\mathcal{S}_\lambda$. In particular, we will show that it approximates $\mathcal{S}_\lambda$ in the following sense:

**Definition E.3** ($\sqrt{\delta}$-smoothed soft threshold)**.** *An odd function $\mathcal{S}_\lambda^\delta[\cdot]:\mathbb{R}\to\mathbb{R}$ is a $\sqrt{\delta}$-smoothed soft thresholding function with parameter $\delta>0$ if it is a strictly monotone odd function and is differentiable everywhere, whose function value satisfies*

$$0\le\operatorname{sign}(z)\left(\mathcal{S}_\lambda^\delta[z]-\mathcal{S}_\lambda[z]\right)\le\sqrt{\lambda\delta},\qquad\forall z\in\mathbb{R} \tag{E.12}$$

*and its derivative satisfies for any given $B\in(0,\lambda)$:*

$$\left|\nabla\mathcal{S}_\lambda^\delta[z]-\nabla\mathcal{S}_\lambda[z]\right|\le\sqrt{\lambda\delta}/B,\qquad||z|-\lambda|\ge B. \tag{E.13}$$

If $\rho$ is a $\delta$-smooth $\ell^1$ function, then for all $i\in[n]$, we have that $\operatorname{prox}_{\lambda\rho}[\boldsymbol{z}]_i$ is a $\sqrt{\delta}$-smoothed soft threshold function of $\boldsymbol{z}_i$. This can be proven with the following lemma:

**Lemma E.4** (Proximal operator for smoothed $\ell^1$)**.** *Suppose $\rho$ is a $\delta$-smoothed $\ell^1$ function, then $\boldsymbol{z}_i\mapsto\operatorname{prox}_{\lambda\rho}[\boldsymbol{z}]_i$ is a $\sqrt{\delta}$-smoothed soft threshold function.*

**Proof.** We know that

$$\boldsymbol{x}_z:=\operatorname{prox}_{\lambda\rho}[\boldsymbol{z}]=\operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n}\lambda\rho(\boldsymbol{x})+\tfrac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2. \tag{E.14}$$

This optimization problem is strongly convex, and so the minimizer $\boldsymbol{x}_z$ is unique. Using the stationarity condition and since $\rho$ is separable, for all $i\in[n]$, we have $\lambda\nabla\rho_i(\boldsymbol{x}_{zi})+\boldsymbol{x}_{zi}-\boldsymbol{z}_i=0$, implies

$$\boldsymbol{x}_{zi}=(\mathrm{Id}+\lambda\nabla\rho_i)^{-1}(\boldsymbol{z}_i). \tag{E.15}$$

Since $\rho_i$ is convex and even , $\nabla \rho_i$ is monotone increasing and odd. By inverse function theorem, we know that strict monotonicity and differentiability of $\mathrm{Id} + \lambda \nabla \rho_i$ implies its inverse is differentiable and is a strictly monotone increasing odd function. Furthermore, it implies $\nabla \boldsymbol{x}_{zi}$ has the form

$$\nabla \boldsymbol{x}_{zi} = \nabla_i (\mathrm{Id} + \lambda \nabla \rho_i)^{-1}(\boldsymbol{z}_i) = \frac{1}{\lambda \nabla^2 \rho_i(\boldsymbol{x}_{zi}) + 1} < 1. \tag{E.16}$$

Notice that since $\nabla^2 \rho_i(x)$ is monotone decreasing when $x \geq 0$, hence $\nabla \boldsymbol{x}_{zi}$ is monotone increasing in $\boldsymbol{z}_i \geq 0$.

Now we are left to show that (E.12) and (E.13) hold, and since $\mathrm{prox}_{\lambda\rho}[\cdot]_i$ is an odd function it suffices to consider the case when the input vector $\boldsymbol{z}_i$ is nonnegative. Firstly, via convexity and entrywise bounded difference $|\rho_i(x) - |x|| \leq \delta/2$ we are going to show

$$|\nabla \rho_i(x)| \leq 1 \quad \forall x \in \mathbb{R}, \qquad \nabla \rho_i(x) \geq 1 - \sqrt{\delta/\lambda} \quad \forall x \geq \sqrt{\lambda\delta}. \tag{E.17}$$

Consider a positive $x$ with $\nabla \rho_i(x) > 1 + \varepsilon$ for some $\varepsilon > 0$, by convexity if $\widetilde{x} > x$ then $\nabla \rho_i(\widetilde{x}) > 1 + \varepsilon$, hence

$$\rho_i(x + \delta/\varepsilon) \geq \rho_i(x) + \nabla \rho_i(x) \cdot (\delta/\varepsilon) > x - \delta/2 + (1+\varepsilon) \cdot (\delta/\varepsilon) = (x + \delta/\varepsilon) + \delta/2,$$

contradicts the boundedness condition. Secondly, use mean value theorem we know for all $x \geq \sqrt{\lambda\delta}$:

$$\nabla \rho_i(x) \geq \frac{\rho_i(\sqrt{\lambda\delta}) - \rho_i(0)}{\sqrt{\lambda\delta} - 0} \geq \frac{(\sqrt{\lambda\delta} - \delta/2) - (0 + \delta/2)}{\sqrt{\lambda\delta} - 0} \geq 1 - \sqrt{\frac{\delta}{\lambda}}.$$

To prove (E.12), when $0 \leq \boldsymbol{z}_i \leq \lambda$, then $\mathcal{S}_\lambda[\boldsymbol{z}_i] = 0$ and $\boldsymbol{x}_{zi} \leq \sqrt{\lambda\delta}$ since if $\boldsymbol{x}_{zi} > \sqrt{\lambda\delta}$, by (E.17):

$$\lambda \nabla \rho_i(\boldsymbol{x}_{zi}) + \boldsymbol{x}_{zi} > \lambda(1 - \sqrt{\delta/\lambda}) + \sqrt{\lambda\delta} = \lambda \geq \boldsymbol{z}_i$$

then $\boldsymbol{x}_{zi}$ violate the stationary condition in (E.15), resulting $0 \leq \boldsymbol{x}_{zi} - \mathcal{S}_\lambda[\boldsymbol{z}_i] \leq \sqrt{\lambda\delta}$ whenever $0 \leq \boldsymbol{z}_i \leq \lambda$. Likewise in the case of $\boldsymbol{z}_i \geq \lambda$ where $\mathcal{S}_\lambda[\boldsymbol{z}_i] = \boldsymbol{z}_i - \lambda$, (E.17) provides:

$$\begin{cases} \forall\, \boldsymbol{x}_{zi} > \boldsymbol{z}_i - \lambda + \sqrt{\lambda\delta}, & \lambda \nabla \rho_i(\boldsymbol{x}_{zi}) + \boldsymbol{x}_{zi} > \lambda(1 - \sqrt{\delta/\lambda}) + \boldsymbol{z}_i - \lambda + \sqrt{\lambda\delta} = \boldsymbol{z}_i \\ \forall\, \boldsymbol{x}_{zi} < \boldsymbol{z}_i - \lambda, & \lambda \nabla \rho_i(\boldsymbol{x}_{zi}) + \boldsymbol{x}_{zi} < \lambda + \boldsymbol{z}_i - \lambda = \boldsymbol{z}_i \end{cases}$$

again violates (E.15) and therefore (E.12) holds for all $\boldsymbol{z}_i \in \mathbb{R}$.

Lastly (E.13) is a direct result of (E.12). For all $\boldsymbol{z}_i \leq \lambda - B$, recall that $\nabla \boldsymbol{x}_{zi}$ is monotone increasing in $\boldsymbol{z}_i$:

$$\nabla \boldsymbol{x}_{zi} \leq \min_{y \in [\lambda - B, \lambda]} \nabla \boldsymbol{x}_{yi} \leq \frac{\boldsymbol{x}_{\lambda i} - \boldsymbol{x}_{(\lambda - B)i}}{\lambda - (\lambda - B)} \leq \frac{(\sqrt{\lambda\delta} + \mathcal{S}_\lambda[\lambda]) - \mathcal{S}_\lambda[\lambda - B]}{B} = \frac{\sqrt{\lambda\delta}}{B};$$

and similarly for all $\boldsymbol{z}_i > \lambda + B$:

$$\nabla \boldsymbol{x}_{zi} \geq \max_{y \in [\lambda, \lambda + B]} \nabla \boldsymbol{x}_{yi} \geq \frac{\boldsymbol{x}_{(\lambda + B)i} - \boldsymbol{x}_{\lambda i}}{(\lambda + B) - \lambda} \geq \frac{\mathcal{S}_\lambda[\lambda + B] - (\mathcal{S}_\lambda[\lambda] + \sqrt{\lambda\delta})}{B} = 1 - \frac{\sqrt{\lambda\delta}}{B},$$

implies (E.13) holds. ■

**Approximate geometry of $\varphi_\rho$ using $\varphi_{\ell^1}$** Based on (E.9)-(E.10) and denote $\breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{\iota} \boldsymbol{a} = \breve{\boldsymbol{a}} * \boldsymbol{y}$, the only differences of Riemannian gradient and Hessian between $\varphi_\rho$ and $\varphi_{\ell^1}$ comes from the difference of $\mathrm{prox}_{\lambda\rho}[\breve{\boldsymbol{a}} * \boldsymbol{y}]$ and $\mathrm{prox}_{\lambda\|\cdot\|_1}[\breve{\boldsymbol{a}} * \boldsymbol{y}]$. Thus for the purpose of obtaining good geometric approximation of $\varphi_\rho$ with that of objective $\varphi_{\ell^1}$, we may apply both Definition E.3 and Lemma E.4, together suggest if $\rho$ is a $\delta$-smoothed $\ell^1$ function, then the $i$-th entry of $\mathrm{prox}_{\lambda\rho}[\breve{\boldsymbol{a}} * \boldsymbol{y}]$ will be $\sqrt{\lambda\delta}$-close to the authentic soft thresholding function $\mathcal{S}_\lambda[\breve{\boldsymbol{a}} * \boldsymbol{y}]_i$, and its gradient $\nabla\, \mathrm{prox}_{\lambda\rho}[\breve{\boldsymbol{a}} * \boldsymbol{y}]$ is $\sqrt{\lambda\delta}/B$-close to $\nabla \mathcal{S}_\lambda[\breve{\boldsymbol{a}} * \boldsymbol{y}]$ as long as $(\breve{\boldsymbol{a}} * \boldsymbol{y})_i$ is not close to $\pm\lambda$ by distance $B$.

Firsly, we will show by utilizing the random structure of $\boldsymbol{y}$, such that with high probability, only a fraction of entries of $\breve{\boldsymbol{a}} * \boldsymbol{y}$ will be close to $\pm\lambda$.

**Lemma E.5** (Gradients discontinuity entries). *For each $\boldsymbol{a} \in \mathbb{S}^{p-1}$, let*

$$J_B(\boldsymbol{a}) := \left\{ i \;\middle|\; \left( \check{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{\iota} \boldsymbol{a} \right)_i \in [-\lambda - B, -\lambda + B] \cup [\lambda - B, \lambda + B] \right\}. \tag{E.18}$$

*Suppose the subspace dimension is at most $k$ and signal $\boldsymbol{y}$ satisfies Definition B.1. Let $\lambda = c_\lambda/\sqrt{k}$ and $B \leq c'\lambda\theta^2/p\log n$ for some $c_\lambda, c' \in (0,1)$, then there is a numerical constant $C > 0$ such that if $n \geq Cp^5\theta^{-2}\log p$, then with probability at least $1 - 3/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$, we have*

$$|J_B(\boldsymbol{a})| \leq \frac{24c'n\theta^2}{p\log n} \tag{E.19}$$

**Proof.** See Appendix I.3. ∎

The geometric approximation between $\varphi_{\ell^1}$ and $\varphi_\rho$ necessarily consists of three parts: the gradient, the Hessian, and the coefficients. Here we conclude the approximation result with the following lemma:

**Lemma E.6** ($\varphi_{\ell^1}$ approximates $\varphi_\rho$). *Suppose $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Let $\rho \in \mathbb{R}^n \to \mathbb{R}$ be a $\delta$-smoothed $\ell^1$ function with*

$$\lambda = \frac{c_\lambda}{\sqrt{k}}, \qquad \delta \leq \frac{c'^4\theta^8}{p^2\log^2 n}\lambda \tag{E.20}$$

*with some $c'$, $c_\lambda \in (0,1)$, then there is a numerical constant $C, \bar{c} > 0$ such that if $n > Cp^5\theta^{-2}\log p$ and $c_\mu \leq \bar{c}$, then with probability at least $1 - 10/n$, the following statements hold simultaneously for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$:*

*(1). The coefficients has norm difference*

$$\left\| \boldsymbol{\iota}^*_{[\pm p]} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \operatorname{prox}_{\lambda\ell^1}[\check{\boldsymbol{a}} * \boldsymbol{y}] - \boldsymbol{\iota}^*_{[\pm p]} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \operatorname{prox}_{\lambda\rho}[\check{\boldsymbol{a}} * \boldsymbol{y}] \right\|_2 \leq c'n\theta^4. \tag{E.21}$$

*(2). The gradient has norm difference*

$$\left\| \nabla\varphi_{\ell^1}(\boldsymbol{a}) - \nabla\varphi_\rho(\boldsymbol{a}) \right\|_2 \leq c'n\theta^4. \tag{E.22}$$

*(3). The (pesudo) Riemmannian curvature difference is bounded in all directions $\boldsymbol{v} \in \mathbb{S}^{p-1}$ via*

$$\forall\, \boldsymbol{v} \in \mathbb{S}^{p-1}, \qquad \left| \boldsymbol{v}^* \left( \widetilde{\mathrm{Hess}}[\varphi_{\ell^1}](\boldsymbol{a}) - \mathrm{Hess}[\varphi_\rho](\boldsymbol{a}) \right) \boldsymbol{v} \right| \leq 200c'n\theta^2. \tag{E.23}$$

**Proof.** 1. (Coefficients) From Lemma E.4, the proximal $\delta$-smoothed $\ell^1$ function satisfies

$$\left| \mathcal{S}_\lambda[\check{\boldsymbol{a}} * \boldsymbol{y}] - \mathcal{S}^\delta_\lambda[\check{\boldsymbol{a}} * \boldsymbol{y}] \right|_j < \sqrt{\lambda\delta} \qquad \forall j \in [n].$$

Since the support of coefficient vectors are contained in $[\pm p]$, using simple norm inequality:

$$\left\| \boldsymbol{\iota}^*_{[\pm p]} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda[\check{\boldsymbol{a}} * \boldsymbol{y}] - \boldsymbol{\iota}^*_{[\pm p]} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}^\delta_\lambda[\check{\boldsymbol{a}} * \boldsymbol{y}] \right\|_2 \leq \sqrt{\lambda\delta n} \cdot \left\| \boldsymbol{\iota}^*_{[\pm p]} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \right\|_2. \tag{E.24}$$

Apply Lemma A.5 by replacing $\boldsymbol{a}_0$ with standard basis $\boldsymbol{e}_0$ and extend support of $\boldsymbol{\iota}$ to $\boldsymbol{\iota}_{[\pm p]}$, notice that in this case we have $\mu = 0$. Condition on the event

$$\left\| \boldsymbol{\iota}^*_{[\pm p]} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \right\|_2 \leq \left\| \boldsymbol{\iota}^*_{[\pm p]} \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \boldsymbol{C}^*_{\boldsymbol{e}_0} \right\|_2 \leq \sqrt{3(1+2\mu p)n\theta} \leq \sqrt{3n\theta},$$

and we gain

$$\text{(E.24)} \leq \sqrt{\lambda\delta n} \cdot \sqrt{3n\theta} \leq n\sqrt{3\lambda\theta\delta} \leq c'n\theta^4.$$

2. (Gradient) From definition of Riemannian gradient (E.9) and apply similar norm bound of (E.24), and condition on the following events of Lemma A.5 holds, obtain

$$\|\nabla\varphi_{\ell^1}(\boldsymbol{a})-\nabla\varphi_\rho(\boldsymbol{a})\|_2\le\sqrt{\lambda\delta n}\cdot\left\|\boldsymbol{\iota}^*\breve{\boldsymbol{C}}_{\boldsymbol{y}}\right\|_2\le n\sqrt{3\lambda\theta(1+\mu p)\delta}\le c'n\theta^4. \tag{E.25}$$

3. (Hessian) For every realization of $J_B(\boldsymbol{a})$ from $\boldsymbol{a}\in\cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}},\gamma(c_\mu))$, base on Lemma E.5, condition on the event such that

$$B\le\frac{c'\lambda\theta^2}{p\log n},\qquad |J|\le\frac{24c'n\theta^2}{p\log n}; \tag{E.26}$$

and rewrite $J_B(\boldsymbol{a})$ as $J$. Also condition on the event using Lemma A.5 and $(1+\mu p)\theta\log\theta^{-1}<1$

$$\left\|\boldsymbol{\iota}^*\breve{\boldsymbol{C}}_{\boldsymbol{y}}\right\|_2\le\sqrt{3n},\qquad\left\|\boldsymbol{\iota}^*\breve{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{P}_J\right\|_2\le\sqrt{8\,|J|\,p\log n}, \tag{E.27}$$

then the difference of Hessian (E.10), in direction $\boldsymbol{v}\in\mathbb{S}^{p-1}$ can be bounded as

$$\begin{aligned}&\left|\boldsymbol{v}^*\left(\widetilde{\mathrm{Hess}}[\varphi_{\ell^1}](\boldsymbol{a})-\mathrm{Hess}[\varphi_\rho](\boldsymbol{a})\right)\boldsymbol{v}\right|\\ &\le\left|\boldsymbol{v}^*\boldsymbol{\iota}^*\breve{\boldsymbol{C}}_{\boldsymbol{y}}\left(\boldsymbol{P}_{I(\boldsymbol{a})}-\mathrm{diag}\left[\nabla\mathcal{S}_\lambda^\delta\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\right]\right]\right)\breve{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{v}\right|+\|\nabla\varphi_{\ell^1}(\boldsymbol{a})-\nabla\varphi_\rho(\boldsymbol{a})\|_2\end{aligned} \tag{E.28}$$

where $I(\boldsymbol{a})$ is defined in (D.1). Let $\boldsymbol{D}=\boldsymbol{P}_{I(\boldsymbol{a})}-\mathrm{diag}\left[\nabla\mathcal{S}_\lambda^\delta\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\right]\right]$ and notice that $\boldsymbol{D}$ is a diagonal matrix, which suggests (E.28) can be decomposed using

$$(\boldsymbol{P}_J+\boldsymbol{P}_{J^c})\boldsymbol{D}(\boldsymbol{P}_J+\boldsymbol{P}_{J^c})=\boldsymbol{P}_J\boldsymbol{D}\boldsymbol{P}_J+\boldsymbol{P}_{J^c}\boldsymbol{D}\boldsymbol{P}_{J^c},$$

where, from with property of $\sqrt{\delta}$-smoothed $\ell^1$ function Lemma E.4:

$$\max_j|\boldsymbol{P}_J\boldsymbol{D}\boldsymbol{P}_J|_{jj}\le 1,\qquad\max_j|\boldsymbol{P}_{J^c}\boldsymbol{D}\boldsymbol{P}_{J^c}|_{jj}\le\sqrt{\lambda\delta}/B.$$

Finally, once again apply $\delta$ bound from (E.20) and bounds for $B,|J|,\boldsymbol{y}$ from (E.26)-(E.27), we gain

$$\begin{aligned}\text{(E.28)}&\le\left\|\boldsymbol{\iota}^*\breve{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{P}_J\right\|_2^2+\frac{\sqrt{\lambda\delta}}{B}\left\|\boldsymbol{\iota}^*\breve{\boldsymbol{C}}_{\boldsymbol{y}}\right\|_2^2+\|\nabla\varphi_{\ell^1}(\boldsymbol{a})-\nabla\varphi_\rho(\boldsymbol{a})\|_2\\ &\le 8\,|J|\,p\log n+\frac{3n\sqrt{\lambda\delta}}{B}+c'n\theta^2\\ &\le 8\cdot\frac{24c'n\theta^2}{p\log n}\cdot p\log n+\frac{3n\left(c'^4\lambda^2\theta^8/p^2\log^2 n\right)^{1/2}}{c'\lambda\theta^2/p\log p}+c'n\theta^2\\ &\le 200c'n\theta^2,\end{aligned}$$

where all above result holds with probability at least $1-10/n$ from Lemma E.5 and Lemma A.5. ■

# F Analysis of geometry

In this section we prove major geometrical result in Theorem 4.1. This lemma consists of three parts of geometry of $\varphi_\rho$; including the negative curvature region Corollary F.2, large gradient region Corollary F.4, strong convexity region near shift Corollary F.6, and retraction to subspace Corollary F.8, which are respectively base on geometric properties of $\varphi_{\ell^1}$ in Lemma F.1, Lemma F.3, Lemma F.5 and Lemma F.7. We will handle each individual region in the following subsections. To shed light on the technical detail of the proof, we will begin with two figures for illustration of a toy example, which demonstrate the geometry near a two dimension solution subspace $\mathcal{S}_{\{i,j\}}$, as follows:

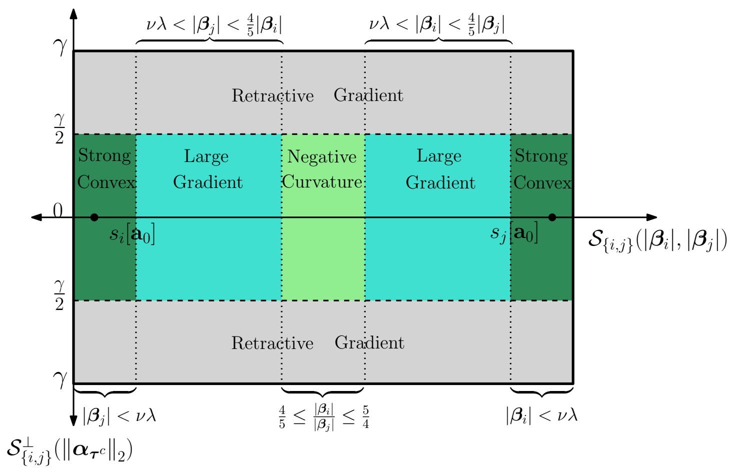


**Figure 15: The top view of geometry over subspace $\mathcal{S}_{\{i,j\}}$.** We display the geometric properties in the neighborhood of subspace $\mathcal{S}_{\{i,j\}}$ (horizontal axis) which contains the solutions $s_i[\boldsymbol{a}_0]$ and $s_j[\boldsymbol{a}_0]$. When $\boldsymbol{a}$ lies near middle of two shifts (light green region) such that $|\boldsymbol{\beta}_i| \approx |\boldsymbol{\beta}_j|$, then there exists a negative curvature direction in subspace $\mathcal{S}_{\{i,j\}}$. When $\boldsymbol{a}$ leans closer to one of the shifts $s_i[\boldsymbol{a}_0]$ (blue green region), its negative gradient direction points at that nearest shift. When $\boldsymbol{a}$ is in the neighborhood of the shift $s_i[\boldsymbol{a}_0]$ (dark green region) such that $|\boldsymbol{\beta}_i| \ll \lambda$, it will be strongly convex at $\boldsymbol{a}$, and the unique minimizer within the convex region will be close to $s_i[\boldsymbol{a}_0]$. Finally, the negative gradient will be pointing back toward the subspace $\mathcal{S}_{\{i,j\}}$ if near boundary (grey region).

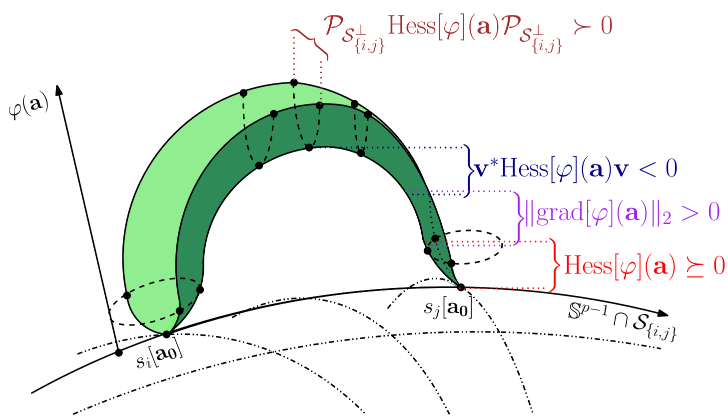


**Figure 16: The side view of geometry of subspace $\mathcal{S}_{\{i,j\}}$ on sphere.** We illustrate the geometry of $\mathcal{S}_{\{i,j\}}$ over the sphere, in which the properties of the three regions are denoted. In negative curvature region, there exists a direction $\boldsymbol{v}$ such that $\boldsymbol{v}^*\mathrm{Hess}[\varphi](\boldsymbol{a})\boldsymbol{v}$ is negative. In large gradient region, the norm of Riemannian gradient $\|\mathrm{grad}[\varphi](\boldsymbol{a})\|_2$ will be strictly greater then 0 and pointing at the nearest shift. Finally there is a convex region near all shifts such that $\mathrm{Hess}[\varphi](\boldsymbol{a})$ is positive semidefinite.

## F.1 Negative curvature

For any $\boldsymbol{a} \in \mathbb{S}^{p-1}$ near the subspace $\mathcal{S}_{\boldsymbol{\tau}}$ such that the entries of leading correlation vector $\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}$ have balanced magnitude, the Hessian of $\varphi_\rho(\boldsymbol{a})$ exhibits negative curvature in the span of $s_{(0)}[\boldsymbol{a}_0], s_{(1)}[\boldsymbol{a}_0]$. We will first demonstrate the pseudo negative curvature of $\varphi_{\ell^1}$ in Lemma F.1, then show $\varphi_\rho$ approximates $\varphi_{\ell^1}$ in terms of Hessian in Corollary F.2 when $\rho$ is properly defined as in Appendix E.

**Lemma F.1** (Negative curvature for $\varphi_{\ell^1}$). *Suppose that $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Set $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_{\ell^1}$ with $c_\lambda \in \left[\frac{1}{5}, \frac{1}{4}\right]$. There exist numerical constants $C, c, c', \bar{c} > 0$ such that if $n > Cp^5\theta^{-2}\log p$, and $c_\mu \le \bar{c}$, then with probability at least $1 - c'/n$ the following holds at every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ satisfying $\left|\boldsymbol{\beta}_{(1)}\right| \ge \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|$: for $\boldsymbol{v} \in \mathcal{S}_{\{(0),(1)\}} \cap \mathbb{S}^{p-1} \cap \boldsymbol{a}^\perp$,*

$$\boldsymbol{v}^*\widetilde{\mathrm{Hess}}[\varphi_{\ell^1}](\boldsymbol{a})\boldsymbol{v} \le -cn\theta\lambda. \tag{F.1}$$

**Proof.** First of all the regional condition $\left|\frac{\boldsymbol{\beta}_{(0)}}{\boldsymbol{\beta}_{(1)}}\right| \le \frac{5}{4}$ provides a two side bound for the two leading $\beta$'s

$$0.79 \ge \frac{\left|\boldsymbol{\beta}_{(0)}\right|}{\sqrt{\boldsymbol{\beta}_{(0)}^2 + \boldsymbol{\beta}_{(1)}^2}}\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_2 \ge \left|\boldsymbol{\beta}_{(0)}\right| \ge \left|\boldsymbol{\beta}_{(1)}\right| \ge \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right| \ge \frac{4}{5}\cdot\frac{\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_2}{\sqrt{|\boldsymbol{\tau}|}} \ge \frac{0.79}{\sqrt{|\boldsymbol{\tau}|}} \tag{F.2}$$

Set $J = \{(0),(1)\}$, choose $\boldsymbol{v} = \boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\iota}_J\boldsymbol{\gamma}$ with $\|\boldsymbol{v}\|_2 = 1$ then $\left|\|\boldsymbol{\gamma}\|_2^2 - 1\right| \le \mu$. There exists such $\boldsymbol{v}$ satisfies condition above with $\boldsymbol{a} \perp \boldsymbol{v}$ by choosing $\boldsymbol{\gamma}$ as

$$\boldsymbol{a}^*\boldsymbol{v} = \boldsymbol{a}^*\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\iota}_J\boldsymbol{\gamma} = \boldsymbol{\gamma}_{(0)}\boldsymbol{\beta}_{(0)} + \boldsymbol{\gamma}_{(1)}\boldsymbol{\beta}_{(1)} = 0,$$

hence $\left|\frac{\boldsymbol{\gamma}_{(1)}}{\boldsymbol{\gamma}_{(0)}}\right| = \left|\frac{\boldsymbol{\beta}_{(0)}}{\boldsymbol{\beta}_{(1)}}\right| \le \frac{5}{4}$. This implies $\boldsymbol{\gamma}_{(0)}^2 \ge \frac{16}{25}\boldsymbol{\gamma}_{(1)}^2 \ge \frac{16}{25}(1 - \mu - \boldsymbol{\gamma}_{(0)}^2)$ where $\mu \le \frac{c_\mu}{4} \le \frac{1}{100}$, it gives the lower bound of $\boldsymbol{\gamma}_{(0)}$ as

$$\boldsymbol{\gamma}_{(0)}^2 \ge \frac{(1-\mu)\cdot 16}{25+16} \ge 0.385 \tag{F.3}$$

1. (<u>Expand the Hessian</u>) The (pseudo) curvature along direction $\boldsymbol{v}$ is written as

$$\boldsymbol{v}^*\widetilde{\mathrm{Hess}}[\varphi_{\ell^1}](\boldsymbol{a})\boldsymbol{v} = \boldsymbol{v}^*\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a})\boldsymbol{v} - \langle\nabla\varphi_{\ell^1}(\boldsymbol{a}), \boldsymbol{a}\rangle = -\boldsymbol{\gamma}^*\boldsymbol{\iota}_J^*\boldsymbol{M}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{M}\boldsymbol{\iota}_J\boldsymbol{\gamma} + \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}] \tag{F.4}$$

expand the first term of (F.4) we obtain

$$\begin{aligned}
&-\boldsymbol{\gamma}^*\boldsymbol{\iota}_J^*\boldsymbol{M}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{M}\boldsymbol{\iota}_J\boldsymbol{\gamma} \\
&\quad= -\boldsymbol{\gamma}^*\boldsymbol{\iota}_J^*\boldsymbol{M}\left(\boldsymbol{P}_{(0)} + \boldsymbol{P}_{(1)} + \boldsymbol{P}_{J^c}\right)\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\check{\boldsymbol{C}}_{\boldsymbol{x}}\left(\boldsymbol{P}_{(0)} + \boldsymbol{P}_{(1)} + \boldsymbol{P}_{J^c}\right)\boldsymbol{M}\boldsymbol{\iota}_J\boldsymbol{\gamma} \\
&\quad\le -\sum_{i\in J}\left\|\boldsymbol{P}_{I(\boldsymbol{a})}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_i\right\|_2^2(\boldsymbol{e}_i^*\boldsymbol{M}\boldsymbol{\iota}_J\boldsymbol{\gamma})^2 + 2\sum_{\substack{(i,j)\in\{J,J^c\}\\(i,j)=((0),(1))}}\left|\boldsymbol{e}_i^*\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_j\right|\left|(\boldsymbol{e}_i^*\boldsymbol{M}\boldsymbol{\iota}_J\boldsymbol{\gamma})\left(\boldsymbol{e}_j^*\boldsymbol{M}\boldsymbol{\iota}_J\boldsymbol{\gamma}\right)\right| \\
&\quad\le -\sum_{i\in J}\left\|\boldsymbol{P}_{I(\boldsymbol{a})}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_i\right\|_2^2(|\gamma_i| - \mu)^2 \\
&\qquad\quad + 2\max_{i\ne j\in[\pm p]}\left|\boldsymbol{e}_i^*\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_j\right|\left(\left\|\boldsymbol{\iota}_J^*\boldsymbol{M}\boldsymbol{\iota}_J\boldsymbol{\gamma}\right\|_1\left\|\boldsymbol{\iota}_{J^c}^*\boldsymbol{M}\boldsymbol{\iota}_J\boldsymbol{\gamma}\right\|_1 + \left(\left|\boldsymbol{\gamma}_{(0)}\right| + \mu\right)\left(\left|\boldsymbol{\gamma}_{(1)}\right| + \mu\right)\right)
\end{aligned} \tag{F.5}$$

Consider the following events

$$\begin{cases}\mathcal{E}_{\text{cross}} := \left\{\forall\, \boldsymbol{a} \in \mathbb{S}^{p-1},\ \max_{i\ne j\in[\pm p]}\left|\boldsymbol{e}_i^*\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_j\right| < 4n\theta^2\right\} \\ \mathcal{E}_{\text{ncurv}} := \left\{\forall\, \boldsymbol{a} \in \mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu)),\ \min_{i\in J}\left\|\boldsymbol{P}_{I(\boldsymbol{a})}s_{-i}[\boldsymbol{x}]\right\|_2^2 \ge n\theta\left(1 - \mathbb{E}_{\boldsymbol{s}_i}(\lambda, \boldsymbol{s}_i) + \mathbb{E}_{\boldsymbol{s}_i}(\lambda, \boldsymbol{s}_i)\right) - \frac{c_\mu n\theta}{p}\right\}\end{cases}, \tag{F.6}$$

and from Lemma B.4 we know

$$\left\|\boldsymbol{\iota}_J^* \boldsymbol{M} \boldsymbol{\iota}_J \boldsymbol{\gamma}\right\|_1 \le \|\boldsymbol{\gamma}\|_1 + 2\mu \le 1.5, \quad \left\|\boldsymbol{\iota}_{J^c}^* \boldsymbol{M} \boldsymbol{\iota}_J \boldsymbol{\gamma}\right\|_1 \le \mu p \|\boldsymbol{\gamma}\|_1 \le 1.5\mu p,$$

on the event $\mathcal{E}_{\text{cross}} \cap \mathcal{E}_{\text{ncurv}}$, we have

$$\begin{aligned} &-\boldsymbol{\gamma}^* \boldsymbol{\iota}_J^* \boldsymbol{M} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{M} \boldsymbol{\iota}_J \boldsymbol{\gamma} \\ &\quad \le \underbrace{-n\theta \cdot \sum_{i \in J} (|\gamma_i| - \mu)^2 \left(1 - \mathbb{E}_{\boldsymbol{s}_i} \operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) + \mathbb{E}_{\boldsymbol{s}_i} f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right)}_{g_1(\boldsymbol{\beta})} + (18\mu p + 8)\, n\theta^2 + \frac{2c_\mu n\theta}{\sqrt{|\boldsymbol{\tau}|}} \end{aligned} \tag{F.7}$$

Meanwhile, for the latter term of (F.4), consider the following event $\mathcal{E}_{\overline{\chi}}$ where we write $\sigma_i = \operatorname{sign}(\boldsymbol{\beta}_i)$ as:

$$\mathcal{E}_{\overline{\chi}} := \left\{ \sigma_i \boldsymbol{\chi}[\boldsymbol{\beta}]_i \le \begin{cases} n\theta \cdot |\boldsymbol{\beta}_i| \left(1 - \mathbb{E}_{\boldsymbol{s}_i} \operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right) + \frac{c_\mu n\theta}{p}, & \forall\, i \in \boldsymbol{\tau} \\ n\theta \cdot |\boldsymbol{\beta}_i|\, 4\theta\, |\boldsymbol{\tau}| + \frac{c_\mu n\theta}{p}, & \forall\, i \in \boldsymbol{\tau}^c \end{cases} \right\}, \tag{F.8}$$

and use both $\|\boldsymbol{\beta}\|_1 \le \frac{c_\mu p}{\sqrt{|\boldsymbol{\tau}|}}$, $\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2^2 \le \frac{c_\mu}{\theta |\boldsymbol{\tau}|^2}$. On this event we have

$$\begin{aligned} \boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}] &\le n\theta \cdot \sum_{i \in \boldsymbol{\tau}} \boldsymbol{\beta}_i^2 \left(1 - \mathbb{E}_{\boldsymbol{s}_i} \operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right) + 4n\theta^2\, |\boldsymbol{\tau}|\, \|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2^2 + \frac{c_\mu n\theta}{p} \|\boldsymbol{\beta}\|_1 \\ &\le \underbrace{n\theta \cdot \sum_{i \in \boldsymbol{\tau}} \boldsymbol{\beta}_i^2 \left(1 - \mathbb{E}_{\boldsymbol{s}_i} \operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right)}_{g_2(\boldsymbol{\beta})} + \frac{5c_\mu n\theta}{\sqrt{|\boldsymbol{\tau}|}}. \end{aligned} \tag{F.9}$$

2. (Lower bound $\mathbb{E} f_{\boldsymbol{\beta}_i}$) Combine the first term from each of the (F.7) and (F.9). Use $\mu \le c_\mu \le \frac{1}{300}$ and (F.3) to obtain $\left(\left|\gamma_{(0)}\right| - \mu\right)^2 > 0.38$, we have

$$\begin{aligned} \frac{1}{n\theta} \left(g_1(\boldsymbol{\beta}) + g_2(\boldsymbol{\beta})\right) \le &- \sum_{i \in J} \left[ (|\gamma_i| - \mu)^2 - \boldsymbol{\beta}_i^2 \right] \left(1 - \mathbb{E}_{\boldsymbol{s}_i} \operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right) \\ &+ \sum_{i \in \boldsymbol{\tau} \setminus J} \boldsymbol{\beta}_i^2 \left(1 - \mathbb{E}_{\boldsymbol{s}_i} \operatorname{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right) - 0.38 \sum_{i \in J} \mathbb{E}_{\boldsymbol{s}_i} f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i), \end{aligned} \tag{F.10}$$

now use Taylor expansion [11] for $f_{\boldsymbol{\beta}_i}$ and apply upper bound $\mathbb{E} \boldsymbol{s}_i^2 \le \theta \|\boldsymbol{\beta}\|_2^2 \le \theta \left(1 + \frac{c_\mu}{\sqrt{|\boldsymbol{\tau}|}} + \frac{c_\mu}{\theta |\boldsymbol{\tau}|^2}\right) \le \frac{3c_\mu}{|\boldsymbol{\tau}|}$,

$$\mathbb{E}_{\boldsymbol{s}_i} f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) \ge \mathbb{E}_{\boldsymbol{s}_i} \frac{1}{\sqrt{2\pi}} \cdot \left( \frac{2\lambda}{|\boldsymbol{\beta}_i|} - \frac{\lambda^3}{|\boldsymbol{\beta}_i|^3} \left(1 + \frac{3\boldsymbol{s}_i^2}{\lambda^2}\right) \right) \ge \frac{1}{\sqrt{2\pi}} \cdot \underbrace{\left( \frac{2\lambda}{|\boldsymbol{\beta}_i|} - \frac{1}{|\boldsymbol{\beta}_i|^3} \left(\lambda^3 + \frac{9c_\mu \lambda}{|\boldsymbol{\tau}|}\right) \right)}_{f(\beta)},$$

where $f(\beta)$ is concave at stationary point since

$$\begin{cases} f'(\beta_*) = 0 \implies 2\lambda \beta_*^2 = 3\lambda \left(\lambda^2 + \frac{9c_\mu}{|\boldsymbol{\tau}|}\right) \\ f''(\beta_*) = \frac{1}{|\beta_*|^3} \left(4\lambda - \frac{12\lambda}{\beta_*^2} \left(\lambda^2 + \frac{9c_\mu}{|\boldsymbol{\tau}|}\right)\right) = \frac{1}{|\beta_*|^3} \left(4\lambda - \frac{12}{3/2}\lambda\right) < 0 \end{cases},$$

then combine with regional condition (F.2), and also apply assumption $c_\lambda \le \frac{1}{3}$ and $c_\mu \le \frac{1}{300}$, we gain

$$0.38 \sum_{i \in J} \mathbb{E}_{\boldsymbol{s}_i} f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) \ge 0.3 \min_{\beta = \frac{0.79}{\sqrt{|\boldsymbol{\tau}|}}, 0.79} f(\beta)$$

[11] Apply $\exp\left[-x^2/2\right] > 1 - x^2/2$

$$
\begin{aligned}
&\geq 0.3 \min\left\{\frac{2c_\lambda}{0.79} - \frac{c_\lambda^3 + 9c_\mu c_\lambda}{0.79^3},\ \lambda\left(\frac{2}{0.79} - \frac{c_\lambda^2 + 9c_\mu}{0.79^3}\right)\right\} \\
&\geq 0.3 \min\{2c_\lambda, 2\lambda\} \geq 0.6\lambda. \qquad \text{(F.11)}
\end{aligned}
$$

3. (Upper bound $\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i$) When $\boldsymbol{\beta}_{(0)}^2 = \left(\left|\gamma_{(0)}\right| - \mu\right)^2 - \eta$ for some $\eta > 0$. With monotonicity Lemma C.3, which implies:

$$
\left(1 - \mathbb{E}_{\boldsymbol{s}_{(0)}} \mathrm{erf}_{\boldsymbol{\beta}_{(0)}}(\lambda, \boldsymbol{s}_{(0)})\right) \geq \left(1 - \mathbb{E}_{\boldsymbol{s}_{(1)}} \mathrm{erf}_{\boldsymbol{\beta}_{(1)}}(\lambda, \boldsymbol{s}_{(1)})\right) \geq \left(1 - \mathbb{E}_{\boldsymbol{s}_i} \mathrm{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right), \qquad \text{(F.12)}
$$

then combine (F.11)-(F.12) and use $\mu \leq \frac{c_\mu}{4\sqrt{|\boldsymbol{\tau}|}}$ from Lemma B.5

$$
\begin{aligned}
\text{(F.10)} \leq &- \underbrace{\left(\left(\left|\gamma_{(0)}\right|^2 - \mu\right)^2 - \boldsymbol{\beta}_{(0)}^2 - \eta\right)}_{=0} \left(1 - \mathbb{E}_{\boldsymbol{s}_{(0)}} \mathrm{erf}_{\boldsymbol{\beta}_{(0)}}(\lambda, \boldsymbol{s}_{(0)})\right) \\
&+ \left(\sum_{i \in \boldsymbol{\tau}\setminus(0)} \boldsymbol{\beta}_i^2 - \left(\left|\gamma_{(1)}\right| - \mu\right)^2 - \eta\right) \underbrace{\left(1 - \mathbb{E}_{\boldsymbol{s}_{(1)}} \mathrm{erf}_{\boldsymbol{\beta}_{(1)}}(\lambda, \boldsymbol{s}_{(1)})\right)}_{<1} - 0.38 \sum_{i \in J} \mathbb{E}_{\boldsymbol{s}_i} f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) \\
\leq &\left(\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 - \|\boldsymbol{\gamma}\|_2^2 + 2\mu \|\boldsymbol{\gamma}\|_1\right) - 0.6\lambda \\
\leq &\frac{2c_\mu}{\sqrt{|\boldsymbol{\tau}|}} - 0.6\lambda. \qquad \text{(F.13)}
\end{aligned}
$$

On the other hand, when $\boldsymbol{\beta}_{(0)}^2 \geq \left(\left|\gamma_{(0)}\right| - \mu\right)^2 > 0.38$, combining (F.11)-(F.12) gives:

$$
\begin{aligned}
\text{(F.10)} \leq &\left(\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 - \|\boldsymbol{\gamma}\|_2^2 + 2\mu \|\boldsymbol{\gamma}\|_1\right) + \left(\left(\left|\gamma_{(0)}\right| - \mu\right)^2 - \boldsymbol{\beta}_{(0)}^2\right) \mathbb{E}_{\boldsymbol{s}_{(0)}} \mathrm{erf}_{\boldsymbol{\beta}_{(0)}}(\lambda, \boldsymbol{s}_{(0)}) \\
&+ \left(\left(\left|\gamma_{(1)}\right| - \mu\right)^2 - \sum_{i \in \boldsymbol{\tau}\setminus(0)} \beta_i^2\right) \mathbb{E}_{\boldsymbol{s}_{(1)}} \mathrm{erf}_{\boldsymbol{\beta}_{(1)}}(\lambda, \boldsymbol{s}_{(1)}) - 0.38 \sum_{i \in J} \mathbb{E}_{\boldsymbol{s}_i} f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) \\
\leq &\left(\frac{c_\mu}{\sqrt{|\boldsymbol{\tau}|}} + 4\mu\right) + \left(\gamma_{(1)}^2 - \|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 + \boldsymbol{\beta}_{(0)}^2\right) \mathbb{E}_{\boldsymbol{s}_{(1)}} \mathrm{erf}_{\boldsymbol{\beta}_{(1)}}(\lambda, \boldsymbol{s}_{(1)}) - 0.6\lambda, \qquad \text{(F.14)}
\end{aligned}
$$

where Lemma C.2 provides the upper bound for $\mathbb{E}_{\boldsymbol{s}_{(1)}} \mathrm{erf}_{\boldsymbol{\beta}_{(1)}}(\lambda, \boldsymbol{s}_{(1)})$ as

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{s}_{(1)}} \mathrm{erf}_{\boldsymbol{\beta}_{(1)}}(\lambda, \boldsymbol{s}_{(1)}) &= 1 - \frac{1}{n\theta \boldsymbol{\beta}_{(1)}} \mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_{(1)} \leq 1 - \frac{\sigma_{(1)}}{n\theta \left|\boldsymbol{\beta}_{(1)}\right|} \underline{\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]}_{(1)} = 1 - \frac{1}{\left|\boldsymbol{\beta}_{(1)}\right|} \left(\left|\boldsymbol{\beta}_{(1)}\right| - \sqrt{\frac{2}{\pi}} \lambda\right) \\
&\leq \sqrt{\frac{2}{\pi}} \cdot \frac{\lambda}{\left|\boldsymbol{\beta}_{(1)}\right|}, \qquad \text{(F.15)}
\end{aligned}
$$

then calculate the constant for the second term in (F.14) by writing $\kappa = \left|\frac{\gamma_{(1)}}{\gamma_{(0)}}\right| = \left|\frac{\boldsymbol{\beta}_{(0)}}{\boldsymbol{\beta}_{(1)}}\right| \leq \frac{5}{4}$, which provides $\gamma_{(1)}^2 \leq \frac{(1+\mu)\kappa^2}{\kappa^2+1}$ and $\boldsymbol{\beta}_{(0)}^2 \leq \frac{\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 \kappa^2}{\kappa^2+1}$ where $\mu < \frac{c_\mu}{4}$, and by applying $\left|\boldsymbol{\beta}_{(1)}\right| > \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right| \geq 0.3$, we have

$$
\frac{(\gamma_{(1)}^2 - 1) + c_\mu + \boldsymbol{\beta}_{(0)}^2}{\left|\boldsymbol{\beta}_{(1)}\right|} \leq -\frac{\kappa}{(\kappa^2+1)\left|\boldsymbol{\beta}_{(0)}\right|} + \kappa \left|\boldsymbol{\beta}_{(0)}\right| + \frac{\mu + c_\mu}{0.3} \leq \frac{\kappa^2 - 1}{\sqrt{\kappa^2+1}} + \kappa \left(\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 - 1\right) + 4.2c_\mu \leq 0.36 + 6c_\mu, \qquad \text{(F.16)}
$$

and finally combine (F.15)-(F.16), follow from (F.14) and use $c_\lambda \leq \frac{1}{3}$:

$$
\text{(F.10)} \leq \frac{2c_\mu}{\sqrt{|\boldsymbol{\tau}|}} + \sqrt{\frac{2}{\pi}} \left(\gamma_{(1)}^2 - 1 + c_\mu + \boldsymbol{\beta}_{(0)}^2\right) \frac{\lambda}{\left|\boldsymbol{\beta}_{(1)}\right|} - 0.6\lambda
$$

$$
\begin{aligned}
&\le \frac{2c_\mu}{\sqrt{|\boldsymbol{\tau}|}} + \sqrt{\frac{2}{\pi}}\left(0.36\lambda + \frac{6c_\mu c_\lambda}{0.3}\right) - 0.6\lambda \\
&\le \frac{4c_\mu}{\sqrt{|\boldsymbol{\tau}|}} - 0.3\lambda
\end{aligned} \tag{F.17}
$$

3. (Collect all results) Combine the components of pseudo Hessian (F.7), (F.9) with bounds for $g_1 + g_2$ from (F.13) and (F.17), and use Lemma B.5 which provides both $\mu p\theta\,|\boldsymbol{\tau}| < \frac{c_\mu}{4}$ and $\theta\,|\boldsymbol{\tau}| < \frac{c_\mu}{4}$ where $c_\mu < \frac{1}{300}$ and $c_\lambda \ge \frac{1}{5}$, we can obtain:

$$
\begin{aligned}
\boldsymbol{v}^*\widetilde{\mathrm{Hess}}_{\varphi_{\ell^1}}[\boldsymbol{a}]\boldsymbol{v} &\le g_1(\boldsymbol{\beta}) + g_2(\boldsymbol{\beta}) + \frac{7c_\mu n\theta}{\sqrt{|\boldsymbol{\tau}|}} + (18\mu p + 8)\,n\theta^2 \\
&\le n\theta\cdot\left(\frac{4c_\mu}{\sqrt{|\boldsymbol{\tau}|}} - 0.3\lambda\right) + n\theta\cdot\frac{7c_\mu}{\sqrt{|\boldsymbol{\tau}|}} + n\theta\cdot\frac{6.5c_\mu}{|\boldsymbol{\tau}|} \\
&\le \frac{n\theta}{\sqrt{|\boldsymbol{\tau}|}}(0.059 - 0.06) \le -0.001 n\theta\lambda
\end{aligned} \tag{F.18}
$$

Finally, the curvature is negative along $\boldsymbol{v}$ direction with probability at least

$$
1 - \underbrace{\mathbb{P}\left[\mathcal{E}^c_{\mathrm{cross}}\right]}_{\text{Lemma A.4}} - \underbrace{\mathbb{P}\left[\mathcal{E}^c_{\mathrm{ncurv}}\right]}_{\text{Corollary D.3}} - \underbrace{\mathbb{P}\left[\mathcal{E}^c_{\overline{\mathsf{X}}}\right]}_{\text{Corollary C.4}}. \tag{F.19}
$$

■

Similarly for objective $\varphi_\rho$, we have that

**Corollary F.2** (Negative curvature for $\varphi_\rho$)**.** *Suppose that $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_\rho$ where $c_\lambda \in \left[\frac{1}{5}, \frac{1}{4}\right]$, then there exists some numerical constants $C, c, c', c'', \bar{c} > 0$ such that if $\rho$ is $\delta$-smoothed $\ell^1$ function where $\delta \le c''\lambda\theta^8/p^2\log^2 n$, $n > Cp^5\theta^{-2}\log p$ and $c_\mu \le \bar{c}$, then with probability at least $1 - c'/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ satisfying $\left|\boldsymbol{\beta}_{(1)}\right| \ge \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|$: for $\boldsymbol{v} \in \mathcal{S}_{\{(0),(1)\}} \cap \mathbb{S}^{p-1} \cap \boldsymbol{a}^\perp$,*

$$
\boldsymbol{v}^*\widetilde{\mathrm{Hess}}[\varphi_\rho](\boldsymbol{a})\boldsymbol{v} \le -cn\theta\lambda \tag{F.20}
$$

**Proof.** Choose $\boldsymbol{v} \in \mathbb{S}^{p-1}$ according to Lemma F.1 and (E.23) from Lemma E.6 with constant multiplier $\delta$ satisfies $c''^{1/4} < 10^{-3}c$, we gain

$$
\boldsymbol{v}^*\mathrm{Hess}[\varphi_\rho](\boldsymbol{a})\boldsymbol{v} \le -cn\theta\lambda + 200c'n\theta^2 \le -cn\theta\lambda/2 \tag{F.21}
$$

■

## F.2 Large gradient

For any $\boldsymbol{a} \in \mathbb{S}^{p-1}$ near subspace and the second largest correlation $\boldsymbol{\beta}_{(1)}$ much smaller then the first correlation $\boldsymbol{\beta}_{(0)}$ while not being near 0, the negative gradient of $\varphi_\rho(\boldsymbol{a})$ will point at the largest shift. We show this in Lemma F.3, and the $\varphi_\rho$ version in Corollary F.4 when $\rho$ is properly defined as in Appendix E.

**Lemma F.3** (Large gradient for $\varphi_{\ell^1}$)**.** *Suppose that $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_{\ell^1}$ with some $c_\lambda \in \left[\frac{1}{5}, \frac{1}{4}\right]$, then there exists some numerical constants $C, c', c, \bar{c} > 0$, such that if $n > Cp^5\theta^{-2}\log p$ and $c_\mu \le \bar{c}$, then with probability at least $1 - c'/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ satisfying $\frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right| > \left|\boldsymbol{\beta}_{(1)}\right| > \frac{1}{4\log\theta^{-1}}\lambda$,*

$$
\left\langle \boldsymbol{\sigma}_{(0)}\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], -\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a})\right\rangle \ge cn\theta\left(\log^{-2}\theta^{-1}\right)\lambda^2 \tag{F.22}
$$

*where $\boldsymbol{\sigma}_i = \mathrm{sign}(\boldsymbol{\beta}_i)$.*

**Proof.** 1. (Properties for $\boldsymbol{\alpha}, \boldsymbol{\beta}$) Define $\theta_{\log} = \frac{1}{\log \theta^{-1}}$, we first derive upper bound on the dominant entry $\left|\boldsymbol{\beta}_{(0)}\right|$ as follows. Write the geodesic distance between $\boldsymbol{a}$ and $\boldsymbol{\iota}^* s_i[\boldsymbol{a}_0]$ as a function of $\boldsymbol{\beta}_i$ as $d_{\mathbb{S}}(\boldsymbol{a}, \pm\boldsymbol{\iota}^* s_i[\boldsymbol{a}_0]) = \cos^{-1}(\boldsymbol{\beta}_i)$, then by triangle inequality we have:

$$
\begin{aligned}
&d_{\mathbb{S}}(\boldsymbol{a}, \pm\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0]) \geq d_{\mathbb{S}}(\pm\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], \boldsymbol{\iota}^* s_{(1)}[\boldsymbol{a}_0]) - d_{\mathbb{S}}(\boldsymbol{a}, \boldsymbol{\iota}^* s_{(1)}[\boldsymbol{a}_0]) \\
\Longrightarrow\ &\cos^{-1} \pm\boldsymbol{\beta}_{(0)} \geq \cos^{-1}\mu - \cos^{-1}\left|\boldsymbol{\beta}_{(1)}\right| \\
\Longrightarrow\ &\pm\boldsymbol{\beta}_{(0)} \leq \cos\left(\cos^{-1}\mu - \cos^{-1}\left|\boldsymbol{\beta}_{(1)}\right|\right) = \mu\left|\boldsymbol{\beta}_{(1)}\right| + \sqrt{(1-\mu^2)\left(1-\boldsymbol{\beta}_{(1)}^2\right)} \leq 1 - \tfrac{1}{2}\left(\left|\boldsymbol{\beta}_{(1)}\right| - \mu\right)^2.
\end{aligned}
$$

Use the regional condition $\left|\boldsymbol{\beta}_{(1)}\right| \geq \frac{\theta_{\log}}{4}\lambda$ and since $\mu\left|\boldsymbol{\tau}\right|^{3/2} < \frac{c_\lambda}{100}\theta_{\log}$ from Definition B.1, implies

$$
\left|\boldsymbol{\beta}_{(0)}\right| \leq 1 - \frac{\boldsymbol{\beta}_{(1)}^2}{2}\left(1 - \frac{4\mu\sqrt{|\boldsymbol{\tau}|}}{\theta_{\log} c_\lambda}\right) \leq 1 - 0.49\boldsymbol{\beta}_{(1)}^2 =: \beta_{\mathrm{ub}}. \tag{F.23}
$$

Meanwhile a lower bound for $\boldsymbol{\beta}_{(0)}$ can be easily determined by the other side of regional condition:

$$
\left|\boldsymbol{\beta}_{(0)}\right| \geq \tfrac{5}{4}\left|\boldsymbol{\beta}_{(1)}\right| =: \beta_{\mathrm{lb}}. \tag{F.24}
$$

Also since $\boldsymbol{\beta} = \boldsymbol{M}\boldsymbol{\alpha}$, based on properties of $\boldsymbol{M}$ from Lemma B.4. When $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 \leq 1 + c_\mu$ and $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \leq \gamma \leq \frac{c_\mu \theta_{\log}^2}{4\mu\sqrt{p}|\boldsymbol{\tau}|}$, we gain:

$$
\begin{aligned}
&\boldsymbol{\beta}_{(0)} = \boldsymbol{\alpha}_{(0)} + \boldsymbol{e}_{(0)}^* \boldsymbol{M}\boldsymbol{\alpha}_{\backslash(0)} \\
\Longrightarrow\ &\left|\boldsymbol{\alpha}_{(0)} - \boldsymbol{\beta}_{(0)}\right| \leq \mu\sqrt{|\boldsymbol{\tau}|}\,\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 + \mu\sqrt{p}\,\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \leq \tfrac{c_\mu\theta_{\log}^2(1+c_\mu)}{4|\boldsymbol{\tau}|} + \mu\sqrt{p}\gamma \leq \tfrac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}.
\end{aligned} \tag{F.25}
$$

and therefore $\left|\boldsymbol{\alpha}_{(0)}\right| \leq \left|\boldsymbol{\beta}_{(0)}\right| + \frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|} \leq 1 - .49\left(\frac{\theta_{\log}}{4}\lambda\right)^2 + \frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|} < 1$.

2. (Upper bound of $\boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]$) Define a piecewise smooth convex upper bound $h$ for $\boldsymbol{\beta}_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i$ as:

$$
h(\boldsymbol{\beta}_i) := \begin{cases} \boldsymbol{\beta}_i^2 - \frac{\nu_1\lambda}{2}|\boldsymbol{\beta}_i| & |\boldsymbol{\beta}_i| \geq \nu_1\lambda \\ \frac{1}{2}\boldsymbol{\beta}_i^2 & |\boldsymbol{\beta}_i| \leq \nu_1\lambda \end{cases},
$$

then Lemma J.7 tells us since $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}\backslash(0)}\right\|_\infty \leq \boldsymbol{\beta}_{(1)}$:

$$
\begin{aligned}
\sum_{i\in\boldsymbol{\tau}\backslash(0)} h(\boldsymbol{\beta}_i) &\leq \left\|\boldsymbol{\beta}_{\boldsymbol{\tau}\backslash(0)}\right\|_2^2\left(1 - \frac{\nu_1\lambda\boldsymbol{\beta}_{(1)}}{2\boldsymbol{\beta}_{(1)}^2}\right) \leq \left(1 + \frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|} - \boldsymbol{\beta}_{(0)}^2\right)\left(1 - \frac{\nu_1\lambda}{2\boldsymbol{\beta}_{(1)}}\right) \\
&\leq \left(1 - \frac{\nu_1\lambda}{2\boldsymbol{\beta}_{(1)}}\right)\left(1 - \boldsymbol{\beta}_{(0)}^2\right) + \frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|},
\end{aligned}
$$

then condition on the following event using Corollary C.4,

$$
\mathcal{E}_{\overline{\chi}} := \left\{\boldsymbol{\beta}_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i \leq \begin{cases} n\theta\cdot h(\boldsymbol{\beta}_i) + \frac{c_\mu\theta}{p^{3/2}}|\boldsymbol{\beta}_i|, & \forall i \in \boldsymbol{\tau}\setminus(0) \\ n\theta\cdot 4\boldsymbol{\beta}_i^2\theta|\boldsymbol{\tau}| + \frac{c_\mu\theta}{p^{3/2}}|\boldsymbol{\beta}_i|, & \forall i \in \boldsymbol{\tau}^c \end{cases}\right\},
$$

which provides the upper bound of $\boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]$ by applying $5p > \log^{8/3}(p\log^2 p) > (\theta_{\log}^2)^{4/3}$ from lower bound of $\theta$ from Definition B.1, $\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2 \leq \frac{c_\mu\theta_{\log}}{\sqrt{\theta}|\boldsymbol{\tau}|}$ from Lemma B.5 , $|\boldsymbol{\tau}| \leq \sqrt{p}$ from lemma assumption and let $c_\mu < \frac{1}{100}$:

$$
\boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}] \leq \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)}\boldsymbol{\beta}_{(0)} + \sum_{i\in\boldsymbol{\tau}\backslash(0)}\boldsymbol{\beta}_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i + \langle\boldsymbol{\beta}_{\boldsymbol{\tau}^c}, \boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\rangle
$$

$$
\begin{aligned}
&\leq \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)}\boldsymbol{\beta}_{(0)} + n\left(\theta \sum_{i\in\boldsymbol{\tau}\setminus(0)} h(\boldsymbol{\beta}_i) + 4\theta^2\left|\boldsymbol{\tau}\right|\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\right\|_2^2 + \frac{c_\mu\theta}{p^{3/2}}\left(\sqrt{|\boldsymbol{\tau}|}\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_2 + \sqrt{p}\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\right\|_2\right)\right)\\
&\leq \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)}\boldsymbol{\beta}_{(0)} + n\left(\theta\cdot\eta(1-\boldsymbol{\beta}_{(0)}^2) + \theta\cdot\frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|} + \frac{4\theta^2\left|\boldsymbol{\tau}\right|c_\mu^2\theta_{\log}^2}{\theta\left|\boldsymbol{\tau}\right|^2} + c_\mu\theta\left(\frac{1+c_\mu}{p^{3/4}\left|\boldsymbol{\tau}\right|} + \frac{c_\mu\theta_{\log}}{p\sqrt{\theta}\left|\boldsymbol{\tau}\right|}\right)\right)\\
&\leq \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)}\boldsymbol{\beta}_{(0)} + n\theta\left(\eta(1-\boldsymbol{\beta}_{(0)}^2) + \frac{6c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\right),
\end{aligned}
\tag{F.26}
$$

where $\eta = 1 - \frac{\nu_1\lambda}{2\boldsymbol{\beta}_{(1)}}$.

3. (Align the gradient with $\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0]$) Base on the definition $\boldsymbol{\beta}$, since $\boldsymbol{\beta}_{(0)} = \left\langle \boldsymbol{a}, \boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0]\right\rangle$, we can expect that the negative gradient is likely aligned with direction toward one of the candidate solution $\pm\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0]$. Wlog assume that both $\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}$ are positive, then expand the gradient and use incoherent property for $\boldsymbol{a}_0$ Lemma B.4 we have:

$$
\begin{aligned}
\left\langle \boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], -\operatorname{grad}_{\varphi_{\ell_1}}[\boldsymbol{a}]\right\rangle &= \left\langle \boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], \boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\left(\boldsymbol{\chi}[\boldsymbol{\beta}] - \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{\alpha}\right)\right\rangle\\
&\geq \left(\boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)} - \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{\alpha}_{(0)}\right) - \mu\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\setminus(0)} - \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{\alpha}_{\setminus(0)}\right\|_1,
\end{aligned}
\tag{F.27}
$$

where $\setminus(0)$ is an abbreviation of the complement set $[\pm 2p_0]\setminus(0)$. The latter part of (F.27) has an upper bound using bounds of $\boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}] < \frac{3n\theta}{2}$, $\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2 < \frac{n\theta\gamma_2}{20}$ from (F.62), and $\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}\setminus(0)}\right\|_2 \leq n\theta\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}\setminus(0)}\right\|_2$ in event $\mathcal{E}_{\overline{\mathcal{X}}}$, we obtain:

$$
\begin{aligned}
&\mu\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\setminus(0)} - \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{\alpha}_{\setminus(0)}\right\|_1\\
&\leq \mu\left(\sqrt{|\boldsymbol{\tau}|}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}\setminus(0)}\right\|_2 + \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\sqrt{|\boldsymbol{\tau}|}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}\setminus(0)}\right\|_2 + \sqrt{p}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\right\|_2 + \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\sqrt{p}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\right\|_2\right)\\
&\leq n\theta\cdot\left[\mu\sqrt{|\boldsymbol{\tau}|}\left(\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2 - \left|\boldsymbol{\beta}_{(0)}\right|\right) + \mu\sqrt{|\boldsymbol{\tau}|}\left(\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\|_2 - \left|\boldsymbol{\alpha}_{(0)}\right|\right) + \frac{1}{20}\mu\sqrt{p}\gamma_2 + \frac{3}{2}\mu\sqrt{p}\gamma_2\right]\\
&\leq n\theta\cdot\frac{c_\mu\theta_{\log}^2}{4\left|\boldsymbol{\tau}\right|}\left[2\left(1+c_\mu\right) - \left|\boldsymbol{\beta}_{(0)}\right| - \left|\boldsymbol{\alpha}_{(0)}\right| + \left(\frac{1}{20}+\frac{3}{2}\right)c_\mu\right]\\
&\leq n\theta\cdot\frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\left(0.5 + c_\mu - 0.5\boldsymbol{\beta}_{(0)}\right).
\end{aligned}
\tag{F.28}
$$

On the other hand, the former term of (F.27) possesses a lower bound using (F.25)-(F.26), $\boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)} > n\theta\left(\boldsymbol{\beta}_{(0)} - \frac{\nu_1}{2}\lambda - \frac{c_\mu}{p}\right) \geq n\theta\left(\boldsymbol{\beta}_{(0)} - 0.51\nu_1\lambda\right)$ and $\boldsymbol{\alpha}_{(0)} \leq 1$:

$$
\begin{aligned}
&\boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)} - \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{\alpha}_{(0)}\\
&\geq \left(1 - \boldsymbol{\alpha}_{(0)}\boldsymbol{\beta}_{(0)}\right)\boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)} - n\theta\cdot\left[\eta\left(1-\boldsymbol{\beta}_{(0)}^2\right) + \frac{6c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\right]\boldsymbol{\alpha}_{(0)}\\
&\geq \underbrace{n\theta\left(1 - \left(\boldsymbol{\beta}_{(0)} + \frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\right)\boldsymbol{\beta}_{(0)}\right)\left(\boldsymbol{\beta}_{(0)} - 0.51\nu_1\lambda\right)}_{(a)} - n\theta\underbrace{\left[\eta\left(1-\boldsymbol{\beta}_{(0)}^2\right)\left(\boldsymbol{\beta}_{(0)} + \frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\right) + \frac{6c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\boldsymbol{\alpha}_{(0)}\right]}_{(b)}\\
&\geq n\theta\left[\underbrace{\left(1-\boldsymbol{\beta}_{(0)}^2\right)\left(\boldsymbol{\beta}_{(0)} - 0.51\nu_1\lambda\right) - \frac{c_\mu\theta_{\log}^2\boldsymbol{\beta}_{(0)}^2}{|\boldsymbol{\tau}|}}_{(a)}\underbrace{-\left(1-\boldsymbol{\beta}_{(0)}^2\right)\eta\boldsymbol{\beta}_{(0)} - \eta\frac{c_\mu\theta_{\log}\left(1-\boldsymbol{\beta}_{(0)}^2\right)}{|\boldsymbol{\tau}|} - \frac{6c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}}_{(b)}\right]
\end{aligned}
$$

$$\geq n\theta\left[\left(1-\boldsymbol{\beta}_{(0)}^2\right)\left(\left(1-\eta\right)\boldsymbol{\beta}_{(0)}-0.51\nu_1\lambda\right)-\frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\left((1-\eta)\boldsymbol{\beta}_{(0)}^2+7\right)\right], \tag{F.29}$$

combine (F.27) with (F.28)-(F.29) and $\eta > 0$, we have

$$\begin{aligned}(\text{F.27}) &\geq n\theta\left[\left(1-\boldsymbol{\beta}_{(0)}^2\right)\left(\left(1-\eta\right)\boldsymbol{\beta}_{(0)}-0.51\nu_1\lambda\right)-\frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\left((1-\eta)\boldsymbol{\beta}_{(0)}^2+7\right)\right]-n\theta\cdot\frac{c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\left(0.5+c_\mu-0.5\boldsymbol{\beta}_{(0)}\right)\\ &\geq n\theta\left[\underbrace{\left(1-\boldsymbol{\beta}_{(0)}^2\right)\left(\frac{\nu_1\lambda}{2\boldsymbol{\beta}_{(1)}}\boldsymbol{\beta}_{(0)}-0.51\nu_1\lambda\right)}_{f(\beta)}-\frac{8c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\right]. \end{aligned}\tag{F.30}$$

4. (Lower bound of $f(\beta)$) Given a fixed $\boldsymbol{\beta}_{(1)}$, the cubic function $f(\boldsymbol{\beta}_{(0)})$ has zeros set $\boldsymbol{\beta}_{(0)} \in \left\{\pm 1, 1.02\boldsymbol{\beta}_{(1)}\right\}$ and has negative leading coefficient. Combine with the condition of $\boldsymbol{\beta}_{(0)} \in \{\beta_{\mathrm{lb}}, \beta_{\mathrm{ub}}\}$ from (F.23)-(F.24), we can observe that

$$\boldsymbol{\beta}_{(0)} \in [\beta_{\mathrm{lb}}, \beta_{\mathrm{ub}}] = \left[\frac{5}{4}\boldsymbol{\beta}_{(1)}, 1-0.49\boldsymbol{\beta}_{(1)}^2\right] \subseteq \left[1.02\boldsymbol{\beta}_{(1)}, 1\right],$$

therefore the cubic term is always positive and minimizer is either one of the boundary point. When $\boldsymbol{\beta}_{(0)} = \beta_{\mathrm{lb}}$, use $\left(1+\frac{25}{16}\right)\boldsymbol{\beta}_{(1)}^2 < 1.01$, and use $\nu_1\lambda < \frac{\sqrt{\theta_{\log}}}{2\sqrt{|\boldsymbol{\tau}|}} \leq \frac{1}{2\sqrt{2}}$, since $|\boldsymbol{\tau}| \geq 2$, we have:

$$f(\beta_{\mathrm{lb}}) \geq \left(1-\beta_{\mathrm{lb}}^2\right)\left(\frac{\nu_1\lambda}{2\boldsymbol{\beta}_{(1)}}\beta_{\mathrm{lb}}-0.51\nu_1\lambda\right) \geq (1-0.616)\cdot\left(\frac{5}{8}-0.51\right)\nu_1\lambda \geq \frac{1}{16\sqrt{2}}\nu_1\lambda \geq \frac{\theta_{\log}^2}{32}\lambda^2. \tag{F.31}$$

On the other hand when $\boldsymbol{\beta}_{(0)} = \beta_{\mathrm{ub}}$:

$$f(\beta_{\mathrm{ub}}) \geq \left(1-\beta_{\mathrm{ub}}^2\right)\left(\frac{\nu_1\lambda}{2\boldsymbol{\beta}_{(1)}}\beta_{\mathrm{ub}}-0.51\nu_1\lambda\right) \geq 0.49\boldsymbol{\beta}_{(1)}^2\cdot\left(\frac{\nu_1\lambda}{2\boldsymbol{\beta}_{(1)}}\left(1-0.49\boldsymbol{\beta}_{(1)}^2\right)-0.51\nu_1\lambda\right),$$

which is a cubic function of $\boldsymbol{\beta}_{(1)}$ with negative leading coefficient, whose zeros set is $\{-0.73, 0, 2.81\}$. Thus it minimizes at the boundary points of $\boldsymbol{\beta}_{(1)} \in \left[\frac{\lambda}{4\log\theta^{-1}}, 1\right] \subset [0, 2.81]$, thus assign $\boldsymbol{\beta}_{(1)} = \frac{\lambda}{4\log\theta^{-1}}$, we have:

$$f(\beta_{\mathrm{ub}}) \geq 0.49\left(\frac{\lambda}{4\log\theta^{-1}}\right)^2\cdot\left(\frac{1}{2}\left(1-0.49\left(\frac{\lambda}{4\log\theta^{-1}}\right)^2\right)-0.51\nu_1\lambda\right) \geq \frac{1}{6}\left(\frac{\lambda}{4\log\theta^{-1}}\right)^2 \geq \frac{\theta_{\log}^2}{96}\lambda^2. \tag{F.32}$$

Finally combine (F.30) with the lower bound of cubic function (F.31)-(F.32) together with condition $c_\mu < \frac{c_\lambda^2}{800}$ and $\nu_1 = \frac{\sqrt{\theta_{\log}}}{2}$, obtain

$$\begin{aligned}\left\langle \boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], -\mathrm{grad}_{\varphi_{\ell_1}}[\boldsymbol{a}]\right\rangle &\geq n\theta\cdot\left(\min\left\{f(\beta_{\mathrm{ub}}), f(\beta_{\mathrm{lb}})\right\}-\frac{8c_\mu\theta_{\log}^2}{|\boldsymbol{\tau}|}\right)\\ &\geq n\theta\left(\frac{\theta_{\log}^2c_\lambda^2}{96\,|\boldsymbol{\tau}|}-\frac{8\theta_{\log}^2c_\lambda^2}{800\,|\boldsymbol{\tau}|}\right) \geq 6\times 10^{-3}n\theta\theta_{\log}^2c_\lambda^2. \end{aligned}\tag{F.33}$$

The proof for the case where $\boldsymbol{\beta}_{(0)}$ negative can be derived in the same manner. ■

As a consequence, we have that

**Corollary F.4** (Large gradient for $\varphi_\rho$)**.** *Suppose that $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_\rho$ with $c_\lambda \in \left[\frac{1}{5}, \frac{1}{4}\right]$, then there exists some numerical constants $C, c, c', c'', \bar{c} > 0$ such that if $\rho$ is $\delta$-smoothed $\ell^1$ function where $\delta \le c''\lambda\theta^8/p^2\log^2 n$ with $n > Cp^5\theta^{-2}\log p$ and $c_\mu \le \bar{c}$, then with probability at least $1 - c'/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ satisfying $\frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right| > \left|\boldsymbol{\beta}_{(1)}\right| > \frac{1}{4\log\theta^{-1}\lambda}$,*

$$\left\langle \boldsymbol{\sigma}_{(0)}\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], -\mathrm{grad}[\varphi_\rho](\boldsymbol{a})\right\rangle \ge cn\theta\left(\log^{-2}\theta^{-1}\right)\lambda^2 \tag{F.34}$$

*where $\boldsymbol{\sigma}_i = \mathrm{sign}(\boldsymbol{\beta}_i)$.*

**Proof.** Choose $\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0]$ as in Lemma F.3, and apply (E.22) from Lemma E.6 with the constant multiplier of $\delta$ satisfies $c''^4 < c/4$, then utilize $\theta\,|\boldsymbol{\tau}|\log^2\theta^{-1} < c_\mu$ from Definition B.1 we have

$$\left\langle \boldsymbol{\sigma}_{(0)}\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], -\mathrm{grad}[\varphi_\rho](\boldsymbol{a})\right\rangle \ge cn\theta(\log^{-2}\theta^{-1})\lambda - c''n\theta^2 \ge cn\theta(\log^{-2}\theta^{-1})\lambda/2 \tag{F.35}$$

■

## F.3 Convex near solutions

For any $\boldsymbol{a} \in \mathbb{S}^{p-1}$ near subspace and the second largest correlation $\boldsymbol{\beta}_{(1)}$ smaller then $\frac{1}{4\log\theta^{-1}}\lambda$, then $\varphi_\rho$ will be strongly convex at $\boldsymbol{a}$. We show this in Lemma F.5, and the $\varphi_\rho$ version in Corollary F.6 when $\rho$ is properly defined as in Appendix E.

**Lemma F.5** (Strong convexity of $\varphi_{\ell^1}$ near shift)**.** *Suppose that $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_{\ell^1}$ with $c_\lambda \in \left[\frac{1}{4}, \frac{1}{5}\right]$, then there exists some numerical constants $C, c, c'\bar{c} > 0$ such that if $n > Cp^5\theta^{-2}\log p$ and $c_\mu \le \bar{c}$, then with probability at least $1 - c'/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ satisfying $\left|\boldsymbol{\beta}_{(1)}\right| < \frac{1}{4\log\theta^{-1}}\lambda$: for all $\boldsymbol{v} \in \mathbb{S}^{p-1} \cap \boldsymbol{v}^\perp$,*

$$\boldsymbol{v}^*\widetilde{\mathrm{Hess}}[\varphi_{\ell^1}](\boldsymbol{a})\boldsymbol{v} > cn\theta; \tag{F.36}$$

*furthermore, there exists $\bar{\boldsymbol{a}}$ as an local minimizer such that*

$$\min_\ell \left\|\bar{\boldsymbol{a}} - s_\ell[\boldsymbol{a}_0]\right\|_2 \le \tfrac{1}{2}\max\left\{\mu, p^{-1}\right\}. \tag{F.37}$$

**Proof.** 1. (Expectation of $\boldsymbol{\chi}$ near shifts) We will write $\boldsymbol{x}$ as $\boldsymbol{x}_0$ through out this proof. When $\boldsymbol{a}$ is near one of the shift, the $\boldsymbol{\chi}$ operator shrinks all other smaller entries of correlation vector $\boldsymbol{\beta}_{\setminus(0)}$ in an even larger shrinking ratio. Firstly we can show $\left|\left\langle\boldsymbol{\beta}_{\setminus(0)}, \boldsymbol{x}_{\setminus(0)}\right\rangle\right|$ is no larger then $\lambda/2$ with probability at least $1 - 4\theta$, since

$$\mathbb{P}\left[\left|\left\langle\boldsymbol{\beta}_{\setminus(0)}, \boldsymbol{x}_{\setminus(0)}\right\rangle\right| > \frac{\lambda}{2}\right] \le \mathbb{P}\left[\left|\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau}\setminus(0)}, \boldsymbol{x}_{\boldsymbol{\tau}\setminus(0)}\right\rangle\right| > \frac{2\lambda}{5}\right] + \mathbb{P}\left[\left|\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau}^c}, \boldsymbol{x}_{\boldsymbol{\tau}^c}\right\rangle\right| > \frac{\lambda}{10}\right] \le 4\theta \tag{F.38}$$

via Corollary B.6 and Corollary B.7. Now recall from Lemma C.2 and the derivation of (C.10)-(C.11), we know for every $i \ne (0)$,

$$\begin{aligned}
\boldsymbol{\sigma}_i\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i &= n\theta\left|\boldsymbol{\beta}_i\right|\mathbb{E}_{\boldsymbol{s}_i}\left[1 - \mathrm{erf}_{\boldsymbol{\beta}_i}\left(\lambda, \boldsymbol{s}_i\right)\right] \\
&\le n\theta\left|\boldsymbol{\beta}_i\right|\mathbb{E}_{g,\boldsymbol{x}_{\setminus i}}\left[g^2\mathbf{1}_{\left\{\left|\boldsymbol{\beta}_i g + \boldsymbol{\beta}_{(0)}\boldsymbol{x}_{(0)} + \boldsymbol{\beta}^*_{\setminus\{(0),i\}}\boldsymbol{x}_{\setminus\{(0),i\}}\right| > \lambda\right\}}\right] \\
&\le n\theta\left|\boldsymbol{\beta}_i\right|\left(\mathbb{E}g^2\mathbf{1}_{\{|\boldsymbol{\beta}_i g|>\lambda/2\}} + \mathbb{P}\left[\boldsymbol{x}_{(0)} \ne 0\right] + \mathbb{P}\left[\left|\left\langle\boldsymbol{\beta}_{\setminus\{(0),i\}}, \boldsymbol{x}_{\setminus\{(0),i\}}\right\rangle\right| > \lambda/2\right]\right) \\
&\le n\theta\left|\boldsymbol{\beta}_i\right|\left(\left(\mathbb{E}g^2\right)^{1/2}\mathbb{P}\left[\left|\boldsymbol{\beta}_{(1)}g\right| > \lambda/2\right]^{1/2} + \theta + 4\theta\right) \\
&\le n\theta\left|\boldsymbol{\beta}_i\right|\left(\exp\left(-\log^2\theta^{-1}\right) + 5\theta\right) \\
&\le 6n\theta^2\left|\boldsymbol{\beta}_i\right|
\end{aligned} \tag{F.39}$$

where the third inequality is derived using union bound; the the fourth inequality is the result of (F.38), and the fifth inequality is derived from Gaussian tail bound lemma J.1.

2. (Local strong convexity) Let $\boldsymbol{\gamma} = \boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{\iota}\boldsymbol{v}$, for any $\|\boldsymbol{v}\|_2 = 1$ we have $\|\boldsymbol{\gamma}\|_2^2 \le 1 + \mu p$. Furthermore:

$$\begin{aligned} \left|\gamma_{(0)}\right| &= \left|\left\langle \boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], \boldsymbol{v}\right\rangle\right| = \left|\left\langle \boldsymbol{P}_{\boldsymbol{a}^\perp}\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], \boldsymbol{v}\right\rangle\right| = \left|\left\langle \boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0] - \boldsymbol{\beta}_{(0)}\boldsymbol{a}, \boldsymbol{v}\right\rangle\right| \\ &\le \left\|\boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0] - \boldsymbol{\beta}_{(0)}\boldsymbol{a}\right\|_2 \le \sqrt{1-\boldsymbol{\beta}_{(0)}^2}. \end{aligned} \tag{F.40}$$

Consider any such $\boldsymbol{v}$, the pseudo Hessian can be lower bounded as

$$\begin{aligned} \boldsymbol{v}^*\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a})\boldsymbol{v} &= -\boldsymbol{\gamma}^*\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\gamma} \\ &\ge -\gamma_{(0)}^2\left\|\boldsymbol{P}_{I(\boldsymbol{a})}\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_{(0)}\right\|_2^2 - \sum_{i\ne(0)}\left\|\boldsymbol{P}_{I(\boldsymbol{a})}\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_i\right\|_2^2\gamma_i^2 - 2\sum_{i\ne j}\left|\boldsymbol{e}_i^*\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_j\right|\left|\gamma_i\right|\left|\gamma_j\right| \\ &\ge -\left(1-\boldsymbol{\beta}_{(0)}^2\right)\|\boldsymbol{x}\|_2^2 - \max_{i\ne(0)}\left\|\boldsymbol{P}_{I(\boldsymbol{a})}s_{-i}[\boldsymbol{x}]\right\|_2^2\|\boldsymbol{\gamma}\|_2^2 - 2\max_{i\ne j}\left|\boldsymbol{e}_i^*\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_j\right|\|\boldsymbol{\gamma}\|_1^2, \end{aligned} \tag{F.41}$$

where the second term is bounded by using its expectation derived in Lemma D.2, and utilize $\mathbb{P}\left[|\boldsymbol{s}_i| > \lambda/2\right] < 4\theta$ from (F.38), $\mathbb{E}\boldsymbol{\chi}$ from (F.39) and regional condition $\left|\boldsymbol{\beta}_{(1)}\right| \le \frac{\lambda}{4\log\theta^{-1}}$ to acquire

$$\begin{aligned} \mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})}s_{-i}[\boldsymbol{x}]\right\|_2^2 &= n\theta\left[1 - \mathbb{E}_{\boldsymbol{s}_i}\mathrm{erf}_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) + \mathbb{E}_{\boldsymbol{s}_i}f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i)\right] \\ &\le \frac{|\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i|}{|\boldsymbol{\beta}_i|} + n\theta\cdot\left(\max_{|\boldsymbol{s}_i|\le\frac{\lambda}{2}} f_{\boldsymbol{\beta}_i}(\lambda, \boldsymbol{s}_i) + \mathbb{P}\left[|\boldsymbol{s}_i| > \frac{\lambda}{2}\right]\right) \\ &\le 6n\theta^2 + \frac{2n\theta}{\sqrt{2\pi}}\max_{|\boldsymbol{s}_i|\le\frac{\lambda}{2}}\left(\frac{\lambda+|\boldsymbol{s}_i|}{|\boldsymbol{\beta}_i|}\cdot\exp\left[-\frac{(\lambda-|\boldsymbol{s}_i|)^2}{2\boldsymbol{\beta}_i^2}\right]\right) + 4n\theta^2 \\ &\le 10n\theta^2 + n\theta\cdot\log\theta^{-1}\exp\left(-2\log^2\theta^{-1}\right) \\ &\le 11n\theta^2, \end{aligned} \tag{F.42}$$

and define the events $\mathcal{E}_{\|\boldsymbol{x}\|_2}$, $\mathcal{E}_{\mathrm{cross}}$ and $\mathcal{E}_{\mathrm{pcurv}}$ as follows:

$$\begin{cases} \mathcal{E}_{\mathrm{pcurv}} := \left\{\forall\,\boldsymbol{a}\in\cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu)),\ \left\|\boldsymbol{P}_{I(\boldsymbol{a})}s_{-i}[\boldsymbol{x}]\right\|_2^2 \le 11n\theta^2 + \frac{c_\mu n\theta}{p}\right\} \\ \mathcal{E}_{\mathrm{cross}} := \left\{\forall\,\boldsymbol{a}\in\cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu)), \left|\boldsymbol{\beta}_{(1)}\right| \le \frac{\lambda}{4\log\theta^{-1}},\ \max_{i\ne j\in[\pm p]}\left|\boldsymbol{e}_i^*\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{P}_{I(\boldsymbol{a})}\breve{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{e}_j\right| \le 8n\theta^3\right\} \\ \mathcal{E}_{\|\boldsymbol{x}\|_2} := \left\{\|\boldsymbol{x}\|_2^2 \le n\theta + 3\sqrt{n\theta}\log n\right\} \end{cases}. \tag{F.43}$$

For the Hessian term, on the event $\mathcal{E}_{\mathrm{pcurv}}\cap\mathcal{E}_{\mathrm{cross}}\cap\mathcal{E}_{\|\boldsymbol{x}\|_2}$, and use all $\mu p^2\theta^2$, $\mu p\theta\,|\boldsymbol{\tau}|$ and $\theta\sqrt{p}$ are all less then $\frac{c_\mu}{4\log^2\theta^{-1}}$, from Lemma B.5, and from lemma assumption with sufficiently large $C$ we have $n > \theta^{-1}36\log^2 n$, thus $\boldsymbol{v}^*\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a})\boldsymbol{v}$ can be lower bounded from (F.41) as

$$\begin{aligned} \boldsymbol{v}^*\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a})\boldsymbol{v} &\ge -\left(1-\boldsymbol{\beta}_{(0)}^2\right)\left(n\theta + 3\sqrt{n\theta}\log n\right) - (1+\mu p)\left(11n\theta^2 + \frac{c_\mu n\theta}{p}\right) - 8p\,(1+\mu p)\cdot 8n\theta^3 \\ &\ge -\frac{1}{2}n\theta\cdot(1-\boldsymbol{\beta}_{(0)}^2) - n\theta\cdot\left(\frac{11c_\mu}{4} + c_\mu^2 + \frac{64c_\mu}{4} + \frac{64c_\mu}{4}\right) \\ &\ge -\frac{1}{2}n\theta\cdot\left(1-\boldsymbol{\beta}_{(0)}^2 + 20c_\mu\right). \end{aligned} \tag{F.44}$$

The bounds of $\boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]$ can be derive on the event whose expectation is drawn from Lemma C.2 and (F.39) as

$$\mathcal{E}_\chi := \left\{\begin{cases} \boldsymbol{\sigma}_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i \ge n\theta\mathcal{S}_{\nu_2\lambda}\left[|\boldsymbol{\beta}_i|\right] - \frac{c_\mu n\theta}{p}, & \forall\, i\in[\pm p] \\ \boldsymbol{\sigma}_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i \le 6n\theta^2\left|\boldsymbol{\beta}_i\right| + \frac{c_\mu n\theta}{p^{3/2}}, & \forall\, i\ne(0) \end{cases}\right\},$$

then use $\|\boldsymbol{\beta}\|_1 \le 1 + \frac{\lambda p}{4\log\theta^{-1}} \le \frac{\lambda p}{2}$, implies:

$$\begin{aligned}\boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}] &\ge n\theta\left|\boldsymbol{\beta}_{(0)}\right|\left(\left|\boldsymbol{\beta}_{(0)}\right| - \nu_2\lambda\right) - c_\mu \|\boldsymbol{\beta}\|_1 \tfrac{n\theta}{p} \\ &\ge n\theta\left(\boldsymbol{\beta}_{(0)}^2 - \sqrt{\tfrac{2}{\pi}}\lambda - \tfrac{c_\mu}{2}\lambda\right) \\ &\ge n\theta\left(\boldsymbol{\beta}_{(0)}^2 - \lambda\right). \end{aligned} \tag{F.45}$$

Finally via the regional condition $\left|\boldsymbol{\beta}_{(1)}\right| \le \frac{\lambda}{4\log\theta^{-1}}$, the absolute value of leading correlation

$$\boldsymbol{\beta}_{(0)}^2 \ge \|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 - |\boldsymbol{\tau}|\,\boldsymbol{\beta}_{(1)}^2 \ge 1 - 2c_\mu - 0.1^2 > 0.9, \tag{F.46}$$

then we collect all above results and obtain:

$$\boldsymbol{v}^*\widetilde{\mathrm{Hess}}[\varphi_{\ell_1}](\boldsymbol{a})\boldsymbol{v} = \boldsymbol{v}^*\widetilde{\nabla}^2\varphi_{\ell^1}(\boldsymbol{a})\boldsymbol{v} - \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}] \ge \left(1.5\boldsymbol{\beta}_{(0)}^2 - 0.5 - \lambda - 20c_\mu\right)n\theta \ge 0.3n\theta, \tag{F.47}$$

with probability at least

$$1 - \underbrace{\mathbb{P}\left[\mathcal{E}_{\mathrm{cross}}^c\right]}_{\text{Lemma D.4}} - \underbrace{\mathbb{P}\left[\mathcal{E}_{\mathrm{pcurv}}^c\right]}_{\text{Corollary D.3}} - \underbrace{\mathbb{P}\left[\mathcal{E}_{\|\boldsymbol{x}\|_2}^c\right]}_{\text{Lemma A.2}} - \underbrace{\mathbb{P}\left[\mathcal{E}_{\chi}^c\right]}_{\text{Corollary C.4}} \ge 1 - c'/n. \tag{F.48}$$

3. (Identify local minima) Wlog let $\boldsymbol{a}_*$ be a local minimum where its gradient is zero that is close to $\boldsymbol{a}_0$. The strong convexity (F.47), provides the upper bound on $\|\boldsymbol{a}_* - \boldsymbol{a}_0\|_2^2$ via

$$\begin{aligned} &\varphi_{\ell^1}(\boldsymbol{a}_*) \ge \varphi_{\ell^1}(\boldsymbol{a}_0) + \langle \boldsymbol{a}_* - \boldsymbol{a}_0, \mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}_0)\rangle + \tfrac{0.3}{2}n\theta\,\|\boldsymbol{a}_* - \boldsymbol{a}_0\|_2^2 \\ \Longrightarrow\ &\|\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}_0)\|_2 \ge 0.15n\theta\,\|\boldsymbol{a}_* - \boldsymbol{a}_0\|_2 \end{aligned} \tag{F.49}$$

Thus we only require to bound the gradient at $\boldsymbol{a}_0$, whose coefficients $\boldsymbol{\alpha} = \boldsymbol{e}_0$ and correlation $\boldsymbol{\beta}$ has properties $\boldsymbol{\beta}_0 = 1$ and $\left\|\boldsymbol{\beta}_{\setminus 0}\right\|_\infty \le \mu$ hence $\left\|\boldsymbol{\beta}_{\setminus 0}\right\|_{\le} \sqrt{2p}\mu$. Expand the gradient term and condition on $\mathcal{E}_\chi$, since $\mu p^2\theta^2 \le \frac{c_\mu}{4}$ and $\theta < \frac{c_\mu}{4\sqrt{p}}$, we can upper bound the gradient at $\boldsymbol{a}_0$ as

$$\begin{aligned}\|\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}_0)\|_2 &= \|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\left(\boldsymbol{\chi}[\boldsymbol{\beta}] - \boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{e}_0\right)\|_2 \le \|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\|_2\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\setminus 0}\right\|_2 \\ &\le \sqrt{1+\mu p}\left(6n\theta^2\left\|\boldsymbol{\beta}_{\setminus 0}\right\|_2 + n\theta\cdot\tfrac{c_\mu}{p^{3/2}}\cdot\sqrt{2p}\right) \\ &\le n\theta\sqrt{1+\mu p}\left(6\mu\sqrt{2p}\cdot\theta + \tfrac{2c_\mu}{p}\right) \\ &\le n\theta\left(3c_\mu\mu + 6\mu\cdot\sqrt{2\mu}\cdot p\theta + \tfrac{2c_\mu}{p} + \tfrac{2c_\mu\sqrt{\mu}}{\sqrt{p}}\right) \\ &\le 7\sqrt{c_\mu}n\theta\cdot\max\left\{\mu, \tfrac{1}{p}\right\}. \end{aligned} \tag{F.50}$$

Thus we conclude that with sufficiently small $c_\mu$:

$$\|\boldsymbol{a}_* - \boldsymbol{a}_0\|_2 \le 50\sqrt{c_\mu}\max\left\{\mu, p^{-1}\right\} \le \tfrac{1}{2}\max\left\{\mu, p^{-1}\right\}. \tag{F.51}$$

and we complete the proof by generalize this result from minima near $\boldsymbol{a}_0$ to any of its shifts $s_i[\boldsymbol{a}_0]$.

∎

Similarly, for objective $\varphi_\rho$ we have

**Corollary F.6** (Strong convexity of $\varphi_\rho$ of near shift)**.** *Suppose that $\boldsymbol{x}_0 \sim_{\mathrm{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_\rho$ with $c_\lambda \in \left[\frac{1}{5}, \frac{1}{4}\right]$, then there* *exists some numerical constant $C, c, c', c'', \bar{c} > 0$ such that if $\rho$ is $\delta$-smoothed $\ell^1$ function where $\delta \le c'\lambda\theta^8/p^2\log^2 n$ and $n > Cp^5\theta^{-2}\log p$ and $c_\mu \le \bar{c}$, then with probability at least $1 - c''/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\le k}\Re(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ satisfying $\left|\boldsymbol{\beta}_{(1)}\right| < \nu_1\lambda$: for all $\boldsymbol{v} \in \mathbb{S}^{p-1} \cap \boldsymbol{a}^\perp$,*

$$\boldsymbol{v}^*\widetilde{\mathrm{Hess}}[\varphi_\rho](\boldsymbol{a})\boldsymbol{v} > cn\theta; \tag{F.52}$$

*furthermore, there exists $\bar{\boldsymbol{a}}$ as an local minimizer such that*

$$\min_\ell \|\bar{\boldsymbol{a}} - s_\ell[\boldsymbol{a}_0]\|_2 \le \tfrac{1}{2}\max\left\{\mu, p^{-1}\right\} \tag{F.53}$$

**Proof.** The strong convexity (F.52) is derived by combining (F.36) and (E.23) by letting constant multiplier of $\delta$ satisfies $c'^{1/4} < 10^{-3}c$. On the other hand the local minimizer near solution (F.53) is derived via combining (F.49), (E.21) and utilize both $\theta\sqrt{p} < c_\mu$ and $\mu p^2\theta^2 < c_\mu$ such that:

$$\begin{aligned}
\|\mathrm{grad}[\varphi_\rho](\boldsymbol{a})\|_2 &\le \|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\|_2 \left\|\boldsymbol{\chi}[\boldsymbol{\beta}] - \breve{\boldsymbol{C}}_{\boldsymbol{x}_0}\mathcal{S}_\lambda^\delta\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}\boldsymbol{a}\right]\right\|_2 + \|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\|_2 \left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\setminus 0}\right\|_2 \\
&\le \sqrt{1+\mu p}\cdot n\theta^3 + 7\sqrt{c_\mu}n\theta\cdot\max\left\{\mu, p^{-1}\right\} \\
&\le 8n\theta\sqrt{c_\mu}\cdot\max\left\{\mu, p^{-1}\right\}
\end{aligned} \tag{F.54}$$

■

## F.4 Retraction toward subspace

As in Figure 16, the function value grows in direction away from subspace $\mathcal{S}_{\boldsymbol{\tau}}$, we will illustrate this phenomenon by proving the negative gradient direction $-\boldsymbol{g}$ will point toward the subspace $\mathcal{S}_{\boldsymbol{\tau}}$. To show this, we prove for every coefficients of $\boldsymbol{a}$ as $\boldsymbol{\alpha}$, there exists coefficients of $\boldsymbol{g}$ as $\boldsymbol{\zeta}$ satisfies

$$\langle\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}(\boldsymbol{g}), \boldsymbol{\alpha}_{\boldsymbol{\tau}^c}(\boldsymbol{a})\rangle > c\,\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2\,\|\boldsymbol{\zeta}_{\boldsymbol{\tau}^c}\|_2 \tag{F.55}$$

whenever $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \in \left[\frac{\gamma}{2}, \gamma\right]$. Apparently, the gradient will decrease $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}})$, hence being addressed as *retractive toward subspace $\mathcal{S}_{\boldsymbol{\tau}}$*. This retractive phenomenon is true for gradient of both $\varphi_{\ell^1}$ and $\varphi_\rho$.

**Lemma F.7** (Retraction of $\varphi_{\ell^1}$ toward subspace)**.** *Suppose that $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_{\ell^1}$ with $c_\lambda \in \left(0, \frac{1}{3}\right]$, then there exists some numerical constants $C, c, \bar{c} > 0$ such that if $n > Cp^5\theta^{-2}\log p$ and $c_\mu \le \bar{c}$, then with probability at least $1 - c'/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\le k}\Re(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ such that if*

$$d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \ge \gamma(c_\mu)/2 \tag{F.56}$$

*then for every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{a} = \boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\alpha}$, there exists some $\boldsymbol{\zeta}$ satisfying $\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}) = \boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\zeta}$ that*

$$\langle\boldsymbol{\zeta}_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\rangle \ge \tfrac{1}{4n\theta}\,\|\boldsymbol{\zeta}_{\boldsymbol{\tau}^c}\|_2^2. \tag{F.57}$$

**Proof.** Write $\gamma = \gamma(c_\mu)$ Recall the gradient can be derived as

$$\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}) = -\boldsymbol{P}_{\boldsymbol{a}^\perp}\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\chi}[\boldsymbol{\beta}] = (\boldsymbol{a}\boldsymbol{a}^* - \boldsymbol{I})\,\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\chi}[\boldsymbol{\beta}] = \boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\left(\boldsymbol{\beta}^*\boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{\alpha} - \boldsymbol{\chi}[\boldsymbol{\beta}]\right), \tag{F.58}$$

for every $\boldsymbol{\alpha}$ satisfies $\boldsymbol{a} = \boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\alpha}$. Now via Corollary C.4, condition on the event:

$$\mathcal{E}_\chi := \left\{\sigma_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i \le \begin{cases} n\theta\cdot|\boldsymbol{\beta}_i| + \frac{c_\mu n\theta}{p}, & \forall\, i \in \boldsymbol{\tau} \\ n\theta\cdot|\boldsymbol{\beta}_i|\,4\theta\,|\boldsymbol{\tau}| + \frac{c_\mu n\theta}{p}, & \forall\, i \in \boldsymbol{\tau}^c \end{cases},\quad \sigma_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i \ge n\theta\cdot\mathcal{S}_{\sqrt{2/\pi}\lambda}\left[|\boldsymbol{\beta}_i|\right]\right\}, \tag{F.59}$$

and on this event, utilize Lemma B.5, bounds of $\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}]$ and $\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2$ can be derived with $c_\mu < \frac{1}{100}$ as:

$$\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}] \le n\theta \left( \|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 + 4\theta \left|\boldsymbol{\tau}\right| \|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2^2 + c_\mu \right) \ge n\theta \left(1 + c_\mu + 4c_\mu^2 + c_\mu\right) \le \tfrac{3}{2} n\theta \tag{F.60}$$

$$\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}] \ge n\theta \left( \|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_2^2 - \sqrt{2/\pi}\lambda \|\boldsymbol{\beta}_{\boldsymbol{\tau}}\|_1 - c_\mu \right) \ge n\theta \left(1 - 4c_\mu - \sqrt{2/\pi} c_\lambda - c_\mu \right) \ge \tfrac{1}{2} n\theta \tag{F.61}$$

$$\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2 \le 4n\theta^2 \left|\boldsymbol{\tau}\right| \|\boldsymbol{\beta}_{\boldsymbol{\tau}^c}\|_2 + \tfrac{c_\mu n\theta}{p}\sqrt{p} \le n\theta \left(4c_\mu\gamma + c_\mu\gamma\right) \le \tfrac{1}{20} n\theta\gamma. \tag{F.62}$$

Let $\boldsymbol{\alpha}(\boldsymbol{g}) = \boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{\alpha} - \boldsymbol{\chi}[\boldsymbol{\beta}]$, derive

$$\begin{aligned} &\langle \boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \rangle - \tfrac{1}{4n\theta} \|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c}\|_2^2 \\ &\quad = \ \boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}] \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2^2 - \langle \boldsymbol{\alpha}_{\boldsymbol{\tau}^c}, \boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c} \rangle - \tfrac{1}{4n\theta} \|\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}]\boldsymbol{\alpha}_{\boldsymbol{\tau}^c} - \boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2^2 \\ &\quad \ge \ \boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}] \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2^2 - \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2 - \tfrac{1}{2n\theta} \left|\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}]\right|^2 \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2^2 - \tfrac{1}{2n\theta} \|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2^2 \\ &\quad \ge \ \left(\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}] - \tfrac{1}{2n\theta} (\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}])^2\right) \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2^2 - \tfrac{1}{20} n\theta\gamma \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 - \tfrac{1}{1000} n\theta\gamma^2, \end{aligned} \tag{F.63}$$

notice that this is a quadratic function of $\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}]$ with negative leading coefficient and zeros at $\{0, 2n\theta\}$, hence (F.63) is minimized when $\boldsymbol{\beta}^* \boldsymbol{\chi}[\boldsymbol{\beta}] = \frac{1}{2} n\theta$. Plugging in,

$$\text{(F.63)} \ \ge \ \tfrac{3}{8} n\theta \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2^2 - \tfrac{1}{20} n\theta\gamma \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 - \tfrac{1}{1000} n\theta\gamma^2 \tag{F.64}$$

then again this is a quadratic function of $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2$ with positive leading coefficient and zeros at $\left\{0, \frac{8}{60}\gamma\right\}$, thus (F.64) is minimized at $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 = \frac{\gamma}{2}$. Plugging in again,

$$\text{(F.64)} \ \ge \ \tfrac{3}{8} n\theta \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2^2 - \tfrac{1}{20} n\theta\gamma \|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 - \tfrac{1}{1000} n\theta\gamma^2 \ge \left(\tfrac{3}{32} - \tfrac{1}{80} - \tfrac{1}{1000}\right) n\theta\gamma^2 > 0 \tag{F.65}$$

which concludes our proof. ■

As a consequence, we have that

**Corollary F.8** (Retraction of $\varphi_\rho$ toward the subspace)**.** *Suppose that $\boldsymbol{x}_0 \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$. Define $\lambda = c_\lambda/\sqrt{|k|}$ in $\varphi_\rho$ with $c_\lambda \in \left(0, \frac{1}{3}\right]$, then there exists some numerical constants $C, c, c', c'', \bar{c} > 0$ such that if $\rho$ is $\delta$-smoothed $\ell^1$ function where $\delta \le c''\lambda\theta^8/p^2 \log^2 n$ and $n > Cp^5\theta^{-2}\log p$ and $c_\mu \le \bar{c}$, then with probability at least $1 - c'/n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \le k} \mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ such that if*

$$d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \ge \gamma(c_\mu)/2 \tag{F.66}$$

*then for every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}$, there exists some $\boldsymbol{\zeta}$ satisfying $\mathrm{grad}[\varphi_\rho](\boldsymbol{a}) = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\zeta}$ that*

$$\langle \boldsymbol{\zeta}_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \rangle \ge \tfrac{1}{6n\theta} \|\boldsymbol{\zeta}_{\boldsymbol{\tau}^c}\|_2^2 . \tag{F.67}$$

**Proof.** Write $\gamma = \gamma(c_\mu)$. Define

$$\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] = \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda \left[\check{\boldsymbol{a}} * \boldsymbol{y}\right], \qquad \boldsymbol{\chi}_\rho[\boldsymbol{\beta}] = \check{\boldsymbol{C}}_{\boldsymbol{x}_0} \mathcal{S}_\lambda^\delta \left[\check{\boldsymbol{a}} * \boldsymbol{y}\right],$$

which, and on event (F.59) and Lemma E.6, we know

$$\boldsymbol{\beta}^* \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] \le \tfrac{3}{2} n\theta, \tag{F.68}$$

$$\|\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2 \le \tfrac{1}{20} n\theta\gamma, \tag{F.69}$$

$$\|\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] - \boldsymbol{\chi}_\rho[\boldsymbol{\beta}]\|_2 \le c_1 n\theta^4, \tag{F.70}$$

for some constant $c_1 > 0$. Now given any $\boldsymbol{\alpha}$ satisfies $\boldsymbol{a} = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha}$, the gradient of both objective can be derived as

$$\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}) = -\boldsymbol{P}_{\boldsymbol{a}^\perp} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \operatorname{prox}_{\lambda\|\cdot\|_1} \left[\check{\boldsymbol{a}} * \boldsymbol{y}\right] = \left(\boldsymbol{a}\boldsymbol{a}^* - \boldsymbol{I}\right) \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]$$

$$
\begin{aligned}
&= \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \left( \boldsymbol{\beta}^* \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] \boldsymbol{\alpha} - \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] \right), && \text{(F.71)} \\
\operatorname{grad}[\varphi_\rho](\boldsymbol{a}) &= -\boldsymbol{P}_{\boldsymbol{a}^\perp} \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \operatorname{prox}_{\lambda\rho}[\check{\boldsymbol{a}} * \boldsymbol{y}] = (\boldsymbol{a}\boldsymbol{a}^* - \boldsymbol{I})\, \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\chi}_\rho[\boldsymbol{\beta}] \\
&= \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \left( \boldsymbol{\beta}^* \boldsymbol{\chi}_\rho[\boldsymbol{\beta}] \boldsymbol{\alpha} - \boldsymbol{\chi}_\rho[\boldsymbol{\beta}] \right). && \text{(F.72)}
\end{aligned}
$$

In the same spirit, define the coefficient of each gradient vector

$$
\begin{aligned}
\boldsymbol{\zeta}_{\ell^1} &= \boldsymbol{\beta}^* \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] \boldsymbol{\alpha} - \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}], && \text{(F.73)} \\
\boldsymbol{\zeta}_\rho &= \boldsymbol{\beta}^* \boldsymbol{\chi}_\rho[\boldsymbol{\beta}] \boldsymbol{\alpha} - \boldsymbol{\chi}_\rho[\boldsymbol{\beta}], && \text{(F.74)}
\end{aligned}
$$

which, by norm inequality from (F.68)-(F.70) and Lemma F.7 , we can derive

$$
\begin{aligned}
\left\| \boldsymbol{\zeta}_{\ell^1} - \boldsymbol{\zeta}_\rho \right\|_2 &\le \left\| (\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\beta}^*) \left( \boldsymbol{\chi}_\rho[\boldsymbol{\beta}] - \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] \right) \right\|_2 \le c_1 n \theta^4, && \text{(F.75)} \\
\left\| (\boldsymbol{\zeta}_{\ell^1})_{\boldsymbol{\tau}^c} \right\|_2 &\ge \left| \boldsymbol{\beta}^* \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] \right| \left\| \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \right\|_2 - \left\| \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c} \right\|_2 \ge \tfrac{1}{5} n \theta \gamma, && \text{(F.76)} \\
\left\langle (\boldsymbol{\zeta}_{\ell^1})_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \right\rangle &\ge \tfrac{1}{4n\theta} \left\| (\boldsymbol{\zeta}_{\ell^1})_{\boldsymbol{\tau}^c} \right\|_2^2, && \text{(F.77)}
\end{aligned}
$$

where the first inequality is derived by observing $(\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\beta}^*)$ is a projection operator, as such:

$$
\begin{aligned}
\boldsymbol{\beta}^* \boldsymbol{\alpha} &= \boldsymbol{a}^* \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\alpha} = \boldsymbol{a}^* \boldsymbol{a} = 1, \\
(\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\beta}^*)^2 &= \boldsymbol{I} - 2\boldsymbol{\alpha}\boldsymbol{\beta}^* + \boldsymbol{\alpha}(\boldsymbol{\beta}^*\boldsymbol{\alpha})\boldsymbol{\beta}^* = \boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\beta}^*.
\end{aligned}
$$

Now we are ready to derive (F.67):

$$
\begin{aligned}
\left\langle (\boldsymbol{\zeta}_\rho)_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \right\rangle &\ge \left\langle (\boldsymbol{\zeta}_{\ell^1})_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \right\rangle - \left\| \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \right\|_2 \left\| \boldsymbol{\zeta}_\rho - \boldsymbol{\zeta}_{\ell^1} \right\|_2 \\
&\ge \tfrac{1}{4n\theta} \left\| (\boldsymbol{\zeta}_{\ell^1})_{\boldsymbol{\tau}^c} \right\|_2^2 - c_1 n \theta^4 \gamma \\
&\ge \tfrac{1}{12n\theta} \left\| (\boldsymbol{\zeta}_{\ell^1})_{\boldsymbol{\tau}^c} \right\|_2^2 \\
&\quad + \tfrac{1}{6n\theta} \left( \left\| (\boldsymbol{\zeta}_\rho)_{\boldsymbol{\tau}^c} \right\|_2^2 - 2 \left\| (\boldsymbol{\zeta}_{\ell^1})_{\boldsymbol{\tau}^c} \right\|_2 \left\| \boldsymbol{\zeta}_{\ell^1} - \boldsymbol{\zeta}_\rho \right\|_2 - \left\| \boldsymbol{\zeta}_{\ell^1} - \boldsymbol{\zeta}_\rho \right\|_2^2 \right) - c_1 n \theta^4 \gamma \\
&\ge \tfrac{1}{6n\theta} \left\| (\boldsymbol{\zeta}_\rho)_{\boldsymbol{\tau}^c} \right\|_2^2 + \tfrac{1}{12n\theta} \left( \tfrac{1}{5} n\theta\gamma \right)^2 - \tfrac{1}{3n\theta} \left( \tfrac{1}{5} n\theta\gamma \right) \left( c_1 n \theta^4 \right) - \tfrac{1}{6n\theta} \left( c_1 n \theta^4 \right)^2 - c_1 n \theta^4 \gamma \\
&\ge \tfrac{1}{6n\theta} \left\| (\boldsymbol{\zeta}_\rho)_{\boldsymbol{\tau}^c} \right\|_2^2. && \text{(F.78)}
\end{aligned}
$$

where the last inequality is true since $\theta^3 \ll \gamma$. ∎

## F.5 Proof of Theorem 4.1

By collecting result from above, we are ready to prove the acclaimed geometric result in Theorem 4.1. It guarantees that for every $\boldsymbol{a}$ near $\mathcal{S}_{\boldsymbol{\tau}}$, either one of the following in true

$$
\begin{aligned}
\lambda_{\min} \left( \operatorname{Hess}[\varphi_\rho](\boldsymbol{a}) \right) &\le -c_1 n \theta \lambda, && \text{(F.79)} \\
\left\langle \boldsymbol{\sigma}_{(0)} \boldsymbol{\iota}^* s_{(0)}[\boldsymbol{a}_0], -\operatorname{grad}[\varphi_\rho](\boldsymbol{a}) \right\rangle &\ge c_2 n \theta \left( \log^{-2} \theta^{-1} \right) \lambda^2, && \text{(F.80)} \\
\operatorname{Hess}[\varphi_\rho](\boldsymbol{a}) &\succ c_3 n \theta \cdot \boldsymbol{P}_{\boldsymbol{a}^\perp}, && \text{(F.81)}
\end{aligned}
$$

all local minimizer $\bar{\boldsymbol{a}}$ satisfies for some $\boldsymbol{a}_* \in \left\{ \pm \boldsymbol{\iota}^* s_\ell[\boldsymbol{a}] \,\middle|\, \ell \in [\pm p_0] \right\}$,

$$
\left\| \bar{\boldsymbol{a}} - \boldsymbol{a}_* \right\|_2 \le c_4 \sqrt{c_\mu} \max \left\{ \mu, p_0^{-1} \right\}, \qquad \text{(F.82)}
$$

and whenever $\frac{\gamma}{2} \le d_\alpha \left( \boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}} \right) \le \gamma$, coefficient of $\boldsymbol{a}$ and its gradient $\boldsymbol{g}$, $\boldsymbol{\alpha}$, written as $\boldsymbol{\zeta}$, satisfies

$$
\left\langle \boldsymbol{\zeta}_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \right\rangle \ge \tfrac{c_5}{n\theta} \left\| \boldsymbol{\zeta}_{\boldsymbol{\tau}^c} \right\|_2^2. \qquad \text{(F.83)}
$$

To connect the geometric results introduced in Lemma F.1, Lemma F.3, Lemma F.5 and Lemma F.7, we are only required to prove the required signal condition claimed in Theorem 4.1 is necessary from Definition B.1.

In particular, when the subspace dimension $|\boldsymbol{\tau}| \le 4p_0\theta$. On top of that, we are also required to show the chosen smooth parameter $\delta$ in the pseudo-Huber penalty $\rho(x) = \sqrt{x^2+\delta^2}$ approximate $|x|$ sufficiently well, hence results of Corollary F.2, Corollary F.4, Corollary F.6 and Corollary F.8 also holds.

**Proof.** Firstly we will show when largest solution subspace dimension $k = 4p_0\theta$, the signal condition of Definition B.1 will be satisfied. Recall that the signal condition of Theorem 4.1 requests

$$\frac{2}{p_0 \log^2 p_0} \le \theta \le \frac{c}{\left(p_0\sqrt{\mu} + \sqrt{p_0}\right)\log^2 p_0}, \tag{F.84}$$

since $p = 3p_0 - 2$, this implies the lower bounds for sparsity $\theta$ as:

$$\theta \ge \frac{1}{2p_0\left(\frac{1}{2}\log p_0\right)^2} \ge \frac{1}{p \log^2 \theta^{-1}}; \tag{F.85}$$

the upper bound of $\theta$ via $\theta\sqrt{p_0}\log^2 p_0 \le c$ :

$$\theta \le \frac{9c}{\sqrt{p_0}(3\log p_0)^2} \le \frac{16c}{\sqrt{p}\log^2\theta^{-1}}, \qquad \theta \le \frac{4c^2}{k\log^4 p_0} \le \frac{36c^2}{k(3\log p_0)^2} \le \frac{36c^2}{k\log^2\theta^{-1}}; \tag{F.86}$$

and the upper bound for coherence $\mu$ as:

$$\mu \max\left\{k^2, (p\theta)^2\right\}\log^2\theta^{-1} \le \mu\max\left\{16(p_0\theta)^2, 9(p_0\theta)^2\right\}\log^2\theta^{-1} \le 16\left(\sqrt{\mu}p_0\theta\right)^2\log^2 p_0 \le 16c. \tag{F.87}$$

Therefore Definition B.1 holds if $\max\left\{16c, 36c^2\right\} \le c_\mu/4$ via (F.85)-(F.87).

Furthermore, we know from lemma assumption all interested $\boldsymbol{a}$ are near subspace $\mathcal{S}_{\boldsymbol{\tau}}$ by

$$d_\alpha(\boldsymbol{a}.\mathcal{S}_{\boldsymbol{\tau}}) \le \frac{c}{\sqrt{p_0}\log^2\theta^{-1}}\cdot\min\left\{\frac{1}{\sqrt{\theta}}, \frac{1}{\sqrt{\mu}}.\frac{1}{\mu\,(p_0\theta)^{3/2}}\right\} \le \frac{c}{\log^2\theta^{-1}}\min\left\{\frac{2}{\sqrt{k}}, \frac{1}{\sqrt{p_0\mu}}, \frac{4}{\mu p_0\sqrt{\theta}k}\right\} \le \gamma \tag{F.88}$$

where $\gamma$ is defined in Definition B.3 of widened subspace $\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$.

Lastly, the pseudo-Huber function $\rho(x) = \sqrt{x^2+\delta^2}$ is an $\ell^1$ smoothed sparse surrogate defined in Definition E.2, by observing that it is convex, smooth, even, whose second order derivative (according to Table 1) $\nabla^2\rho(x) = \frac{\delta^2}{(x^2+\delta^2)^{3/2}}$ is monotone decreasing in $|x|$. More importantly

$$\sup_{x\in\mathbb{R}} |\rho(x) - |x|| = |\rho(0) - |0|| = \delta. \tag{F.89}$$

Hence, by choosing $\delta \le \frac{c'^4\theta^8}{p^2\log^2 n}\lambda$, for some sufficiently small constant $c'$ and letting $\lambda = 0.2\sqrt{k} = 0.1/\sqrt{p_0\theta}$ in $\varphi_\rho$. We obtain the geometrical results in Corollary F.2 when $\left|\boldsymbol{\beta}_{(1)}\right| \ge \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|$, Corollary F.4 when $\frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right| \ge \left|\boldsymbol{\beta}_{(1)}\right| \ge \frac{\lambda}{4\log^2\theta^{-1}}$ and Corollary F.6 when $\frac{\lambda}{4\log^2\theta^{-1}} \ge \left|\boldsymbol{\beta}_{(1)}\right|$, and the retraction result in Corollary F.8. ∎

# G Analysis of algorithm — minimization within widened subspace

In this section, we prove convergence of the first part of our algorithm—minimization of $\varphi_\rho$ near $\mathcal{S}_{\boldsymbol{\tau}}$. We begin by proving the initialization method guarantees that $\boldsymbol{a}^{(0)}$ is near $\mathcal{S}_\tau$, in the sense that

$$d_\alpha(\boldsymbol{a}^{(0)}, \mathcal{S}_{\boldsymbol{\tau}}) \le \gamma, \tag{G.1}$$

where the distance $d_\alpha$ is defined in (4.15). We then demonstrate that small-stepping curvilinear search converges to a desired local minimum of $\varphi_\rho$ at rate $O(1/k)$, where $k$ is the iteration number. To do this, it is important to utilize(i) the *retractive* property to show that the iterates stay near $\mathcal{S}_{\boldsymbol{\tau}}$ and (ii) the geometric properties of $\varphi_\rho$ near $\mathcal{S}_{\boldsymbol{\tau}}$.

## G.1 Initialization near subspace

The following lemma shows that the initialization $\boldsymbol{a}^{(0)} = \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\nabla\varphi_{\ell^1}(\boldsymbol{a}^{(-1)})\right]$, where

$$\boldsymbol{a}^{(-1)} = \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\textstyle\sum_{\ell\in\boldsymbol{\tau}} \boldsymbol{x}_{0\ell}\boldsymbol{\iota}_{p_0}^* s_\ell[\boldsymbol{a}_0]\right], \tag{G.2}$$

and is very close to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$:

**Lemma G.1** (Initialization from a piece of data)**.** *Let $\overline{\boldsymbol{x}} \in \mathbb{R}^{2p_0-1}$ indexed by $[\pm p_0]$, with $\overline{\boldsymbol{x}}_i \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$. Define $\overline{\boldsymbol{y}} = \overline{\boldsymbol{x}} * \boldsymbol{a}_0$, and $\boldsymbol{a}^{(0)}$ as*

$$\boldsymbol{a}^{(0)} = -\boldsymbol{P}_{\mathbb{S}^{p-1}}\nabla\varphi_{\ell^1}\left(\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{0}^{p_0-1}; [\overline{\boldsymbol{y}}_0; \cdots; \overline{\boldsymbol{y}}_{p_0-1}]; \boldsymbol{0}^{p_0-1}\right]\right), \tag{G.3}$$

*with $\lambda = 0.2/\sqrt{p\theta}$ in $\varphi_1$. Set $\boldsymbol{\tau} = \mathrm{supp}(\overline{\boldsymbol{x}})$. Suppose that $(\boldsymbol{a}_0, \theta, k)$ satisfies the sparsity-coherence condition $\mathrm{SCC}(c_\mu)$ and $\boldsymbol{a}_0$ satisfies $\max_{i\neq j}\left|\left\langle \boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0], \boldsymbol{\iota}_{p_0}^* s_j[\boldsymbol{a}_0]\right\rangle\right| \le \mu$. Then there exists some constant $c, \overline{c} > 0$ such that if $p_0\theta > 1000c$ and $c_\mu \le \overline{c}$, then with probability at least $1 - 1/c$, we have*

$$d_\alpha\left(\boldsymbol{a}^{(0)}, \mathcal{S}_{\boldsymbol{\tau}}\right) \le \frac{c_\mu}{4\log^2\theta^{-1}} \min\left\{\frac{1}{\sqrt{|\boldsymbol{\tau}|}}, \frac{1}{\sqrt{\mu p}}, \frac{1}{\mu p\sqrt{\theta}\,|\boldsymbol{\tau}|}\right\}. \tag{G.4}$$

**Proof.** 1. (Distance to $\mathcal{S}_{\boldsymbol{\tau}}$ from $\boldsymbol{a}^{(0)}$) Let $\eta = \left\|\boldsymbol{\iota}_{p_0}^*(\boldsymbol{a}_0 * \boldsymbol{x})\right\|_2 = \left\|\boldsymbol{\iota}_{p_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{x}\right\|_2$ and $\gamma = \gamma(c_\mu)$, as in (G.4). Expand the expression of $\boldsymbol{a}^{(0)}$ from (G.3) we have

$$\boldsymbol{a}^{(0)} = \boldsymbol{P}_{\mathbb{S}^{p-1}}\boldsymbol{\iota}^*\breve{\boldsymbol{C}}_{\boldsymbol{y}}\mathcal{S}_\lambda\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}}\boldsymbol{\iota}_{p_0}\boldsymbol{P}_{\mathbb{S}^{p_0-1}}\boldsymbol{\iota}_{p_0}^*(\boldsymbol{a}_0 * \boldsymbol{x})\right] = \boldsymbol{P}_{\mathbb{S}^{p-1}}\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\chi}\left[\tfrac{1}{\eta}\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{\iota}_{p_0}\boldsymbol{\iota}_{p_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{x}\right] \tag{G.5}$$

To relate $\boldsymbol{a}^{(0)}$ to its coefficient, introduce the truncated autocorrelation matrix $\widetilde{\boldsymbol{M}} = \boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{\iota}_{p_0}\boldsymbol{\iota}_{p_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}$, define $\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}$ as

$$\widetilde{\boldsymbol{\beta}} = \tfrac{1}{\eta}\widetilde{\boldsymbol{M}}\boldsymbol{x}, \quad \widetilde{\boldsymbol{\alpha}} = \boldsymbol{\chi}\left[\tfrac{1}{\eta}\widetilde{\boldsymbol{M}}\boldsymbol{x}\right] = \boldsymbol{\chi}[\widetilde{\boldsymbol{\beta}}] \tag{G.6}$$

and note that $\widetilde{\boldsymbol{M}}$ is bounded entrywise as

$$\left|\widetilde{\boldsymbol{M}}_{ij}\right| \le \begin{cases} 1 & i = j \in [-p_0+1, p_0-1] \\ \mu & i \neq j \in [-p_0+1, p_0-1],\ |i-j| < p_0 \\ 0 & \text{otherwise} \end{cases}. \tag{G.7}$$

From (G.5), we can write $\boldsymbol{a}^{(0)} = \boldsymbol{P}_{\mathbb{S}^{p-1}}\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\widetilde{\boldsymbol{\alpha}}$, meaning that the normalized version of $\widetilde{\boldsymbol{\alpha}}$ is a valid coefficient vector for $\boldsymbol{a}^{(0)}$. Let $\boldsymbol{\tau}^c = [\pm 2p_0] \setminus \boldsymbol{\tau}$. The distance $d_{\boldsymbol{\alpha}}$ to subspace $\mathcal{S}_{\boldsymbol{\tau}}$ (4.15) is upper bounded as

$$d_\alpha(\boldsymbol{a}^{(0)}, \mathcal{S}_{\boldsymbol{\tau}}) \le \frac{\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2}{\|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\widetilde{\boldsymbol{\alpha}}\|_2} \le \frac{\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2}{\|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\|_2 - \|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2}$$

$$\leq \frac{\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2}{\sqrt{1-\mu\,|\boldsymbol{\tau}|}\,\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\|_2 - \sqrt{1+\mu p}\,\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2}$$

where the last inequality is derived with Lemma B.4. Therefore, it is sufficient to show

$$\left(1+\gamma\sqrt{1+\mu p}\right)\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2 \leq \gamma\sqrt{1-\mu\,|\boldsymbol{\tau}|}\,\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\|_2 \tag{G.8}$$

to complete the proof that $d_\alpha(\boldsymbol{a}^{(0)}, \mathcal{S}_{\boldsymbol{\tau}}) \leq \gamma$.

2. (Bound $\eta$) Condition on the following two events

$$\mathcal{E}_\tau := \{|\boldsymbol{\tau}| < 4p_0\theta\}, \quad \mathcal{E}_{\|\boldsymbol{x}\|_2} := \left\{\sqrt{p_0\theta} \leq \|\boldsymbol{x}\|_2 \leq \sqrt{3p_0\theta}\right\} \tag{G.9}$$

and utilize $\mu$ bound from Lemma B.5 such that $\mu\,|\boldsymbol{\tau}| < 0.1$. An upper bound on $\eta$ can be obtained using properties of $\widetilde{\boldsymbol{M}}$ of (G.7):

$$\eta = \left\|\boldsymbol{\iota}_{p_0}^* \boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{x}\right\|_2 \leq \|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{x}\|_2 \leq \sqrt{1+\mu\,|\boldsymbol{\tau}|}\,\|\boldsymbol{x}\|_2 \leq 2\sqrt{p_0\theta} \tag{G.10}$$

To lower bound $\eta$, use $\eta^2 = \boldsymbol{g}^*\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}}\boldsymbol{g}$ where $\boldsymbol{g}$ is the standard Gaussian vector. Observe the submatrix of $\widetilde{\boldsymbol{M}}$ is diagonal dominant:

$$\begin{cases} \widetilde{\boldsymbol{M}}_{ii} = \left\|\boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0]\right\|_2^2 \in [0,1] \\ \operatorname{tr}\left(\widetilde{\boldsymbol{M}}\right) = \sum\limits_{i\in[\pm p_0]} \left\|\boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0]\right\|_2^2 = \|\boldsymbol{a}_0\|_2^2 + \sum\limits_{i=1}^{p_0-1}\left(\left\|\boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0]\right\|_2^2 + \left\|\boldsymbol{\iota}_{p_0}^* s_{i-p_0}[\boldsymbol{a}_0]\right\|_2^2\right) = p_0 \end{cases}. \tag{G.11}$$

Write $\boldsymbol{x} = \boldsymbol{g}\circ\boldsymbol{w}$ where $\boldsymbol{w}$ and $\boldsymbol{g}$ are Bernoulli and Gaussian vector respectively with $\operatorname{supp}(\boldsymbol{w}) = \boldsymbol{\tau}$, then the trace of $\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}}$ can be written as sum of independent r.v.s as:

$$\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}}\right) = \sum_{i\in[\pm p_0]} w_i \left\|\boldsymbol{\iota}_{p_0}^* s_i[\boldsymbol{a}_0]\right\|_2^2,$$

Bernstein inequality Lemma J.4 and (G.11) gives

$$\mathbb{P}\left[\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}}\right) < \frac{3p_0\theta}{4}\right] \leq \mathbb{P}\left[\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}}\right) - p_0\theta \leq -\frac{p_0\theta}{4}\right] \leq 2\exp\left(\frac{-(p_0\theta/4)^2}{2p_0\theta + p_0\theta/2}\right) \leq 2\exp\left(\frac{-p_0\theta}{40}\right), \tag{G.12}$$

thus condition on $\boldsymbol{\omega}$ satisfies $\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}}\right) \geq 3p_0\theta/4$ and $\mathcal{E}_{\boldsymbol{\tau}}$, expectation $\eta^2$ has lower bound

$$\mathbb{E}_{\boldsymbol{g}|\boldsymbol{w}}\eta^2 = \mathbb{E}_{\boldsymbol{g}|\boldsymbol{w}}\left[\boldsymbol{g}^*\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}}\boldsymbol{g}\right] = \operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}}\right) \geq \frac{3p_0\theta}{4}$$

then apply Bernstein inequality again by first writing svd of $\boldsymbol{P}_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{P}_{\boldsymbol{\tau}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^*$ with $\boldsymbol{\Sigma}$ being rank $|\boldsymbol{\tau}| < 4p_0\theta$ and square orthobasis $\boldsymbol{U}$. Let $\boldsymbol{g}' = \boldsymbol{U}^*\boldsymbol{g}$, then $\boldsymbol{g}'$ is standard i.i.d. Gaussian vector, provides alternative expression $\eta^2 < \sum_{i=1}^{4p_0\theta} g_i'^2\sigma_i$ where $\sigma_i \leq 1+\mu\,|\boldsymbol{\tau}| \leq 1.1$. We obtain probability of $\eta^2$ to be small as

$$\mathbb{P}_{\boldsymbol{g}|\boldsymbol{w}}\left[\eta^2 < \frac{p_0\theta}{2}\right] \leq \mathbb{P}_{\boldsymbol{g}|\boldsymbol{w}}\left[\eta^2 - \mathbb{E}_{\boldsymbol{g}|\boldsymbol{w}}\eta^2 < -\frac{p_0\theta}{4}\right] \leq 2\exp\left(\frac{-(p_0\theta/4)^2}{2(1+\mu\,|\boldsymbol{\tau}|)(12p_0\theta + p_0\theta/2)}\right) \leq 2\exp\left(\frac{-p_0\theta}{440}\right) \tag{G.13}$$

by applying moment bounds $(\sigma^2, R) = (12p_0\theta(1+\mu|\boldsymbol{\tau}|), 2(1+\mu|\boldsymbol{\tau}|))$. We thereby define event

$$\mathcal{E}_\eta = \left\{ \sqrt{p_0\theta/2} \le \eta \le 2\sqrt{p_0\theta} \right\}, \tag{G.14}$$

which holds w.h.p. based on (G.9), (G.12) and (G.13).

3. (Bound $\widetilde{\boldsymbol{\alpha}}$) Condition on $\mathcal{E}_\eta \cap \mathcal{E}_{\|\boldsymbol{x}\|_2} \cap \mathcal{E}_\tau$. Use definition $\widetilde{\boldsymbol{\beta}} = \frac{1}{\eta}\widetilde{\boldsymbol{M}}\boldsymbol{x}$ from (G.6), and properties of $\widetilde{\boldsymbol{M}}$ from (G.7) we can obtain:

$$\begin{cases} \|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}^c}\|_2 \le \frac{1}{\eta}\left\|\boldsymbol{\iota}^*_{\boldsymbol{\tau}^c}\widetilde{\boldsymbol{M}}\boldsymbol{\iota}_{\boldsymbol{\tau}}\right\|_2 \|\boldsymbol{x}\|_2 \le \frac{\mu\sqrt{p_0|\boldsymbol{\tau}|}}{\sqrt{p_0\theta/2}} \cdot \sqrt{3p_0\theta} \le 3\mu\sqrt{p_0|\boldsymbol{\tau}|} \\ \|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}}\|_2 \ge \frac{1}{\eta}\left\|\boldsymbol{\iota}^*_{\boldsymbol{\tau}}\widetilde{\boldsymbol{M}}\boldsymbol{\iota}_{\boldsymbol{\tau}}\right\|_2 \|\boldsymbol{x}\|_2 \ge \frac{\sqrt{1-\mu|\boldsymbol{\tau}|}}{2\sqrt{p_0\theta}} \cdot \sqrt{p_0\theta} \ge 0.45 \end{cases}. \tag{G.15}$$

Use definition $\|\widetilde{\boldsymbol{\alpha}}\|_2 = \|\boldsymbol{\chi}[\widetilde{\boldsymbol{\beta}}]\|_2$, condition on event

$$\mathcal{E}_\chi := \left\{ \begin{cases} \sigma_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i \ge n\theta\mathcal{S}_{\nu_2\lambda}\left[|\boldsymbol{\beta}_i|\right] - \frac{c_\mu^2 n\theta}{p}, & \forall i \in \boldsymbol{\tau} \\ \sigma_i\boldsymbol{\chi}[\boldsymbol{\beta}]_i \le 4n\theta^2|\boldsymbol{\tau}||\boldsymbol{\beta}_i| + \frac{c_\mu n\theta}{p}, & \forall i \in \boldsymbol{\tau}^c \end{cases} \right\},$$

also from Definition B.1 we have $\mu(p\theta)^{1/2}|\boldsymbol{\tau}|^{3/2} < \frac{c_\mu}{4\log^2\theta^{-1}}$ and from lemma assumption $\lambda = \frac{1}{5\sqrt{p\theta}}$, provides bounds of $\|\widetilde{\boldsymbol{\alpha}}\|_2$ via triangle inequality as:

$$\begin{cases} \|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2 \le 4n\theta^2|\boldsymbol{\tau}| \cdot \|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}^c}\|_2 + \frac{c_\mu n\theta}{p} \cdot \sqrt{2p_0} \le 3c_\mu n\theta\left(\frac{\sqrt{\theta}}{\log^2\theta^{-1}} + \frac{c_\mu}{p}\right) \\ \|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\|_2 \ge n\theta\left(\|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}}\|_2 - \nu_2\lambda\sqrt{|\boldsymbol{\tau}|} - \frac{c_\mu}{p}\sqrt{|\boldsymbol{\tau}|}\right) \ge n\theta\left(0.45 - \sqrt{\frac{2}{\pi}} \cdot \frac{1}{5} - c_\mu\right) \ge 0.2n\theta \end{cases}, \tag{G.16}$$

since both $\theta|\boldsymbol{\tau}|$, $\mu p\theta|\boldsymbol{\tau}| < c_\mu$, we have

$$\begin{cases} \sqrt{1+\mu p}\,\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2 \le 3c_\mu n\theta\sqrt{1+\mu p}\left(\sqrt{\theta} + p^{-1}\right) \le 6c_\mu n\theta \\ \|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2 \le \frac{6c_\mu^{3/2}n\theta}{\log^2\theta^{-1}} \min\left\{\frac{1}{\sqrt{|\boldsymbol{\tau}|}}, \frac{1}{\sqrt{\mu p}}, \frac{1}{\mu p\sqrt{\theta}|\boldsymbol{\tau}|}\right\} \le 24\sqrt{c_\mu}n\theta\gamma \end{cases},$$

which satisfies (G.8), since $\mu|\boldsymbol{\tau}| < c_\mu < \frac{1}{1000}$,

$$\left(1 + \gamma\sqrt{1+\mu p}\right)\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^c}\|_2 \le \left(24\sqrt{c_\mu} + 6c_\mu\right)n\theta\gamma \le 0.1n\theta\gamma \le \gamma\sqrt{1-\mu|\boldsymbol{\tau}|}\,\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\|_2\,. \tag{G.17}$$

Finally, given $p_0\theta > 1000c$, this result holds with probability at least

$$1 - \underbrace{\mathbb{P}\left[\mathcal{E}_\tau^c\right]}_{\text{Lemma A.1}} - \underbrace{\mathbb{P}\left[\mathcal{E}^c_{\|\boldsymbol{x}\|_2}\right]}_{\text{Lemma A.2}} - \underbrace{\mathbb{P}\left[\mathcal{E}_\eta^c\right]}_{\text{(G.14)}} - \underbrace{\mathbb{P}\left[\mathcal{E}_\chi^c\right]}_{\text{Corollary C.4}} \ge 1 - \frac{2}{p_0\theta} - \frac{1}{n} - 4\exp\left(\frac{-p_0\theta}{440}\right) \ge 1 - \frac{1}{c} \tag{G.18}$$

■

## G.2 Minimization near subspace (Proof of Theorem 5.1)

Before we start the proof of theorem, writing $\boldsymbol{g} = \mathrm{grad}[\varphi_\rho](\boldsymbol{a})$ and $\boldsymbol{H} = \mathrm{Hess}[\varphi_\rho](\boldsymbol{a})$, we will first restate the results of Theorem 4.1 in simplified terms. The theorem shows that for any $\boldsymbol{a} \in \mathbb{S}^{p-1}$ whose distance to subspace $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \le \gamma$, then at least one of the the following statement hold:

$$\|\boldsymbol{g}\|_2 \ge \eta_g \tag{G.19}$$

$$\lambda_{\min}(\boldsymbol{H}) \le -\eta_v \tag{G.20}$$
$$\boldsymbol{H} \succ \eta_c \cdot \boldsymbol{P}_{\boldsymbol{a}^\perp}. \tag{G.21}$$

Furthermore, $\varphi_\rho$ is retractive near $\mathcal{S}_{\boldsymbol{\tau}}$: wherever $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \ge \frac{\gamma}{2}$, writing $\boldsymbol{\alpha}(\boldsymbol{a})$, $\boldsymbol{\alpha}(\boldsymbol{g})$ to be the coefficient of $\boldsymbol{a}$, $\boldsymbol{g}$, we have

$$\langle \boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c} \rangle \ge \eta_r \left\| \boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c} \right\|_2. \tag{G.22}$$

Also, the the gradient, Hessian and the third order derivative are all bounded as follows:

**Remark G.2.** *With high probability, for every $\boldsymbol{a}$ whose $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) < \gamma$, its $\max\{\|\boldsymbol{g}\|_2, \|\boldsymbol{H}\|_2, \|\nabla \boldsymbol{H}\|_2\} \le \bar{\eta} = \operatorname{poly}(n,p)$.*

We state Remark G.2 without explicit proof since its derivation is similar to the proof in Theorem 4.1.

We prove that if the negative curvature direction $-\boldsymbol{v}$ is chosen to be the least eigenvector with $\boldsymbol{v}^*\boldsymbol{H}\boldsymbol{v} < -\eta_v$ and $\boldsymbol{v}^*\boldsymbol{g}$ (if cannot, let $\boldsymbol{v} = \boldsymbol{0}$), then the iterates

$$\boldsymbol{a}^{(k+1)} = \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a}^{(k)} - t\boldsymbol{g}^{(k)} - t^2\boldsymbol{v}^{(k)}\right] \tag{G.23}$$

converges toward the minimizer $\bar{\boldsymbol{a}}$ in $\ell^2$-norm with rate $O(1/k)$. Notice that here all $\eta_g$, $\eta_v$, $\eta_c$, $\eta_r$, $\bar{\eta}$ are all greater then $0$ and are rational functions of the dimension parameters $n, p$.

Finally, we should note that $\boldsymbol{a}_0$ being $\mu$-truncated shift coherent implies that $\boldsymbol{a}_0$ is at at most $2\mu$-shift coherent. Hence we utilize the usual incoherence condition in the proof.

**Proof.** Notice that when $\boldsymbol{a}$ is in the region near some signed shift $\bar{\boldsymbol{a}}$ of $\boldsymbol{a}_0$, the function $\varphi_\rho$ is strongly convex, and the iterates coincide with the Riemannian gradient method, which converges at a linear rate. Indeed, if for all $k$ larger than some $\bar{k}$, $\boldsymbol{a}^{(k)}$ is in this region, then $\left\|\boldsymbol{a}^{(k)} - \bar{\boldsymbol{a}}\right\|_2 \le (1 - t\eta_c)^{-(k-\bar{k})} \|\boldsymbol{a}^{(\bar{k})} - \bar{\boldsymbol{a}}\|_2$ [AMS09](Theorem 4.5.6) where the step size $t = \Omega(1/n\theta)$ hence $t\eta_c = \Omega(1)$. We will argue that the iterates $\boldsymbol{a}^{(k)}$ remain close to the subspace $\mathcal{S}_{\boldsymbol{\tau}}$ and that after $\bar{k} = \operatorname{poly}(n,p)$ iterations they indeed remain in the strongly convex region around some $\bar{\boldsymbol{a}}$.

1. (Existence of Armijo steplength). First, we show there exists a nontrivial step size $t$ at every iteration, in the sense that for all $\boldsymbol{a} \in \mathbb{S}^{p-1}$, there exists $T > 0$ such that for all $t \in (0, T)$, the Armijo step condition (5.11) is satisfied. Note that since $\varphi_\rho$ is a smooth function, $\boldsymbol{a} \to \varphi_\rho \circ \boldsymbol{P}_{\mathbb{S}^{p-1}}(\boldsymbol{a})$ admits a version of Taylor's theorem (see also [AMS09](Section 7.1.3)): for any $\boldsymbol{\xi} \perp \boldsymbol{a}$, writing $\boldsymbol{a}^+ = \boldsymbol{P}_{\mathbb{S}^{p-1}}[\boldsymbol{a} + \boldsymbol{\xi}]$,

$$\left|\varphi_\rho(\boldsymbol{a}^+) - \left(\varphi_\rho(\boldsymbol{a}) + \langle \operatorname{grad}[\varphi_\rho](\boldsymbol{a}), \boldsymbol{\xi}\rangle + \tfrac{1}{2}\boldsymbol{\xi}^*\operatorname{Hess}[\varphi_\rho](\boldsymbol{a})\boldsymbol{\xi}\right)\right| \le \bar{\eta}\,\|\boldsymbol{\xi}\|_2^3, \tag{G.24}$$

using $\|\nabla \boldsymbol{H}\|_2 \le \bar{\eta}$. Now, let $\boldsymbol{\xi} = -t\boldsymbol{g} - t^2\boldsymbol{v}$ as in the iterates (5.10). Suppose the Armijo step condition (5.11) does not hold, so

$$\varphi_\rho(\boldsymbol{a}^+) > \varphi_\rho(\boldsymbol{a}) - \tfrac{1}{2}\left(t\|\boldsymbol{g}\|_2^2 + \tfrac{1}{2}t^4\eta_v\|\boldsymbol{v}\|_2^2\right). \tag{G.25}$$

Since $\boldsymbol{g}^*\boldsymbol{v} \ge 0$ and $\boldsymbol{v}^*\boldsymbol{H}\boldsymbol{v} \le -\eta_v\|\boldsymbol{v}\|_2^2$ or $\boldsymbol{v} = \boldsymbol{0}$, using $\|\boldsymbol{a} + \boldsymbol{b}\|_2^3 \le 4\|\boldsymbol{a}\|_2^3 + 4\|\boldsymbol{b}\|_2^3$ (Hölder's inequality) and $\|\boldsymbol{H}\|_2 < \bar{\eta}$, we can derive

$$\begin{aligned}
&\left\langle \boldsymbol{g}, -t\boldsymbol{g} - t^2\boldsymbol{v}\right\rangle + \tfrac{1}{2}(t\boldsymbol{g} + t^2\boldsymbol{v})^*\boldsymbol{H}\left(t\boldsymbol{g} + t^2\boldsymbol{v}\right) + c\left\|t\boldsymbol{g} + t^2\boldsymbol{v}\right\|_2^3 > -\tfrac{1}{2}\left(t\|\boldsymbol{g}\|_2^2 + \tfrac{1}{2}t^4\eta_v\|\boldsymbol{v}\|_2^2\right) \\
\Longrightarrow\ & -\tfrac{1}{2}t\|\boldsymbol{g}\|_2^2 + \tfrac{1}{2}t^2\boldsymbol{g}^*\boldsymbol{H}\boldsymbol{g} + t^3\boldsymbol{v}^*\boldsymbol{H}\boldsymbol{g} - \tfrac{1}{4}t^4\eta_v\|\boldsymbol{v}\|_2^2 + 4\bar{\eta}t^3\|\boldsymbol{g}\|_2^3 + 4\bar{\eta}t^6\|\boldsymbol{v}\|_2^3 > 0 \\
\Longrightarrow\ & -\tfrac{1}{2}t\|\boldsymbol{g}\|_2^2 + t^2\left(\tfrac{1}{2}\bar{\eta}\|\boldsymbol{g}\|_2^2 + t\bar{\eta}\|\boldsymbol{v}\|_2\|\boldsymbol{g}\|_2 + 4\bar{\eta}t\|\boldsymbol{g}\|_2^3\right) - \tfrac{1}{4}t^4\eta_v\|\boldsymbol{v}\|_2^2 + 4\bar{\eta}t^6\|\boldsymbol{v}\|_2^3 > 0.
\end{aligned} \tag{G.26}$$

If

$$t < T = \min\left\{\frac{\|\boldsymbol{g}\|_2}{\bar{\eta}\|\boldsymbol{g}\|_2 + 2\bar{\eta}t\|\boldsymbol{v}\|_2 + 8\bar{\eta}t\|\boldsymbol{g}\|_2^2}, \sqrt{\frac{\eta_v}{16\bar{\eta}\|\boldsymbol{v}\|_2}}\right\}, \tag{G.27}$$

then (G.26) $< 0$ contradicting (G.25). Using our bounds on $\|\boldsymbol{g}\|_2$, $\bar{\eta}$, $\eta_v$ and $\|\boldsymbol{v}\|$, it follows that $T$ is lower bounded by a polynomial $\mathrm{poly}\left(n^{-1}, p^{-1}\right)$.

2.(Bounds on $d_\alpha(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}})$, $d_\alpha(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}})$) We will show there are numerical constants $c_g$, $c_v$ such that

$$d_\alpha(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}}) \le c_g n\theta\gamma \qquad \text{and} \qquad d_\alpha(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}) \le c_v n\theta p. \tag{G.28}$$

Define

$$\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] = \check{\boldsymbol{C}}_{\boldsymbol{x}_0}\, \mathrm{prox}_{\lambda\ell^1}\left[\check{\boldsymbol{a}} * \boldsymbol{y}\right], \qquad \boldsymbol{\chi}_\rho[\boldsymbol{\beta}] = \check{\boldsymbol{C}}_{\boldsymbol{x}_0}\, \mathrm{prox}_{\lambda\rho}\left[\check{\boldsymbol{a}} * \boldsymbol{y}\right],$$

then the gradient can be written as (F.71)

$$\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}) = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}\left(\boldsymbol{\beta}^* \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]\boldsymbol{\alpha} - \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]\right), \tag{G.29}$$

$$\mathrm{grad}[\varphi_\rho](\boldsymbol{a}) = \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}\left(\boldsymbol{\beta}^* \boldsymbol{\chi}_\rho[\boldsymbol{\beta}]\boldsymbol{\alpha} - \boldsymbol{\chi}_\rho[\boldsymbol{\beta}]\right). \tag{G.30}$$

Use the following inequalities:

$$\begin{aligned} \tfrac{1}{2}n\theta \le \left|\boldsymbol{\beta}^* \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]\right| &\le \tfrac{3}{2}n\theta, \\ \left\|\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\right\|_2 &\le \tfrac{1}{20}n\theta\gamma, \\ \left\|\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\beta}^*\right\|_2 &\le 4\sqrt{p}, \\ \left\|\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}] - \boldsymbol{\chi}_\rho[\boldsymbol{\beta}]\right\|_2 &\le n\theta^4, \end{aligned}$$

where the first and second bounds of $\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]$ based on event (F.59); the third by observing $\|\boldsymbol{\alpha}\|_2 \le 2$ and $\|\boldsymbol{\beta}\|_2 \le 2 + c_\mu\sqrt{p}$; the last from (E.21) of Lemma E.6 when $\delta$ is sufficiently small. Hence, by definition of $d_\alpha(\,\cdot, \mathcal{S}_{\boldsymbol{\tau}})$ (4.15) and knowing $\boldsymbol{a}$ is close to subspace $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \le \gamma$, via triangle inequality, we get

$$\begin{aligned} d_\alpha(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}}) &\le d_\alpha(\mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}), \mathcal{S}_{\boldsymbol{\tau}}) + d_\alpha(\mathrm{grad}[\varphi_\rho](\boldsymbol{a}) - \mathrm{grad}[\varphi_{\ell^1}](\boldsymbol{a}), \mathcal{S}_{\boldsymbol{\tau}}) \\ &\le \left\|\boldsymbol{\beta}^* \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]\boldsymbol{\alpha}_{\boldsymbol{\tau}^c} - \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\right\|_2 + \left\|\left(\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\beta}^*\right)\left(\boldsymbol{\chi}_\rho[\boldsymbol{\beta}] - \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]\right)\right\|_2. \\ &\le \tfrac{3}{2}n\theta\gamma \;+\; \tfrac{1}{20}n\theta\gamma \;+\; 4\sqrt{p}n\theta^4 \\ &\le 3n\theta\gamma. \end{aligned} \tag{G.31}$$

To bound the $d_\alpha$ norm of least eigenvector $\boldsymbol{v}$, note that $\boldsymbol{\beta}^*\boldsymbol{\chi}_\rho[\boldsymbol{\beta}] > 0$, we can conclude

$$\boldsymbol{v}^* \nabla^2 \varphi_\rho(\boldsymbol{a})\boldsymbol{v} \;\le\; \boldsymbol{v}^* \boldsymbol{P}_{\boldsymbol{a}^\perp} \nabla^2 \varphi_\rho(\boldsymbol{a}) \boldsymbol{P}_{\boldsymbol{a}^\perp}\boldsymbol{v} + \boldsymbol{\beta}^*\boldsymbol{\chi}_\rho[\boldsymbol{\beta}] \;=\; \boldsymbol{v}^*\boldsymbol{H}\boldsymbol{v} \;<\; -\eta_v,$$

expand $\nabla^2\varphi_\rho(\boldsymbol{a})$ as in (E.8), and since $\boldsymbol{v}$ is the eigenvector of smallest eigenvalue $\lambda_{\min} < -\eta_v$,

$$\boldsymbol{P}_{\boldsymbol{a}^\perp} \nabla^2 \varphi_\rho(\boldsymbol{a}) \boldsymbol{P}_{\boldsymbol{a}^\perp}\boldsymbol{v} \;=\; \left(\boldsymbol{I} - \boldsymbol{a}\boldsymbol{a}^*\right)\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\nabla\mathrm{prox}_{\lambda\rho}\left[\check{\boldsymbol{a}} * \boldsymbol{y}\right]\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{\iota}\boldsymbol{v} \;=\; \lambda_{\min}\boldsymbol{v}, \tag{G.32}$$

hence there exists $\boldsymbol{\alpha}(\boldsymbol{v})$ satisfies $\boldsymbol{v} = \boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\alpha}(\boldsymbol{v})$ and

$$\boldsymbol{\alpha}(\boldsymbol{v}) \;=\; \lambda_{\min}^{-1}\left[\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\nabla\mathrm{prox}_{\lambda\rho}\left[\check{\boldsymbol{a}} * \boldsymbol{y}\right]\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{\iota}\boldsymbol{v} - \left(\boldsymbol{\beta}^*\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\nabla\mathrm{prox}_{\lambda\rho}\left[\check{\boldsymbol{a}} * \boldsymbol{y}\right]\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{C}^*_{\boldsymbol{a}_0}\boldsymbol{\iota}\boldsymbol{v}\right)\boldsymbol{\alpha}\right].$$

Now since $\nabla\mathrm{prox}_{\lambda\rho}\left[\check{\boldsymbol{a}} * \boldsymbol{y}\right]$ is a diagonal matrix with entries in $[0, 1]$,

$$d_\alpha(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}) \;\le\; \|\boldsymbol{\alpha}(\boldsymbol{v})\|_2 \;\le\; |\lambda_{\min}|^{-1}\,\|\boldsymbol{\iota}\boldsymbol{C}_{\boldsymbol{a}_0}\|_2\,\|\boldsymbol{x}_0\|_1^2\,\|\boldsymbol{v}\|_2\left(1 + \|\boldsymbol{\alpha}\|_2\,\|\boldsymbol{\beta}\|_2\right) \;<\; c_v n\theta p, \tag{G.33}$$

where we use upper bound of $\|\boldsymbol{x}_0\|_1 < cn\theta$ from Lemma A.2 and $|\lambda_{\min}| > \eta_v > cn\theta\lambda$ from Corollary F.2.

3. (Iterates stay within widened subspace). Suppose (G.22) holds. We will show that whenever

$$t \le T' = \frac{1}{10n\theta}, \tag{G.34}$$

then setting $\boldsymbol{a}^+ = \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a} - t\boldsymbol{g} - t^2\boldsymbol{v}\right]$, we have

$$\left|d_\alpha\left(\boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}}\right) - d_\alpha\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)\right| \leq \tfrac{\gamma}{2}, \tag{G.35}$$

and whenever $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) \in \left[\frac{\gamma}{2}, \gamma\right]$

$$d_\alpha^2\left(\boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq d_\alpha^2\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) - t \cdot c' n\theta\gamma^2. \tag{G.36}$$

If both (G.35) and (G.36) hold, then all iterates $\boldsymbol{a}^{(k)}$ will stay near the subspace: $d_\alpha(\boldsymbol{a}^{(k)}, \mathcal{S}_{\boldsymbol{\tau}}) < \gamma$.

To derive (G.35), since both $\boldsymbol{g} \perp \boldsymbol{a}$ and $\boldsymbol{v} \perp \boldsymbol{a}$ we have $\left\|\boldsymbol{a} - t\boldsymbol{g} - t^2\boldsymbol{v}\right\|_2^2 = \|\boldsymbol{a}\|_2^2 + \left\|t\boldsymbol{g} + t^2\boldsymbol{v}\right\|_2^2 > 1$, and since $d_\alpha(\cdot, \mathcal{S}_{\boldsymbol{\tau}})$ is a seminorm Lemma B.2:

$$\begin{aligned} d_\alpha\left(\boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}}\right) &= d_\alpha(\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a} - t\boldsymbol{g} - t^2\boldsymbol{v}\right], \mathcal{S}_{\boldsymbol{\tau}}) \leq d_\alpha\left(\boldsymbol{a} - t\boldsymbol{g} - t^2\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}\right) \\ &\leq d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) + t d_\alpha(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}}) + t^2 d_\alpha(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}) \end{aligned} \tag{G.37}$$

suggests (G.35) holds via (G.28) and let $n > Cp^5\theta^{-2}$, we have

$$t d_\alpha(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}}) + t^2 d_\alpha(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}) \leq \tfrac{c_g n\theta\gamma}{10n\theta} + \tfrac{c_v n\theta p}{(10n\theta)^2} < \tfrac{\gamma}{2} \tag{G.38}$$

with sufficiently large $C$.

To derive (G.36), let $\boldsymbol{\alpha}(\boldsymbol{a})$ to be a coefficient vector satisfying $d_\alpha(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) = \|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}\|_2$, and based on (G.30) and (G.33), define

$$\boldsymbol{\alpha}(\boldsymbol{g}) = \boldsymbol{\beta}^*\boldsymbol{\chi}_\rho[\boldsymbol{\beta}]\boldsymbol{\alpha}(\boldsymbol{a}) - \boldsymbol{\chi}_\rho[\boldsymbol{\beta}] \tag{G.39}$$

$$\boldsymbol{\alpha}(\boldsymbol{v}) = \lambda_{\min}^{-1}\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\nabla\mathrm{prox}_{\lambda\rho}\left[\check{\boldsymbol{a}} * \boldsymbol{y}\right]\check{\boldsymbol{C}}_{\boldsymbol{x}_0}\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{\iota}\boldsymbol{v}. \tag{G.40}$$

By the retraction property and norm bounds,

$$\langle\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c}\rangle \geq \tfrac{1}{6n\theta}\|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c}\|_2^2 \tag{G.41}$$

$$\|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}\|_2 \leq \gamma \tag{G.42}$$

$$\|\boldsymbol{\alpha}(\boldsymbol{v})\|_2 \leq c_v n\theta p. \tag{G.43}$$

Since $\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 > \frac{\gamma}{2}$,

$$\begin{aligned} \|\boldsymbol{a}(\boldsymbol{g})_{\boldsymbol{\tau}^c}\|_2 &\geq \|\boldsymbol{\beta}^*\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]\boldsymbol{\alpha}_{\boldsymbol{\tau}^c} \;-\; \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2 - \|(\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\beta}^*)\left(\boldsymbol{\chi}_\rho[\boldsymbol{\beta}] - \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]\right)\|_2 \\ &\geq |\boldsymbol{\beta}^*\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]|\,\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^c}\|_2 \;-\; \|\boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^c}\|_2 \;-\; \|(\boldsymbol{I} - \boldsymbol{\alpha}\boldsymbol{\beta}^*)\|_2\,\|(\boldsymbol{\chi}_\rho[\boldsymbol{\beta}] - \boldsymbol{\chi}_{\ell^1}[\boldsymbol{\beta}])\|_2 \\ &\geq \tfrac{1}{2}n\theta \times \tfrac{\gamma}{2} \;-\; \tfrac{1}{20}n\theta\gamma \;+\; 2n\theta^4 \\ &\geq \tfrac{1}{10}n\theta\gamma. \end{aligned} \tag{G.44}$$

Finally, we can bound $d_\alpha(\boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}})$ as

$$\begin{aligned} d_\alpha^2(\boldsymbol{a}^+, \mathcal{S}_{\boldsymbol{\tau}}) &\leq d_\alpha^2(\boldsymbol{a} - t\boldsymbol{g} - t^2\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}) \\ &\leq \left\|\left[\boldsymbol{\alpha}(\boldsymbol{a}) - t\boldsymbol{\alpha}(\boldsymbol{g}) - t^2\boldsymbol{\alpha}(\boldsymbol{v})\right]_{\boldsymbol{\tau}^c}\right\|_2^2 \\ &= \|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}\|_2^2 \;-\; 2t\left\langle\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}, \left[\boldsymbol{\alpha}(\boldsymbol{g}) + t\boldsymbol{\alpha}(\boldsymbol{v})\right]_{\boldsymbol{\tau}^c}\right\rangle \;+\; t^2\left\|\left[\boldsymbol{\alpha}(\boldsymbol{g}) + t\boldsymbol{\alpha}(\boldsymbol{v})\right]_{\boldsymbol{\tau}^c}\right\|_2^2 \\ &\leq \|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}\|_2^2 \;-\; 2t\left\langle\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}, \boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c}\right\rangle \;+\; 2t^2\|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^c}\|_2\|\boldsymbol{\alpha}(\boldsymbol{v})\|_2 \;+\; 2t^2\|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c}\|_2^2 \;+\; 2t^4\|\boldsymbol{\alpha}(\boldsymbol{v})\|_2^2 \\ &\leq d^2(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) - 2t\left[\left(\tfrac{1}{3n\theta} - t\right)\|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^c}\|_2^2 - tn\theta p\gamma - t^3(c_v n\theta p)^2\right] \\ &\leq d^2(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}) - t \cdot c' n\theta\gamma^2 \end{aligned} \tag{G.45}$$

where the last inequality holds when $t < \frac{0.1}{n\theta}$ with sufficiently large $n$.

4. (Polynomial time convergence) The iterates $\boldsymbol{a}^{(k)}$ remain within a $\gamma$ neighborhood of $\mathcal{S}_{\boldsymbol{\tau}}$ for all $k$. At any iteration $k$, $\boldsymbol{a}^{(k)}$ is in at least one of three regions: strong gradient, negative curvature, or strong convexity. In the gradient and curvature regions, we obtain a decrease in the function value which is at least some (nonzero) rational function of $n$ and $p$. On the strongly convex region, the function value does not increase. The suboptimality at initialization is bounded by a polynomial in $n$ and $p$,$\mathrm{poly}(n,p)$, and hence the total number of steps in the gradient and curvature regions is bounded by a polynomial in $n, p$. After the iterates reach the strongly convex region, the number of additional steps required to achieve $\|\boldsymbol{a}^{(k)} - \bar{\boldsymbol{a}}\|_2 \leq \varepsilon$ is bounded by $\mathrm{poly}(n,p)\log \varepsilon^{-1}$. In particular, the number of iterations required to achieve $\|\boldsymbol{a}^{(k)} - \bar{\boldsymbol{a}}\|_2 \leq \mu + 1/p$ is bounded by a polynomial in $(n, p)$, as claimed. ■

## H Analysis of algorithm — local refinement

In this section, we describe and analyze an algorithm which refines an estimate $\boldsymbol{a}^{(0)} \approx \boldsymbol{a}_0$ of the kernel to exactly recover $(\boldsymbol{a}_0, \boldsymbol{x}_0)$. Set

$$\lambda^{(0)} \leftarrow 5\kappa_I \widetilde{\mu} \qquad \text{and} \qquad I^{(0)} \leftarrow \operatorname{supp}(\mathcal{S}_\lambda \left[\boldsymbol{C}^*_{\boldsymbol{a}^{(0)}} \boldsymbol{y}\right]), \tag{H.1}$$

where as each iteration of the algorithm consists of the following key steps:

- **Sparse Estimation using Reweighted Lasso:** Set

$$\boldsymbol{x}^{(k+1)} \;\leftarrow\; \operatorname*{argmin}_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{a}^{(k)} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \sum_{i \notin I^{(k)}} \lambda^{(k)} \left|\boldsymbol{x}_i\right| ; \tag{H.2}$$

- **Kernel Estimation using Least Squares:** Set

$$\boldsymbol{a}^{(k+1)} \;\leftarrow\; \boldsymbol{P}_{\mathbb{S}^{p-1}} \big[ \operatorname*{argmin}_{\boldsymbol{a}} \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)} - \boldsymbol{y}\|_2^2 \big]; \tag{H.3}$$

- **Continuation and reweighting by decreasing sparsity regularizer:** Set

$$\lambda^{(k+1)} \leftarrow \tfrac{1}{2}\lambda^{(k)} \qquad \text{and} \qquad I^{(k+1)} \;\leftarrow\; \operatorname{supp}(\boldsymbol{x}^{(k+1)}). \tag{H.4}$$

Our analysis will show that $\boldsymbol{a}^{(k)}$ converges to $\boldsymbol{a}_0$ at a linear rate. In the remainder of this section, we describe the assumptions of our analysis. In subsequent sections, we prove key lemmas analyzing each of the three main steps of the algorithm.

**Modified coherence and rate assumptions** Below, we will write

$$\widetilde{\mu} = \max\left\{\mu, p^{-1}\right\}. \tag{H.5}$$

Our refinement algorithm will demand an initialization satisfying

$$\|\boldsymbol{a}^{(0)} - \boldsymbol{a}_0\|_2 \le \widetilde{\mu}. \tag{H.6}$$

**Support density of $\boldsymbol{x}_0$** Our goal is to show that the proposed annealing algorithm exactly solves the SaS deconvolution problem, i.e., exactly recovers $(\boldsymbol{a}_0, \boldsymbol{x}_0)$ up to a signed shift. We will denote the support sets of true sparse vector $\boldsymbol{x}_0$ and recovered $\boldsymbol{x}^{(k)}$ in the intermediate $k$-th steps as

$$I = \operatorname{supp}(\boldsymbol{x}_0), \qquad\qquad I^{(k)} = \operatorname{supp}(\boldsymbol{x}^{(k)}). \tag{H.7}$$

It should be clear that exact recovery is unlikely if $\boldsymbol{x}_0$ contains many consecutive nonzero entries: in this situation, even *non-blind* deconvolution fails. We introduce the notation $\kappa_I$ as an upper bound for number of nonzero entries of $\boldsymbol{x}_0$ in a length-$p$ window:

$$\kappa_I = 6\max\left\{\theta p, \log n\right\}, \tag{H.8}$$

then in the Bernoulli-Gaussian model, with high probability,

$$\max_{\ell} \left| I \cap ([p] + \ell)\right| \le \kappa_I. \tag{H.9}$$

Here, indexing and addition should be interpreted modulo $n$. The $\log n$ term reflects the fact that as $n$ becomes enormous (exponential in $p$) eventually it becomes likely that some length-$p$ window of $\boldsymbol{x}_0$ is densely occupied. In our main theorem statement, we preclude this possibility by putting an upper bound on $n$ (w.r.t $\widetilde{\mu}$). We find it useful to also track the maximum $\ell^2$ norm of $\boldsymbol{x}_0$ over any length-$p$ window:

$$\|\boldsymbol{x}_0\|_{\square} := \max_{\ell} \left\|\boldsymbol{P}_{([p]+\ell)}\boldsymbol{x}_0\right\|_2 . \tag{H.10}$$

Below, we will sometimes work with the $\square$-induced operator norm:

$$\|\boldsymbol{M}\|_{\square\to\square} = \sup_{\|\boldsymbol{x}\|_{\square}\le 1} \|\boldsymbol{M}\boldsymbol{x}\|_{\square} \tag{H.11}$$

For now, we note that in the Bernoulli-Gaussian model, $\|\boldsymbol{x}_0\|_{\square}$ is typically not large

$$\|\boldsymbol{x}_0\|_{\square} \;\le\; \sqrt{\kappa_I}. \tag{H.12}$$

## H.1 Reweighted Lasso finds the large entries of $\boldsymbol{x}_0$

The following lemma asserts that when $\boldsymbol{a}$ is close to $\boldsymbol{a}_0$, the reweighted Lasso finds all of the large entries of $\boldsymbol{x}_0$. Our reweighted Lasso is modified version from [CWB08], we only penalize $\boldsymbol{x}$ on entries outside of its known support subset. We write $T$ to be the subset of true support $I$, and define the sparsity surrogate as

$$\sum_{i\,\in T^c} |\boldsymbol{x}_i| \tag{H.13}$$

The reweighted Lasso recovers more accurate $\boldsymbol{x}$ on set $T$ compares to the vanilla Lasso problem, it turns out to be very helpful in our analysis which proves convergence of the proposed alternating minimization.

**Lemma H.1** (Accuracy of reweighted Lasso estimate)**.** *Suppose that $\boldsymbol{y} = \boldsymbol{a}_0 * \boldsymbol{x}_0$ with $\boldsymbol{a}_0$ is $\widetilde{\mu}$-shift coherent and $\|\boldsymbol{x}_0\|_{\square} \le \sqrt{\kappa_I}$ with $\kappa_I \ge 1$. If $\widetilde{\mu}\kappa_I^2 \le c_\mu$, then for every $T \subseteq I$ and $\boldsymbol{a}$ satisfying $\|\boldsymbol{a}-\boldsymbol{a}_0\|_2 \le \widetilde{\mu}$, the solution $\boldsymbol{x}^+$ to the optimization problem*

$$\min_{\boldsymbol{x}} \left\{ \tfrac{1}{2}\|\boldsymbol{a} * \boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda \sum_{i\,\in T^c} |\boldsymbol{x}_i| \right\}, \tag{H.14}$$

*with*

$$\lambda \;>\; 5\kappa_I\|\boldsymbol{a}-\boldsymbol{a}_0\|_2, \tag{H.15}$$

*is unique with the form*

$$\boldsymbol{x}^+ \;=\; \boldsymbol{\iota}_J\left(\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}J}\right)^{-1}\boldsymbol{\iota}_J^*\left(\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{y} - \lambda\boldsymbol{P}_{J\setminus T}\boldsymbol{\sigma}\right) \tag{H.16}$$

*where $\boldsymbol{\sigma} = \operatorname{sign}(\boldsymbol{x}^+)$. Its support set $J$ satisfies*

$$(\,T \cup I_{\ge 3\lambda}\,) \;\subseteq\; J \;\subseteq\; I \tag{H.17}$$

*and its entrywise error is bounded as*

$$\left\|\boldsymbol{x}^+ - \boldsymbol{x}_0\right\|_\infty \;\le\; 3\lambda. \tag{H.18}$$

*Above, $c_\mu > 0$ is a positive numerical constant.*

We prove Lemma H.1 below. The proof relies heavily on the fact that when $\boldsymbol{a}_0$ is shift-incoherent and $\boldsymbol{a} \approx \boldsymbol{a}_0$, $\boldsymbol{a}$ is also shift-incoherent, an observation which is formalized in a sequence of calculations in Appendix H.4.

**Proof.** 1. (Restricted support Lasso problem). We first consider the restricted problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^{|I|}} \left\{ \tfrac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{\iota}_I\boldsymbol{w} - \boldsymbol{y}\right\|_2^2 + \lambda \sum_{i\in T^c} |(\boldsymbol{\iota}_I\boldsymbol{w})_i| \right\}. \tag{H.19}$$

Under our assumptions, provided $c < \frac{1}{9}$, Lemma H.6 implies that

$$\boldsymbol{\iota}_I^* \boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{\iota}_I = [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}}]_{I,I} \succ \mathbf{0}, \tag{H.20}$$

and the restricted problem is strongly convex and its solution is unique. The KKT conditions imply that a vector $\boldsymbol{w}_\star$ is the unique optimal solution to this problem if and only if

$$\boldsymbol{\iota}_I^* \boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{\iota}_I \boldsymbol{w}_\star \in \boldsymbol{\iota}_I^* \boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{y} - \lambda \, \partial \left\| \boldsymbol{P}_{T^c} \left[ \cdot \right] \right\|_1 (\boldsymbol{w}_\star). \tag{H.21}$$

Writing $J = \mathrm{supp}(\boldsymbol{\iota}_I \boldsymbol{w}_\star) \subseteq I$, $\boldsymbol{C}_{\boldsymbol{a}\,J} = \boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{\iota}_J$, $\boldsymbol{w}_J = \boldsymbol{\iota}_J^* \boldsymbol{\iota}_I \boldsymbol{w}_\star$ the corresponding sub-vector containing the nonzero entries of $\boldsymbol{w}_\star$ and $\boldsymbol{\sigma}_{J \setminus T} = \boldsymbol{\iota}_J^* \boldsymbol{P}_{T^c} \left[ \mathrm{sign}(\boldsymbol{\iota}_I \boldsymbol{w}_*) \right]$, the condition (H.21) is satisfied if and only if

$$\boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \boldsymbol{w}_J = \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{y} - \lambda \boldsymbol{\sigma}_{J \setminus T}, \tag{H.22}$$

$$\left\| \boldsymbol{C}_{\boldsymbol{a}\,I \setminus J}^* \left( \boldsymbol{C}_{\boldsymbol{a}\,J} \boldsymbol{w}_J - \boldsymbol{y} \right) \right\|_\infty \le \lambda. \tag{H.23}$$

We will argue that under our assumptions, $J$ necessarily contains all of the large entries of $\boldsymbol{x}_0$:

$$I_{>3\lambda} = \{ \ell \in I \mid |\boldsymbol{x}_{0\ell}| > 3\lambda \} \subseteq J. \tag{H.24}$$

We show this by contradiction – namely, if some large entry of $\boldsymbol{x}_0$ is not in $J$, then the dual condition (H.23) is violated, contradicting the optimality of $\boldsymbol{w}_\star$. To this end, note that by Corollary H.7, $\boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J}$ has full rank. From (H.22),

$$\boldsymbol{w}_J = \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{y} - \lambda \boldsymbol{\sigma}_{J \setminus T} \right]. \tag{H.25}$$

Write $\boldsymbol{x}_{0\,J} = \boldsymbol{\iota}_J^* \boldsymbol{x}_0$ and $(\boldsymbol{x}_0)_{I \setminus J} = \boldsymbol{P}_{I \setminus J} \boldsymbol{x}_0$. We can further notice that

$$\begin{aligned} \boldsymbol{C}_{\boldsymbol{a}J} \boldsymbol{w}_J - \boldsymbol{y} \;&=\; \left( \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* - \boldsymbol{I} \right) \boldsymbol{y} - \lambda \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{\sigma}_{J \setminus T} \\ &=\; \left( \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* - \boldsymbol{I} \right) \boldsymbol{C}_{\boldsymbol{a}_0\,J} \boldsymbol{x}_{0\,J} + \left( \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* - \boldsymbol{I} \right) \boldsymbol{C}_{\boldsymbol{a}_0 I \setminus J} (\boldsymbol{x}_0)_{I \setminus J} \\ &\quad - \lambda \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{\sigma}_{J \setminus T} \\ &=\; \left( \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* - \boldsymbol{I} \right) \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}\,J} \boldsymbol{x}_{0\,J} + \left( \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* - \boldsymbol{I} \right) \boldsymbol{C}_{\boldsymbol{a}_0 I \setminus J} (\boldsymbol{x}_0)_{I \setminus J} \\ &\quad - \lambda \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{\sigma}_{J \setminus T}, \end{aligned} \tag{H.26}$$

where in the final line we have used that

$$\left( \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* - \boldsymbol{I} \right) \boldsymbol{C}_{\boldsymbol{a}\,J} = \mathbf{0}. \tag{H.27}$$

Suppose that $J$ is a strict subset of $I$ (otherwise, if $J = I$, we are done). Take any $i \in I \setminus J$ such that $|\boldsymbol{x}_{0i}| = \left\| (\boldsymbol{x}_0)_{I \setminus J} \right\|_\infty$, and let $\xi = \mathrm{sign}(\boldsymbol{x}_{0i})$. Using (H.26), Corollary H.7 and Lemma H.8, we have

$$\begin{aligned} -\xi s_i[\boldsymbol{a}]^* \left( \boldsymbol{C}_{\boldsymbol{a}J} \boldsymbol{w}_J - \boldsymbol{y} \right) \;&=\; \xi s_i[\boldsymbol{a}]^* \left( \boldsymbol{I} - \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* \right) s_i[\boldsymbol{a}_0] \boldsymbol{x}_{0i} \\ &\quad + \xi s_i[\boldsymbol{a}]^* \left( \boldsymbol{I} - \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* \right) \boldsymbol{C}_{\boldsymbol{a}_0} (\boldsymbol{x}_0)_{I \setminus (J \cup \{i\})} \\ &\quad + \xi s_i[\boldsymbol{a}]^* \left( \boldsymbol{I} - \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a}\,J}^* \right) \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}\,J} \boldsymbol{x}_{0\,J} \\ &\quad + \xi \lambda s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}\,J} \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \boldsymbol{\sigma}_{J \setminus T} \qquad\qquad \text{(H.28)} \\ \ge\;& \left( \langle s_i[\boldsymbol{a}], s_i[\boldsymbol{a}_0] \rangle \;-\; \| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}\,J} \|_1 \left\| \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \right\|_{\infty \to \infty} \| \boldsymbol{C}_{\boldsymbol{a}\,J}^* s_i[\boldsymbol{a}_0] \|_\infty \right) \left\| (\boldsymbol{x}_0)_{I \setminus J} \right\|_\infty \\ &\quad - \left( \left\| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}_0 I \setminus \{i\}} \right\|_1 \;+\; \| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}\,J} \|_1 \left\| \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \right\|_{\infty \to \infty} \left\| \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}_0 I \setminus J} \right\|_{\infty \to \infty} \right) \left\| (\boldsymbol{x}_0)_{I \setminus J} \right\|_\infty \\ &\quad - \left( \| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}\,J} \|_2 \;+\; \| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}\,J} \|_2 \left\| \left[ \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}\,J} \right]^{-1} \right\|_{\square \to \square} \| \boldsymbol{C}_{\boldsymbol{a}\,J}^* \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}\,J} \|_{\square \to \square} \right) \sqrt{2} \, \| \boldsymbol{x}_0 \|_\square \end{aligned}$$

$$
\begin{aligned}
& \quad - \lambda \left\| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a} J} \right\|_1 \left\| \left[ \boldsymbol{C}_{\boldsymbol{a} J}^* \boldsymbol{C}_{\boldsymbol{a} J} \right]^{-1} \right\|_{\infty \to \infty} \left\| \boldsymbol{\sigma}_{J \setminus T} \right\|_\infty && \text{(H.29)} \\
\geq & \left( (1 - \|\boldsymbol{a} - \boldsymbol{a}_0\|_2) \; - \; C_1 \kappa_I \widetilde{\mu} \times 1 \times \widetilde{\mu} \right) \left\| (\boldsymbol{x}_0)_{I \setminus J} \right\|_\infty \\
& \quad - C_2 \left( \kappa_I \widetilde{\mu} \; + \; \kappa_I \widetilde{\mu} \times 1 \times \kappa_I \widetilde{\mu} \right) \left\| (\boldsymbol{x}_0)_{I \setminus J} \right\|_\infty \\
& \quad - \left( 2 \sqrt{\kappa_I} \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \; + \; C_3 \sqrt{\kappa_I} \widetilde{\mu} \times 1 \times \kappa_I \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \right) \|\boldsymbol{x}_0\|_\square \\
& \quad - \lambda \, C_4 \kappa_I \widetilde{\mu} && \text{(H.30)} \\
\geq & \left( 1 \; - \; C_1' \kappa_I \widetilde{\mu} - C_2 \left( \kappa_I \widetilde{\mu} \right)^2 \right) \|(\boldsymbol{x}_0)_{I \setminus J}\|_\infty \\
& \quad - 2 \kappa_I \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 - \left( C_3 \kappa_I^{3/2} \widetilde{\mu} \right) \kappa_I \left\| \boldsymbol{a} - \boldsymbol{a}_0 \right\|_2 - \left( C_4 \kappa_I \widetilde{\mu} \right) \lambda && \text{(H.31)} \\
\geq & \; \tfrac{1}{2} \left\| (\boldsymbol{x}_0)_{I \setminus J} \right\|_\infty - \lambda / 2, && \text{(H.32)}
\end{aligned}
$$

where the last line holds provided $\widetilde{\mu} \kappa_I^2 \leq c_\mu$ to be a sufficiently small numerical constants. If $\|(\boldsymbol{x}_0)_{I \setminus J}\|_\infty > 3\lambda$, this is strictly larger than $\lambda$, implying that $\left| \boldsymbol{a}_i^* \left( \boldsymbol{C}_{\boldsymbol{a} J} \boldsymbol{w}_J - \boldsymbol{y} \right) \right| > \lambda$, and contradicting the KKT conditions for the restricted problem. Hence, under our assumptions

$$
\left\| (\boldsymbol{x}_0)_{I \setminus J} \right\|_\infty \leq 3\lambda. \tag{H.33}
$$

2. (Solution of Full Lasso problem) We next argue that the solution of the restricted support Lasso problem, $\boldsymbol{w}_J$, when extended to $\mathbb{R}^n$ as $\boldsymbol{x}^+ = \boldsymbol{\iota}_J \boldsymbol{w}_J$, is the unique optimal solution to the *full* Lasso problem

$$
\min_{\boldsymbol{x}} \; \varphi_{\text{lasso}}(\boldsymbol{x}) \;\; \equiv \;\; \tfrac{1}{2} \left\| \boldsymbol{a} * \boldsymbol{x} - \boldsymbol{y} \right\|_2^2 + \lambda \sum_{i \in T^c} |\boldsymbol{x}_i| \, . \tag{H.34}
$$

To prove that $\boldsymbol{x}^+$ is the unique optimal solution, it suffices to show that for every $i \in I^c$,

$$
\left| \, s_i[\boldsymbol{a}]^* (\boldsymbol{a} * \boldsymbol{x}^+ - \boldsymbol{y}) \, \right| < \lambda. \tag{H.35}
$$

Indeed, suppose that this inequality is in force. Write $\varepsilon = \lambda - \max_{i \in I^c} \left| s_i[\boldsymbol{a}]^* (\boldsymbol{a} * \boldsymbol{x}^+ - \boldsymbol{y}) \right|$, and notice that from the KKT conditions for the restricted problem,

$$
\mathbf{0} \; \in \; \boldsymbol{P}_I \partial_{\boldsymbol{x}} \varphi_{\text{lasso}}(\boldsymbol{x}) \tag{H.36}
$$

Combining with (H.35), we have that for every vector $\boldsymbol{\zeta}$ with $\operatorname{supp}(\boldsymbol{\zeta}) \subseteq I^c$ and $\|\boldsymbol{\zeta}\|_\infty \leq 1$, then $\varepsilon \boldsymbol{\zeta} \in \partial \varphi_{\text{lasso}}(\boldsymbol{x}^+)$. Let $\boldsymbol{x}'$ be any vector with $\boldsymbol{x}'_{I^c} \neq \mathbf{0}$ and set $\boldsymbol{\zeta} = \mathcal{P}_{I^c} \operatorname{sign}(\boldsymbol{x}')$, then from the subgradient inequality,

$$
\begin{aligned}
\varphi_{\text{lasso}}(\boldsymbol{x}') & \geq \varphi_{\text{lasso}}(\boldsymbol{x}^+) + \left\langle \varepsilon \boldsymbol{\zeta}, \boldsymbol{x}' - \boldsymbol{x}^+ \right\rangle \\
& \geq \varphi_{\text{lasso}}(\boldsymbol{x}^+) + \varepsilon \left\| \boldsymbol{x}'_{I^c} \right\|_1 ,
\end{aligned} \tag{H.37}
$$

which is strictly larger than $\varphi_{\text{lasso}}(\boldsymbol{x}^+)$. Hence, when (H.35) holds, any optimal solution $\bar{\boldsymbol{x}}$ to the full Lasso problem must satisfy $\operatorname{supp}(\bar{\boldsymbol{x}}) \subseteq I$. By strong convexity of the restricted problem, the solution to (H.34) is unique and equal to $\boldsymbol{x}^+$.

We finish by showing (H.35). Using the same expansion as above, we obtain

$$
\begin{aligned}
\left| s_i[\boldsymbol{a}]^* (\boldsymbol{C}_{\boldsymbol{a} J} \boldsymbol{w}_J - \boldsymbol{y}) \right| \;\; \leq & \;\; \left| s_i[\boldsymbol{a}]^* \left( \boldsymbol{I} - \boldsymbol{C}_{\boldsymbol{a} J} \left[ \boldsymbol{C}_{\boldsymbol{a} J}^* \boldsymbol{C}_{\boldsymbol{a} J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a} J}^* \right) \boldsymbol{C}_{\boldsymbol{a}_0 I \setminus J} (\boldsymbol{x}_0)_{I \setminus J} \right| \\
& + \; \left| s_i[\boldsymbol{a}]^* \left( \boldsymbol{I} - \boldsymbol{C}_{\boldsymbol{a} J} \left[ \boldsymbol{C}_{\boldsymbol{a} J}^* \boldsymbol{C}_{\boldsymbol{a} J} \right]^{-1} \boldsymbol{C}_{\boldsymbol{a} J}^* \right) \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a} \, J} \boldsymbol{x}_{0 \, J} \right| \\
& + \; \lambda \left| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a} J} \left[ \boldsymbol{C}_{\boldsymbol{a} J}^* \boldsymbol{C}_{\boldsymbol{a} J} \right]^{-1} \boldsymbol{\sigma}_{J \setminus T} \right| && \text{(H.38)} \\
\leq \;\; & \left( \left\| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}_0 I \setminus J} \right\|_1 \; + \; \left\| s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a} J} \right\|_1 \left\| \left[ \boldsymbol{C}_{\boldsymbol{a} J}^* \boldsymbol{C}_{\boldsymbol{a} J} \right]^{-1} \right\|_{\infty \to \infty} \left\| \boldsymbol{C}_{\boldsymbol{a} J}^* \boldsymbol{C}_{\boldsymbol{a}_0 I \setminus J} \right\|_{\infty \to \infty} \right) \left\| (\boldsymbol{x}_0)_{I \setminus J} \right\|_\infty
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \left(\|s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}\, J}\|_2 \,+\, \|s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}\, J}\|_2 \left\| [\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, J}]^{-1} \right\|_{\square\to\square} \|\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}\, J}\|_{\square\to\square}\right) \sqrt{2}\, \|\boldsymbol{x}_0\|_\square \\
&\quad + \lambda \|s_i[\boldsymbol{a}]^* \boldsymbol{C}_{\boldsymbol{a}\, J}\|_1 \left\| [\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, J}]^{-1} \right\|_{\infty\to\infty} \left\|\boldsymbol{\sigma}_{J\setminus T}\right\|_\infty && \text{(H.39)} \\
\leq\; & C_1 \left(\widetilde{\mu}\kappa_I \,+\, \widetilde{\mu}\kappa_I \times 1 \times \widetilde{\mu}\kappa_I\right) \times 2\lambda \\
&\quad + \left(2\sqrt{\kappa_I}\|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \,+\, C_2 \sqrt{\kappa_I}\widetilde{\mu} \times 1 \times \kappa_I \|\boldsymbol{a} - \boldsymbol{a}_0\|_2\right) \times \sqrt{\kappa_I} \\
&\quad + \lambda C_3 \times \widetilde{\mu}\kappa_I && \text{(H.40)} \\
\leq\; & \left((C_1 + C_3)\, \widetilde{\mu}\kappa_I \,+\, C_1 (\widetilde{\mu}\kappa_I)^2\right)\lambda \;+\; \left(2 \,+\, C_2 \widetilde{\mu}\kappa_I\right)\kappa_I \,\|\boldsymbol{a} - \boldsymbol{a}_0\|_2 && \text{(H.41)} \\
<\; & \lambda, && \text{(H.42)}
\end{aligned}
$$

where the last line holds as long as $c_\mu$ is a sufficiently small numerical constant. This establishes that $\boldsymbol{x}^+$ is the unique optimal solution to the full Lasso problem.

3. (Entrywise difference to $\boldsymbol{x}_0$) Finally we will be controlling $\left\|\boldsymbol{x}_J^+ - (\boldsymbol{x}_0)_J\right\|_\infty$. Indeed, from Lemma H.8,

$$
\begin{aligned}
\left\|\boldsymbol{x}_J^+ - (\boldsymbol{x}_0)_J\right\|_\infty \;=\;& \left\| [\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, J}]^{-1} \boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 - \lambda\, [\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, J}]^{-1} \boldsymbol{\sigma}_{J\setminus T} - (\boldsymbol{x}_0)_J \right\|_\infty \\
\leq\;& \left\| [\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, J}]^{-1} \boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}\, J} (\boldsymbol{x}_0)_J \right\|_\infty + \lambda \left\| [\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, J}]^{-1} \boldsymbol{\sigma}_{J\setminus T} \right\|_\infty \\
&\quad + \left\| [\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, J}]^{-1} \boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, I\setminus J} (\boldsymbol{x}_0)_{I\setminus J} \right\|_\infty \\
\leq\;& 2 \left\|\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}\, J}\right\|_{\square\to\infty} \|(\boldsymbol{x}_0)_J\|_\square \,+\, 2\lambda \,+\, 2 \left\|\boldsymbol{C}_{\boldsymbol{a}\, J}^* \boldsymbol{C}_{\boldsymbol{a}\, I\setminus J}\right\|_{\infty\to\infty} \left\|(\boldsymbol{x}_0)_{I\setminus J}\right\|_\infty \\
\leq\;& 2\sqrt{2\kappa_I}\, \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \, \|\boldsymbol{x}_0\|_\square \,+\, 2\lambda \,+\, 2 \times 3\widetilde{\mu} \times 2\kappa_{I\setminus J} \times 3\lambda \\
\leq\;& 3\kappa_I \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \,+\, 2\lambda \,+\, 36\lambda\widetilde{\mu}\kappa_I \\
\leq\;& 3\lambda, && \text{(H.43)}
\end{aligned}
$$

establishing the claim. ■

## H.2 Least squares solution $\boldsymbol{a}^{(k)}$ contracts

**Approximation of least squares solution.** In this section, given $\boldsymbol{x}$ to be the solution to the reweighted Lasso from $\boldsymbol{a}$, we will show the solution of the least squares problem

$$
\boldsymbol{a}^+ \;\leftarrow\; \operatorname*{argmin}_{\boldsymbol{a}' \in \mathbb{R}^p} \tfrac{1}{2} \left\|\boldsymbol{a}' * \boldsymbol{x} \,-\, \boldsymbol{y}\right\|_2^2 \tag{H.44}
$$

is closer to $\boldsymbol{a}_0$ compared to $\boldsymbol{a}$. Observe that in Lemma H.1, the solution of (H.16)

$$
\boldsymbol{x} \;=\; \boldsymbol{\iota}_J \left(\boldsymbol{C}_{\boldsymbol{a} J}^* \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{\iota}_J^* \left(\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 - \lambda \boldsymbol{P}_{J\setminus T} \boldsymbol{\sigma}\right), \tag{H.45}
$$

by assuming $\boldsymbol{C}_{\boldsymbol{a} J}^* \boldsymbol{C}_{\boldsymbol{a} J} \approx \boldsymbol{I}$, $\boldsymbol{a} \approx \boldsymbol{a}_0$ and $J \setminus T \approx \emptyset$, is a good approximation to the true sparse map $\boldsymbol{x}_0$

$$
\boldsymbol{x} \;\approx\; \boldsymbol{I}\,(\boldsymbol{x}_0 - \boldsymbol{0}) \;=\; \boldsymbol{x}_0\,; \tag{H.46}
$$

furthermore, its difference to the true sparse map $\|\boldsymbol{x}_0 - \boldsymbol{x}\|_2$ is proportional to $\|\boldsymbol{a}_0 - \boldsymbol{a}\|_2$ as

$$
\boldsymbol{x} - \boldsymbol{x}_0 \;\approx\; \boldsymbol{P}_I \left(\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 - \boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{x}_0\right) \;\approx\; \boldsymbol{P}_I \left[\boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}(\boldsymbol{a}_0 - \boldsymbol{a})\right]. \tag{H.47}
$$

To this end, since we know the solution of least square problem $\boldsymbol{a}^+$ is simply

$$
\boldsymbol{a}^+ \;=\; \left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{\iota}\right)^{-1} \left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0\right), \tag{H.48}
$$

this implies the difference between the new $\boldsymbol{a}^+$ and $\boldsymbol{a}_0$, has the relationship with $\boldsymbol{a} - \boldsymbol{a}_0$ roughly

$$
\begin{aligned}
\boldsymbol{a}^+ - \boldsymbol{a}_0 \;=\;& \left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{\iota}\right)^{-1} \left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \,-\, \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{\iota} \boldsymbol{a}_0\right) \;\approx\; (n\theta)^{-1}\, \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} (\boldsymbol{x}_0 - \boldsymbol{x}) \\
\approx\;& (n\theta)^{-1}\, \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} (\boldsymbol{a} - \boldsymbol{a}_0). && \text{(H.49)}
\end{aligned}
$$

To make this point precise, we introduce the following lemma:

**Lemma H.2** (Approximation of least square estimate). *Given $\boldsymbol{a}_0 \in \mathbb{R}^{p_0}$ to be $\widetilde{\mu}$-shift coherent and $\boldsymbol{x}_0 \sim \mathrm{BG}(\theta) \in \mathbb{R}^n$. There exists some constants $C, C', c, c', c_\mu$ such that if $\lambda < c'\widetilde{\mu}\kappa_I$, $\widetilde{\mu}\kappa_I^2 \leq c_\mu$ and $n > Cp^2 \log p$, then with probability at least $1 - c/n$, for every $\boldsymbol{a}$ satisfying $\|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \leq \widetilde{\mu}$ and $\boldsymbol{x}$ of the form*

$$\boldsymbol{x} \;=\; \boldsymbol{\iota}_J \left(\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J}\right)^{-1} \boldsymbol{\iota}_J^* \left(\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{y} - \lambda \boldsymbol{P}_{J\setminus T} \boldsymbol{\sigma}\right) \tag{H.50}$$

*where the set $J, T$ satisfies $I_{>6\lambda} \subseteq T \subseteq J \subseteq I$, we have*

$$\frac{1}{n\theta} \left\| \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \;-\; \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}(\boldsymbol{a}_0 - \boldsymbol{a}) \right\|_2 \;\leq\; C'\lambda \left(\widetilde{\lambda} + \widetilde{\mu}\kappa_I\right) + \frac{1}{32} \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \tag{H.51}$$

*with $\widetilde{\lambda} = \lambda + \frac{\log n}{\sqrt{n\theta^2}}$.*

**Proof.** We will begin with listing the conditions we use for both $\boldsymbol{x}$ and $\boldsymbol{x}_0$. First, we know from Lemma H.1 and our assumptions on the set $T$, then $\boldsymbol{x}$ approximates $\boldsymbol{x}_0$ in the sense that

$$\|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty \;\leq\; 3\lambda \tag{H.52}$$

$$\left\|(\boldsymbol{x}_0)_{I\setminus J}\right\|_\infty \;\leq\; 3\lambda \tag{H.53}$$

$$\left\|(\boldsymbol{x}_0)_{I\setminus T}\right\|_\infty \;\leq\; 6\lambda. \tag{H.54}$$

Write $\boldsymbol{x}_0 = \boldsymbol{g} \circ \boldsymbol{\omega}$ with $\boldsymbol{g}$ iid standard normal, $\boldsymbol{\omega}$ iid Bernoulli and $\boldsymbol{g}$ and $\boldsymbol{\omega}$ independent. From (H.53) we know $|I \setminus J| = |\{\, i \mid |\boldsymbol{g}_i| \leq 3\lambda,\ \boldsymbol{\omega}_i \neq 0 \,\}|$. Since $\mathbb{P}\left[\boldsymbol{\omega}_i \neq 0\right] = \theta$ and $\mathbb{P}\left[|\boldsymbol{g}_i| \leq 3\lambda\right] \leq 3\lambda$, Lemma A.1 implies that with probability at least $1 - 2/n$:

$$|I \setminus J| \;\leq\; 3\lambda n\theta + 6\sqrt{\lambda n\theta} \log n \;\leq\; 3\widetilde{\lambda} n\theta \tag{H.55}$$

$$|I \setminus T| \;\leq\; 6\lambda n\theta + 12\sqrt{\lambda n\theta} \log n \;\leq\; 6\widetilde{\lambda} n\theta, \tag{H.56}$$

and

$$|(I \setminus J) \cap s_\ell[I]| \leq 3\lambda n\theta^2 + 6\sqrt{\lambda n\theta^2} \log n \leq 3\widetilde{\lambda} n\theta^2; \tag{H.57}$$

together with base on properties of Bernoulli-Gaussian vector $\boldsymbol{x}_0$ from Appendix A and we conclude with probability at least $1 - c/n$, all the following events hold:

$$\tfrac{1}{2} n\theta \;\leq\; |I| \;\leq\; 2n\theta, \tag{H.58}$$

$$\max_{\ell \neq 0} |I \cap s_\ell[I]| \;\leq\; 2n\theta^2 \tag{H.59}$$

$$\max_{\ell \neq 0} |(I \setminus J) \cap s_\ell[I]| \;\leq\; 6\widetilde{\lambda} n\theta^2, \tag{H.60}$$

$$\|\boldsymbol{x}_0\|_\square^2 \;\leq\; \kappa_I, \tag{H.61}$$

$$\|\check{\boldsymbol{a}}_0 * \boldsymbol{x}_0\|_\square^2 \;\leq\; \kappa_I, \tag{H.62}$$

$$\|\boldsymbol{x}_0\|_2^2 \;\leq\; 2n\theta, \tag{H.63}$$

$$\|\boldsymbol{x}_0\|_1 \;\leq\; 2n\theta, \tag{H.64}$$

$$\max_{\ell \neq 0} \|\boldsymbol{P}_{I \cap s_\ell[I]} \boldsymbol{x}_0\|_2^2 \;\leq\; 2n\theta^2, \tag{H.65}$$

$$\max_{\ell \neq 0} \left\|\boldsymbol{P}_{I \cap s_\ell[I \setminus J]} \boldsymbol{x}_0\right\|_1 \;\leq\; 12\widetilde{\lambda} n\theta^2, \tag{H.66}$$

$$\|\boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}\|_2^2 \;\leq\; 3n\theta, \tag{H.67}$$

provided by $n \geq C\theta^{-2} \log p$ for sufficiently large constant $C$.

1. (Approximate $\boldsymbol{C}_{\boldsymbol{x}}$ with $\boldsymbol{C}_{\boldsymbol{x}_0}$) Since

$$\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \;=\; \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \;+\; \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \tag{H.68}$$

where

$$
\begin{aligned}
\left\| \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \right\|_2 \;&\le\; \left\|\boldsymbol{a}_0\right\|_2 \left\|\boldsymbol{x}-\boldsymbol{x}_0\right\|_2^2 \;+\; \left\|\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\iota}\right\|_2 \sqrt{2p} \max_{\ell\neq 0} \left|\left\langle s_\ell[\boldsymbol{x}-\boldsymbol{x}_0], \boldsymbol{x}-\boldsymbol{x}_0 \right\rangle\right| \\
&\le\; \left\|\boldsymbol{x}-\boldsymbol{x}_0\right\|_\infty^2 \times |I| \;+\; \sqrt{2\widetilde{\mu} p^2}\left( \left\|\boldsymbol{x}-\boldsymbol{x}_0\right\|_\infty^2 \times \max_{\ell\neq 0} \left|I \cap s_\ell[I]\right| \right) \\
&\le\; C_1\left(\lambda^2 n\theta \;+\; \sqrt{2\widetilde{\mu} p^2}\left(\lambda^2 n\theta^2\right)\right) \\
&\le\; 2C_1\lambda^2 n\theta,
\end{aligned}
\tag{H.69}
$$

we have that

$$
\left\| \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \;-\; \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \right\|_2 \;\le\; 2C_1\lambda^2 n\theta. \tag{H.70}
$$

2. (<u>Extract the $\boldsymbol{a}_0 - \boldsymbol{a}$ term</u>) Observe that

$$
\begin{aligned}
&\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0} \boldsymbol{\iota} \boldsymbol{a}_0 \\
&= \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} (\boldsymbol{x}-\boldsymbol{x}_0) \\
&= \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \left( \boldsymbol{\iota}_J \left(\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J}\right)^{-1} \boldsymbol{\iota}_J^* \left(\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{x}_0 - \lambda \boldsymbol{P}_{J\setminus T}\boldsymbol{\sigma}\right) - \boldsymbol{\iota}_J \left(\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J}\right)^{-1} \left(\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J}\right) (\boldsymbol{x}_0)_J - \boldsymbol{P}_{I\setminus J}\boldsymbol{x}_0 \right) \\
&= \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0 J} (\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J})^{-1} \boldsymbol{C}_{\boldsymbol{a}J}^* \left(\boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}} \boldsymbol{x}_0\right) \\
&\qquad +\; \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0 J} (\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J})^{-1} \boldsymbol{C}_{\boldsymbol{a}J}^* \left(\boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{x}_0 - \boldsymbol{C}_{\boldsymbol{a}J} (\boldsymbol{x}_0)_J\right) \\
&\qquad -\; \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_{I\setminus J} \boldsymbol{x}_0 \\
&\qquad -\; \lambda\, \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0 J} (\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J})^{-1} \boldsymbol{\iota}_J^* \boldsymbol{P}_{J\setminus T} \boldsymbol{\sigma},
\end{aligned}
\tag{H.71}
$$

where, the second term in (H.71) is bounded as

$$
\begin{aligned}
&\left\| \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0 J} (\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J})^{-1} \boldsymbol{C}_{\boldsymbol{a}J}^* \left(\boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{x}_0 - \boldsymbol{C}_{\boldsymbol{a}J} (\boldsymbol{x}_0)_J\right) \right\|_2 \\
&\qquad \le\; \left\|\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right\|_2 \;\times\; \left\|\boldsymbol{C}_{\boldsymbol{a}_0 J}\right\|_2 \left\|(\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J})^{-1}\right\|_2 \;\times\; \left\|\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}I\setminus J}\right\|_2 \;\times\; \left\|(\boldsymbol{x}_0)_{I\setminus J}\right\|_2 \\
&\qquad \le\; C_2\left(\sqrt{n\theta} \;\times\; 3 \;\times\; \widetilde{\mu}\kappa_I \;\times\; \lambda\sqrt{\widetilde{\lambda} n\theta}\right) \\
&\qquad \le\; 3C_2 \widetilde{\mu}\kappa_I \lambda n\theta;
\end{aligned}
\tag{H.72}
$$

the third term in (H.71) is bounded as

$$
\begin{aligned}
\left\| \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_{I\setminus J} \boldsymbol{x}_0 \right\|_2 \;&=\; \left\| \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \left(\boldsymbol{P}_{[\pm p]\setminus 0} + \boldsymbol{e}_0\boldsymbol{e}_0^*\right) \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_{I\setminus J} \boldsymbol{x}_0 \right\|_2 \\
&\le\; \left\|\boldsymbol{a}_0\right\|_2 \left\|(\boldsymbol{x}_0)_{I\setminus J}\right\|_2^2 \;+\; \left\|\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{\iota}\right\|_2 \times \sqrt{2p} \times \max_{\ell\neq 0} \left\|\boldsymbol{P}_{I\cap s_\ell[I\setminus J]}\boldsymbol{x}_0\right\|_1 \times \left\|(\boldsymbol{x}_0)_{I\setminus J}\right\|_\infty \\
&\le\; C_3\left(\lambda^2 \times \widetilde{\lambda} n\theta \;+\; \sqrt{\widetilde{\mu} p^2} \times \widetilde{\lambda} n\theta^2 \times \lambda\right) \\
&\le\; 2C_3 \widetilde{\lambda}\lambda n\theta;
\end{aligned}
\tag{H.73}
$$

and finally, write $\boldsymbol{\Delta} = (\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J})^{-1} - \boldsymbol{I}$, then the forth term in (H.71) is bounded as

$$
\begin{aligned}
&\lambda \left\| \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}_J (\boldsymbol{C}_{\boldsymbol{a}J}^* \boldsymbol{C}_{\boldsymbol{a}J})^{-1} \boldsymbol{\iota}_J^* \boldsymbol{P}_{J\setminus T} \boldsymbol{\sigma} \right\|_2 \\
&\quad =\; \lambda \left\| \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0} \left(\boldsymbol{P}_{[\pm p]\setminus 0} + \boldsymbol{e}_0\boldsymbol{e}_0^*\right) \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{\iota}_J \left(\boldsymbol{I} + \boldsymbol{\Delta}\right) \boldsymbol{\iota}_J^* \boldsymbol{P}_{J\setminus T} \boldsymbol{\sigma} \right\|_2 \\
&\quad \le\; \lambda \left\|\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{\iota}\right\|_2 \sqrt{2p} \max_{\ell\neq 0} \left\|\boldsymbol{P}_{I\cap s_\ell[I\setminus T]}\boldsymbol{x}_0\right\|_1 \;+\; \lambda \left\|\boldsymbol{a}_0\right\|_2 \left\|\boldsymbol{P}_{I\setminus T}\boldsymbol{x}_0\right\|_1 \\
&\qquad +\; \lambda \left\|\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{\iota}\right\|_2 \sqrt{2p} \left\|\boldsymbol{P}_{I\cap s_\ell[I]}\boldsymbol{x}_0\right\|_1 \left\|\boldsymbol{\Delta}\right\|_{\infty\to\infty} \;+\; \lambda \left\|\boldsymbol{a}_0\right\|_2 \left\|\boldsymbol{x}_0\right\|_2 \left\|\boldsymbol{\Delta}\right\|_2 \sqrt{|J\setminus T|} \\
&\quad \le\; C_4\lambda\left(\sqrt{\widetilde{\mu} p^2} \times \widetilde{\lambda} n\theta^2 \;+\; \lambda\widetilde{\lambda} n\theta \;+\; \sqrt{\widetilde{\mu} p^2} \times n\theta^2 \times \widetilde{\mu}\kappa_I \;+\; \sqrt{n\theta} \times \widetilde{\mu}\kappa_I \sqrt{\widetilde{\lambda} n\theta}\right)
\end{aligned}
$$

$$\leq 2C_4\left(\widetilde{\lambda}+\widetilde{\mu}\kappa_I\right)\lambda n\theta. \tag{H.74}$$

Therefore, combining (H.72)-(H.74) we obtain

$$\left\|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_0}\boldsymbol{\iota}\boldsymbol{a}_0 \;-\; \boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0 J}(\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}J})^{-1}\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}_0-\boldsymbol{a}}\boldsymbol{x}_0\right\|_2 \;\leq\; C_5\left(\widetilde{\lambda}+\widetilde{\mu}\kappa_I\right)\lambda n\theta. \tag{H.75}$$

3. (Extract the set $J$) Lastly, we will further simplify the term with $\boldsymbol{a}-\boldsymbol{a}_0$ in (H.75) by extracting the set $J$:

$$\begin{aligned}
&\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0 J}(\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}J})^{-1}\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}_0-\boldsymbol{a}}\boldsymbol{x}_0\\
&\quad=\;\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0 J}\left(\boldsymbol{I}+\boldsymbol{\Delta}\right)\boldsymbol{C}_{\boldsymbol{a}_0+(\boldsymbol{a}-\boldsymbol{a}_0)J}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\left(\boldsymbol{a}_0-\boldsymbol{a}\right)\\
&\quad=\;\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{P}_I\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}(\boldsymbol{a}_0-\boldsymbol{a})\\
&\qquad+\;\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0 J}\boldsymbol{\Delta}\boldsymbol{C}_{\boldsymbol{a}_0 J}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}(\boldsymbol{a}_0-\boldsymbol{a})\;+\;\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0 J}\left(\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}J}\right)^{-1}\boldsymbol{C}_{\boldsymbol{a}-\boldsymbol{a}_0 J}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}(\boldsymbol{a}_0-\boldsymbol{a})\\
&\qquad-\;\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{P}_{I\setminus J}\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}(\boldsymbol{a}_0-\boldsymbol{a}),
\end{aligned} \tag{H.76}$$

where, the latter terms in (H.76) are bounded as

$$\begin{aligned}
&\left\|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0 J}\boldsymbol{\Delta}\boldsymbol{C}_{\boldsymbol{a}_0 J}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right\|_2\;\leq\;\left\|\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right\|_2^2\left\|\boldsymbol{C}_{\boldsymbol{a}_0 J}\right\|_2^2\left\|\boldsymbol{\Delta}\right\|_2\;\leq\;C_6\widetilde{\mu}\kappa_I n\theta\\
&\left\|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0 J}\left(\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}J}\right)^{-1}\boldsymbol{C}_{\boldsymbol{a}-\boldsymbol{a}_0 J}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right\|_2\;\leq\;\left\|\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right\|_2^2\left\|\boldsymbol{C}_{\boldsymbol{a}_0 J}\right\|_2\left\|(\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}J})^{-1}\right\|_2\left\|\boldsymbol{C}_{\boldsymbol{a}_0-\boldsymbol{a}}\boldsymbol{\iota}_J\right\|_2\;\leq\;C_7\widetilde{\mu}\sqrt{\kappa_I}n\theta\\
&\left\|\boldsymbol{P}_{I\setminus J}\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right\|_2^2\;\leq\;|I\setminus J|\left\|\check{\boldsymbol{a}}_0*\boldsymbol{x}_0\right\|_\square^2\;\leq\;C_8\widetilde{\lambda}n\theta\times\kappa_I\;\leq\;C_8\left(\lambda\kappa_I+\tfrac{\kappa_I\log n}{\sqrt{n\theta^2}}\right)n\theta,
\end{aligned} \tag{H.77}$$

whence we conclude, that since $c_\mu\kappa_I^2\leq c_\mu$ and $\lambda\kappa_I\leq 5c_\mu$, as long as $c_\mu<\frac{1}{100}\left(\frac{1}{C_6}+\frac{1}{C_7}+\frac{1}{5C_8}\right)$ and $n>10^6C_8^2\theta^{-2}\kappa_I^2\log^2 n$, we gain:

$$\begin{aligned}
&\left\|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0 J}(\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}J})^{-1}\boldsymbol{C}_{\boldsymbol{a}J}^*\boldsymbol{C}_{\boldsymbol{a}_0-\boldsymbol{a}}\boldsymbol{x}_0\;-\;\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{P}_I\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}(\boldsymbol{a}_0-\boldsymbol{a})\right\|_2\\
&\quad\leq\;\left(\tfrac{3}{100}+\tfrac{1}{1000}\right)n\theta\left\|\boldsymbol{a}_0-\boldsymbol{a}\right\|_2\\
&\quad\leq\;\tfrac{1}{32}n\theta\left\|\boldsymbol{a}_0-\boldsymbol{a}\right\|_2.
\end{aligned} \tag{H.78}$$

The claimed result therefore is followed by combining (H.70), (H.75) and (H.78). ■

**Contraction of least square estimate of $\boldsymbol{a}$ toward $\boldsymbol{a}_0$.** The next thing is to show the operator

$$(n\theta)^{-1}\left(\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{P}_I\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right) \tag{H.79}$$

contracts $\boldsymbol{a}$ toward $\boldsymbol{a}_0$. We first will show that

$$(n\theta)^{-1}\left(\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{P}_I\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right)\;\approx\;\boldsymbol{a}_0\boldsymbol{a}_0^* \tag{H.80}$$

by seeing $\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{P}_I\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\;\approx\;(n\theta)\,\boldsymbol{e}_0\boldsymbol{e}_0^*$ via sparsity of $\boldsymbol{x}_0$. Finally since the local perturbation on sphere is close to a quadratic function in $\ell^2$-norm of difference, we have

$$\left|\langle\boldsymbol{a}_0,\boldsymbol{a}-\boldsymbol{a}_0\rangle\right|\;\leq\;\tfrac{1}{2}\left\|\boldsymbol{a}-\boldsymbol{a}_0\right\|_2^2. \tag{H.81}$$

Again, we introduce the following lemma to solidify our claim:

**Lemma H.3** (Contraction of $\boldsymbol{a}$ to $\boldsymbol{a}_0$). *Given $\boldsymbol{a}_0\in\mathbb{R}^{p_0}$ to be $\widetilde{\mu}$-shift coherent and $\boldsymbol{x}_0\sim\mathrm{BG}(\theta)\in\mathbb{R}^n$. There exists some constants $C,C',c,c',c_\mu$ such that if $\lambda<c'\widetilde{\mu}\kappa_I$, $\widetilde{\mu}\kappa_I^2\leq c_\mu$ and $n>C\theta^{-2}p^2\log p$, then with probability at least $1-c/n$, for every $\|\boldsymbol{a}-\boldsymbol{a}_0\|_2\leq\widetilde{\mu}$,*

$$\left\|\,\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}_0}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{P}_I\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}(\boldsymbol{a}_0-\boldsymbol{a})\,\right\|_2\;\leq\;\frac{1}{32}\left\|\boldsymbol{a}-\boldsymbol{a}_0\right\|_2 n\theta. \tag{H.82}$$

**Proof.** Since $\mathbb{E}\left\langle \boldsymbol{P}_I s_i[\boldsymbol{x}_0], s_j[\boldsymbol{x}_0]\right\rangle = 0$ for all $i \neq j$ and set $I$, we calculate

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\iota}_{[\pm p]}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}_{[\pm p]}\right] \;&=\; \sum_{i\in[\pm p]} \mathbb{E}\left[\boldsymbol{e}_i^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{e}_i\right] \boldsymbol{e}_i \boldsymbol{e}_i^* \;=\; \mathbb{E}\left\|\boldsymbol{x}_0\right\|_2^2 \boldsymbol{e}_0 \boldsymbol{e}_0^* \;+\; \sum_{i\in[\pm p]\backslash 0} \mathbb{E}\left\|\boldsymbol{P}_I s_i[\boldsymbol{x}_0]\right\|_2^2 \boldsymbol{e}_i \boldsymbol{e}_i^* \\ &=\; n\theta \boldsymbol{e}_0 \boldsymbol{e}_0^* \;+\; n\theta^2 \boldsymbol{P}_{[\pm p]\backslash 0} \;=\; n\theta^2 \boldsymbol{I} \;+\; n\theta(1-\theta)\, \boldsymbol{e}_0 \boldsymbol{e}_0^*. \end{aligned} \tag{H.83}$$

whence

$$\mathbb{E}\left[\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}\right] \;=\; \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}^* \mathbb{E}\left[\boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0}\right] \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota} \;=\; n\theta^2 \boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota} \;+\; n\theta(1-\theta) \boldsymbol{a}_0 \boldsymbol{a}_0^*, \tag{H.84}$$

implying the expectation is a contraction mapping for $\boldsymbol{a}_0 - \boldsymbol{a}$ when $c_\mu < \frac{1}{200}$:

$$\begin{aligned}\left\|\,\mathbb{E}\left[\,\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}\right](\boldsymbol{a}_0 - \boldsymbol{a})\,\right\|_2 \;&\leq\; n\theta^2 \left\|\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}\right\|_2 \left\|\boldsymbol{a}_0 - \boldsymbol{a}\right\|_2 \;+\; n\theta \left\|\boldsymbol{a}_0\right\|_2 \left|\left\langle \boldsymbol{a}_0, \boldsymbol{a}_0 - \boldsymbol{a}\right\rangle\right| \\ &\leq\; n\theta^2 \times 2\widetilde{\mu} p \times \left\|\boldsymbol{a}_0 - \boldsymbol{a}\right\|_2 + \tfrac{1}{2} n\theta \left\|\boldsymbol{a}_0 - \boldsymbol{a}\right\|_2^2 \\ &\leq\; \left(2c_\mu + \tfrac{1}{2} c_\mu\right) \left\|\boldsymbol{a}_0 - \boldsymbol{a}\right\|_2 n\theta \\ &\leq\; \tfrac{1}{64} \left\|\boldsymbol{a}_0 - \boldsymbol{a}\right\|_2 n\theta. \end{aligned} \tag{H.85}$$

For each entry of $\boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0}$, again from Appendix A we know with probability at least $1 - c/n$:

$$\left|\boldsymbol{e}_i^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{e}_j - \mathbb{E}\left[\boldsymbol{e}_i^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{e}_j\right]\right| \;\leq\; \begin{cases} C'\sqrt{n\theta \log n} & i = j = 0 \\ C'\sqrt{n\theta^2 \log n} & \text{otherwise} \end{cases}.$$

Thus via Gershgorin disc theorem, when $n > 10^3 C'^2 \theta^{-2} p^2 \log n$:

$$\lambda_{\max}\left(\boldsymbol{\iota}_{[\pm p]}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}_{[\pm p]} - \mathbb{E}\left[\boldsymbol{\iota}_{[\pm p]}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}_{[\pm p]}\right]\right) \;\leq\; C' p \sqrt{n\theta^2 \log n} \;\leq\; \tfrac{1}{64} n\theta^2. \tag{H.86}$$

Finally we combine (H.85), (H.86) and get

$$\left\|\,\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}_0}^* \boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{P}_I \boldsymbol{C}_{\boldsymbol{a}_0}^* \boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}(\boldsymbol{a}_0 - \boldsymbol{a})\,\right\|_2 \;\leq\; \left(\tfrac{1}{64} n\theta + \tfrac{1}{64} n\theta^2 \left\|\boldsymbol{C}_{\boldsymbol{a}_0} \boldsymbol{\iota}_{\pm p}\right\|_2^2\right) \left\|\boldsymbol{a}_0 - \boldsymbol{a}\right\|_2 \;\leq\; \tfrac{1}{32} \left\|\boldsymbol{a}_0 - \boldsymbol{a}\right\|_2 n\theta. \tag{H.87}$$

■

Lemma H.1-H.3 together implies the single iterate of alternating minimization contracts $\boldsymbol{a}$ toward $\boldsymbol{a}_0$. We show it with the following lemma:

**Lemma H.4** (Contraction of least square estimate)**.** *Given $\boldsymbol{a}_0 \in \mathbb{R}^{p_0}$ to be $\widetilde{\mu}$-shift coherent and $\boldsymbol{x}_0 \sim \mathrm{BG}(\theta) \in \mathbb{R}^n$. There exists some constants $C, C', c, c_\mu$ such that if $\widetilde{\mu}\kappa_I^2 \leq c_\mu$ and $n > C\theta^{-2} p^2 \log n$, then with probability at least $1 - c/n$, for every $\lambda$ and $\boldsymbol{a}$ satisfying*

$$5\widetilde{\mu}\kappa_I \;\geq\; \lambda \;\geq\; 5\kappa_I \left\|\boldsymbol{a} - \boldsymbol{a}_0\right\|_2, \tag{H.88}$$

*and suppose $\boldsymbol{x}^+$ has the form of (H.16), then the solution $\boldsymbol{a}^+$ to*

$$\min_{\boldsymbol{a}' \in \mathbb{R}^p} \left\{\left\|\boldsymbol{a}' * \boldsymbol{x}^+ - \boldsymbol{y}\right\|_2^2\right\} \tag{H.89}$$

*is unique and satisfies*

$$\left\|\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a}^+\right] - \boldsymbol{a}_0\right\|_2 \;\leq\; \frac{1}{2} \left\|\boldsymbol{a} - \boldsymbol{a}_0\right\|_2. \tag{H.90}$$

**Proof.** Write $\boldsymbol{x}$ as $\boldsymbol{x}^+$, then

$$\begin{aligned}\lambda_p\left(\boldsymbol{\iota}^* \boldsymbol{C}_{\boldsymbol{x}}^* \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{\iota}\right) \;&=\; \sigma_{\min}^2\left(\boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota} + \boldsymbol{C}_{\boldsymbol{x} - \boldsymbol{x}_0} \boldsymbol{\iota}\right) \\ &\geq\; \left[\sigma_{\min}(\boldsymbol{C}_{\boldsymbol{x}_0} \boldsymbol{\iota}) - \left\|\boldsymbol{C}_{\boldsymbol{x} - \boldsymbol{x}_0} \boldsymbol{\iota}\right\|\right]_+^2 \end{aligned}$$

$$
\begin{aligned}
&\geq \left[\sigma_{\min}(\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}) - 2\sqrt{\kappa_I}\,\|\boldsymbol{x}-\boldsymbol{x}_0\|_2\right]_+^2 \\
&\geq \left[\tfrac{2}{3}\sqrt{\theta n} - 8\lambda\sqrt{\kappa_I}\sqrt{\theta n}\right]_+^2 \\
&\geq \tfrac{1}{2}\theta n, \qquad \text{(H.91)}
\end{aligned}
$$

where the fourth inequality is derived from using the upper bound of sparse convolution matrix from Remark A.6, and the last line holds by knowing $\lambda < 5c_\mu\kappa_I^{-1}$. From (H.91) we know the least square problem of (H.89) has unique solution $\boldsymbol{a}^+$, written as

$$
\boldsymbol{a}^+ \;=\; \left(\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{\iota}\right)^{-1}\boldsymbol{\iota}\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{y}, \qquad \text{(H.92)}
$$

whence

$$
\boldsymbol{a}^+ - \boldsymbol{a}_0 \;=\; \left(\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{\iota}\right)^{-1}\left(\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{C}_{\boldsymbol{x}_0}\boldsymbol{\iota}\right)\boldsymbol{a}_0 - \boldsymbol{a}_0 \;=\; \left(\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{\iota}\right)^{-1}\left(\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{C}_{\boldsymbol{x}_0-\boldsymbol{x}}\boldsymbol{\iota}\right)\boldsymbol{a}_0. \qquad \text{(H.93)}
$$

Combine Lemma H.2 and Lemma H.3, we know

$$
\|\,\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{C}_{\boldsymbol{x}_0-\boldsymbol{x}}\boldsymbol{\iota}\,\|_2 \;\leq\; \left(C_1\lambda\left(\widetilde{\lambda}+\widetilde{\mu}\kappa_I\right) + \tfrac{1}{16}\,\|\boldsymbol{a}-\boldsymbol{a}_0\|_2\right)n\theta \qquad \text{(H.94)}
$$

for some constant $C_1$. Combine (H.91), (H.93), (H.94) and since $\lambda < \widetilde{\mu}\kappa_I$, by letting $c_\mu < \frac{1}{4C_1}$, we gain

$$
\left\|\boldsymbol{a}^+ - \boldsymbol{a}_0\right\|_2 \;\leq\; \frac{\|\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{C}_{\boldsymbol{x}_0-\boldsymbol{x}}\boldsymbol{\iota}\|_2}{\lambda_p(\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{x}}^*\boldsymbol{C}_{\boldsymbol{x}}\boldsymbol{\iota})} \;\leq\; 2C_1\lambda\left(\widetilde{\lambda}+\widetilde{\mu}\kappa_I\right) + \frac{1}{8}\,\|\boldsymbol{a}-\boldsymbol{a}_0\|_2 \;\leq\; \frac{1}{4}. \qquad \text{(H.95)}
$$

For the final bound,

$$
\begin{aligned}
\left\|\frac{\boldsymbol{a}^+}{\|\boldsymbol{a}^+\|_2} - \boldsymbol{a}_0\right\|_2 &\leq \frac{\|\boldsymbol{a}^+-\boldsymbol{a}_0\|_2 + |\|\boldsymbol{a}^+\|_2 - 1|}{\|\boldsymbol{a}^+\|_2} \;\leq\; \frac{2\,\|\boldsymbol{a}^+-\boldsymbol{a}_0\|_2}{1-\|\boldsymbol{a}^+-\boldsymbol{a}_0\|_2} \;\leq\; \frac{8}{3}\left\|\boldsymbol{a}^+-\boldsymbol{a}_0\right\|_2, \\
&\leq C_2\lambda\left(\widetilde{\lambda}+\widetilde{\mu}\kappa_I\right) \;+\; \frac{1}{3}\,\|\boldsymbol{a}-\boldsymbol{a}_0\|_2, \qquad \text{(H.96)}
\end{aligned}
$$

and since $\lambda > \kappa_I\,\|\boldsymbol{a}-\boldsymbol{a}_0\|_2$, finally we gain

$$
\begin{aligned}
\text{(H.96)} &\leq C_2\left(\lambda\kappa_I + \frac{p\kappa_I\log n}{n\theta} + \widetilde{\mu}\kappa_I^2\right)\|\boldsymbol{a}-\boldsymbol{a}_0\|_2 \;+\; \frac{1}{3}\,\|\boldsymbol{a}-\boldsymbol{a}_0\|_2 \\
&\leq \frac{1}{2}\,\|\boldsymbol{a}-\boldsymbol{a}_0\|_2 \qquad \text{(H.97)}
\end{aligned}
$$

as long as $n > 20C_2\theta^{-1}p\kappa_I\log n$ and $c_\mu < \frac{1}{20C_2}$.

■

### H.3 Linear convergence of alternating minimization (Proof of Theorem 5.2 )

In the first two sections we have shown the iterate contract $\boldsymbol{a}$ toward $\boldsymbol{a}_0$, under our signal assumption. We tie up these result by showing the following theorem which proves that the iterates produced by alternating minimization converge linearly to $\boldsymbol{a}_0$:

**Proof.** We will prove our claim by induction on $k$. Clearly, when $k=0$, we have $5\,\kappa_I\left\|\boldsymbol{a}^{(0)}-\boldsymbol{a}_0\right\|_2 \leq \lambda^{(0)} = 5\widetilde{\mu}\kappa_I$ and $I^{(0)} = \left\{i : \left|s_i[\boldsymbol{a}^{(0)}]^*\boldsymbol{\iota}^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{x}_0\right| > \lambda^{(0)}\right\}$. Then for all $|\boldsymbol{x}_j| > 6\lambda^{(0)}$, we have

$$
\begin{aligned}
\left|s_j\left[\boldsymbol{a}^{(0)}\right]^*\boldsymbol{C}_{\boldsymbol{a}_0}\boldsymbol{x}_0\right| &\geq \left(1-\left|\langle\boldsymbol{a}^{(0)}\boldsymbol{a}_0\rangle\right|\right)|\boldsymbol{x}_j| \;-\; \left\|\boldsymbol{P}_{[\pm p]\setminus\{j\}}\boldsymbol{C}_{\boldsymbol{a}_0}^*\boldsymbol{\iota}s_j\left[\boldsymbol{a}^{(0)}\right]\right\|_2 \times \sqrt{2}\,\|\boldsymbol{x}_0\|_\square \\
&\geq (1-2\widetilde{\mu})\,6\lambda^{(0)} \;-\; 2\widetilde{\mu}\sqrt{\kappa_I} \times \sqrt{2\kappa_I}
\end{aligned}
$$

$$
\begin{aligned}
&\geq\ 5\lambda^{(0)}\ -\ 4\lambda^{(0)} \\
&=\ \lambda^{(0)}.
\end{aligned} \tag{H.98}
$$

hence $I_{>6\lambda^{(0)}} \subseteq I^{(0)}$, therefore the condition of Lemma H.4 is satisfied, implies (5.32) holds for $k = 0$.

Suppose it is true for $1, 2, \ldots, k-1$, such that

$$
\kappa_I \left\|\boldsymbol{a}^{(k)} - \boldsymbol{a}_0\right\|_2 \ \leq\ \tfrac{1}{2}\lambda^{(k-1)} \ =\ \lambda^{(k)}, \qquad \text{and} \qquad I_{>3\lambda^{(k-1)}} \subseteq I^{(k)} \tag{H.99}
$$

and since $I_{>6\lambda^{(k)}} = I_{>3\lambda^{(k-1)}} \subseteq I^{(k)}$, we can again apply Lemma H.4, resulting

$$
\kappa_I \left\|\boldsymbol{a}^{(k+1)} - \boldsymbol{a}\right\|_2 \ \leq\ \tfrac{1}{2}\kappa_I \left\|\boldsymbol{a}^{(k)} - \boldsymbol{a}_0\right\|_2 \ \leq\ \tfrac{1}{2}\lambda^{(k)} \tag{H.100}
$$

as claimed. ■

## H.4 Supporting lemmas for refinement

The following lemma controls the shift coherence of $\boldsymbol{a}$:

**Lemma H.5** (Coherence of $\boldsymbol{a}$ near $\boldsymbol{a}_0$ )**.** *Suppose that $\boldsymbol{a}_0$ is $\widetilde{\mu}$-shift coherent, and $\|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \leq \widetilde{\mu}$. Then*

$$
\left\|\mathrm{off}\,[\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}_0}]\right\|_\infty \leq 2\widetilde{\mu} \tag{H.101}
$$

$$
\left\|\mathrm{off}\,[\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}}]\right\|_\infty \leq 3\widetilde{\mu} \tag{H.102}
$$

**Proof.** Notice that for any $\ell \neq 0$, $|\langle \boldsymbol{a}, s_\ell[\boldsymbol{a}_0]\rangle| \leq |\langle \boldsymbol{a}_0, s_\ell[\boldsymbol{a}_0]\rangle| + |\langle \boldsymbol{a} - \boldsymbol{a}_0, s_\ell[\boldsymbol{a}_0]\rangle| \leq \widetilde{\mu} + \|\boldsymbol{a}_0 - \boldsymbol{a}\|_2 \leq 2\widetilde{\mu}$. Similarly, $|\langle \boldsymbol{a}, s_\ell[\boldsymbol{a}]\rangle| \leq |\langle \boldsymbol{a} - \boldsymbol{a}_0, s_\ell[\boldsymbol{a}_0]\rangle| + |\langle \boldsymbol{a}, s_\ell[\boldsymbol{a}_0]\rangle| \leq \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 + 2\widetilde{\mu} \leq 3\widetilde{\mu}$, as claimed. ■

From this we obtain the following spectral control on $\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}}$, to simply the notations, we will write

$$
\boldsymbol{C}_{\boldsymbol{a}I}^*\boldsymbol{C}_{\boldsymbol{a}I} = \boldsymbol{\iota}_I^*\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}}\boldsymbol{\iota}_I = [\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}}]_{I,I} \tag{H.103}
$$

in the latter part of this section.

**Lemma H.6** (Off-diagonals of $[\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}}]_{I,I}$ )**.** *Suppose that $\boldsymbol{a}_0$ is $\widetilde{\mu}$-shift coherent and $\|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \leq \widetilde{\mu}$. Then*

$$
\left\|[\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{I,I}\right\|_2 \leq 9\kappa_I\widetilde{\mu}. \tag{H.104}
$$

We prove this lemma by noting that $\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}} = \boldsymbol{C}_{\boldsymbol{r}_{\boldsymbol{a},\boldsymbol{a}}}$ is the convolution matrix associated with the autocorrelation $\boldsymbol{r}_{\boldsymbol{a},\boldsymbol{a}}$ of $\boldsymbol{a}$. Since $\mathrm{supp}(\boldsymbol{r}_{\boldsymbol{a},\boldsymbol{a}}) \subseteq \{-p+1, \ldots, p-1\}$ is confined to a (cyclic) stripe of width $2p-1$, we can tightly control the norm of this matrix by dividing it into three block-diagonal submatrices with blocks of size $p \times p$. Formally:

**Proof.** Divide $I$ into $r = \lceil n/p \rceil$ subsets $I_0, \ldots, I_{r-1}$ such that for all $\ell = 0, \ldots, r-1$:

$$
I_\ell = I \cap \{p\ell,\, p\ell+1,\, \ldots,\, p\ell + (p-1)\} = I \cap ([p] + p\ell).
$$

Notice that for each $\ell$:

$$
\mathrm{supp}\left([\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}}]_{I_\ell, I}\right) \subseteq I_\ell \times \Big(I_{\ell-1} \uplus I_\ell \uplus I_{\ell+1}\Big),
$$

where $\ell+1$ and $\ell-1$ are interpreted cyclically modulo $r$.

For an arbitrary $\boldsymbol{v} \in \mathbb{R}^{|I|}$, we calculate

$$
\left\|[\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{I,I}\,\boldsymbol{v}\right\|_2^2 = \sum_{\ell=0}^{r-1} \left\|[\boldsymbol{C}_{\boldsymbol{a}}^*\boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{I_\ell,I}\,\boldsymbol{v}\right\|_2^2 \tag{H.105}
$$

$$= \sum_{\ell=0}^{r-1} \left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{I_\ell, I_{\ell-1} \uplus I_\ell \uplus I_{\ell+1}} \boldsymbol{v}_{I_{\ell-1} \uplus I_\ell \uplus I_{\ell+1}} \right\|_2^2 \tag{H.106}$$

$$\leq \sum_{\ell=0}^{r-1} \left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{I_\ell, I_{\ell-1} \uplus I_\ell \uplus I_{\ell+1}} \right\|_F^2 \left\| \boldsymbol{v}_{I_{\ell-1} \uplus I_\ell \uplus I_{\ell+1}} \right\|_2^2 \tag{H.107}$$

$$\leq 3\kappa_I^2 \times (3\widetilde{\mu})^2 \times \sum_{\ell=0}^{r-1} \left\| \boldsymbol{v}_{I_{\ell-1} \uplus I_\ell \uplus I_{\ell+1}} \right\|_2^2 \tag{H.108}$$

$$\leq 3\kappa_I^2 \times 9\widetilde{\mu}^2 \times 3 \left\| \boldsymbol{v} \right\|_2^2, \tag{H.109}$$

giving the claimed result. ∎

As a consequence, we have that

**Corollary H.7** (Inverse of $[\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}}]_{J,J}$)**.** *Suppose that $\boldsymbol{a}_0$ is $\mu$-shift coherent, that $\|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \leq \widetilde{\mu}$ and that $\kappa_I \widetilde{\mu} < \frac{1}{18}$. Then for every $J \subseteq I$ and any norm $\|\cdot\|_\diamond \in \{ \|\cdot\|_{\square\to\square}, \|\cdot\|_{\infty\to\infty}, \|\cdot\|_2 \}$, we have*

$$\left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{J,J} \right\|_\diamond \leq 9\kappa_I \widetilde{\mu} \tag{H.110}$$

$$\left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}}]_{J,J}^{-1} - \boldsymbol{I} \right\|_\diamond \leq 18\kappa_I \widetilde{\mu} \tag{H.111}$$

$$\left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}}]_{J,J}^{-1} \right\|_\diamond \leq 2. \tag{H.112}$$

**Proof.** First we prove

$$\left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{J,J} \right\|_2 \leq 9\kappa_I \widetilde{\mu}, \quad \left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{J,J} \right\|_{\infty\to\infty} \leq 6\kappa_I \widetilde{\mu}, \quad \left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{J,J} \right\|_{\square\to\square} \leq 6\kappa_I \widetilde{\mu} \tag{H.113}$$

Where the first claim follows from Lemma H.6. The second follows by noting that the $\ell^\infty$ operator norm is the maximum row $\ell^1$ norm, and that each row has at most $2\kappa_I$ entries, of size at most $3\widetilde{\mu}$. The last follows by noting that

$$\begin{aligned} \left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{J,J} \right\|_{\square\to\square} &\leq \max_{\ell,\ell'} \left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}} - \boldsymbol{I}]_{J\cap([p]+\ell),\, J\cap([2p]+\ell')} \right\|_F \\ &\leq 6\kappa_I \widetilde{\mu}. \end{aligned} \tag{H.114}$$

Then we prove

$$\left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}}]_{J,J}^{-1} - \boldsymbol{I} \right\|_2 \leq 18\kappa_I \widetilde{\mu}, \quad \left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}}]_{J,J}^{-1} - \boldsymbol{I} \right\|_{\infty\to\infty} \leq 12\kappa_I \widetilde{\mu}, \quad \left\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}}]_{J,J}^{-1} - \boldsymbol{I} \right\|_{\square\to\square} \leq 12\kappa_I \widetilde{\mu}, \tag{H.115}$$

which are followed from the fact that if $\|\cdot\|_\diamond$ is a matrix norm and $\|\boldsymbol{\Delta}\|_\diamond < 1$, then

$$\left\| (\boldsymbol{I} + \boldsymbol{\Delta})^{-1} - \boldsymbol{I} \right\|_\diamond \leq \frac{\|\boldsymbol{\Delta}\|_\diamond}{1 - \|\boldsymbol{\Delta}\|_\diamond}.$$

Finally, (H.112) follows from the triangle inequality. ∎

Also, we need to bound the convolution of $\boldsymbol{a}_0 - \boldsymbol{a}$ with $\|\boldsymbol{a}_0 - \boldsymbol{a}\|_2$ requiring for bounds of the lasso solution:

**Lemma H.8** (Convolution of $\boldsymbol{a}_0 - \boldsymbol{a}$)**.** *Suppose that $\boldsymbol{a}_0$ is $\mu$-shift coherent and $\|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \leq \widetilde{\mu}$, then for every $J \subseteq I$,*

$$\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}}]_{J,J} \|_{\square\to\infty} \leq \sqrt{2\kappa_I} \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \tag{H.116}$$

$$\| [\boldsymbol{C}_{\boldsymbol{a}}^* \boldsymbol{C}_{\boldsymbol{a}_0 - \boldsymbol{a}}]_{J,J} \|_{\square\to\square} \leq \sqrt{2\kappa_I} \|\boldsymbol{a} - \boldsymbol{a}_0\|_2 \tag{H.117}$$

**Proof.** For the first inequality, we have

$$\begin{aligned}
\left\|[\boldsymbol{C}_{\boldsymbol{a}}^{*}\boldsymbol{C}_{\boldsymbol{a}_0-\boldsymbol{a}}]_{J,J}\boldsymbol{v}\right\|_{\square\to\infty} &= \max_{j\in J,\,\|\boldsymbol{v}\|_{\square}=1} \left|\langle s_j[\boldsymbol{a}], (\boldsymbol{a}_0-\boldsymbol{a})*\boldsymbol{v}\rangle\right| \\
&\le \max_{j\in[n],\,\|\boldsymbol{v}\|_{\square}=1} \left\|\boldsymbol{P}_{[p]+j}\left[(\boldsymbol{a}_0-\boldsymbol{a})*\boldsymbol{v}\right]\right\|_2 \\
&\le \|\boldsymbol{a}-\boldsymbol{a}_0\|_2 \times \max_{j\in[n],\,\|\boldsymbol{v}\|_{\square}=1} \left\|\boldsymbol{P}_{[\pm p]+j}\boldsymbol{v}\right\|_1 \\
&\le \sqrt{2\kappa_I}\,\|\boldsymbol{a}_0-\boldsymbol{a}\|_2 \qquad\qquad (\text{H.118})
\end{aligned}$$

The second inequality is derived by

$$\begin{aligned}
\left\|[\boldsymbol{C}_{\boldsymbol{a}}^{*}\boldsymbol{C}_{\boldsymbol{a}_0-\boldsymbol{a}}]_{J,J}\right\|_{\square\to\square} &\le \max_{\ell,\ell'} \left\|[\boldsymbol{C}_{\boldsymbol{a}}^{*}\boldsymbol{C}_{\boldsymbol{a}_0-\boldsymbol{a}}]_{J\cap([p]+\ell),J\cap([2p]+\ell')}\right\|_F \\
&\le \sqrt{2\kappa_I^2 \max_{i,j}\left|\langle s_i[\boldsymbol{a}], s_j[\boldsymbol{a}_0-\boldsymbol{a}]\rangle\right|^2} \\
&\le \sqrt{2}\kappa_I\,\|\boldsymbol{a}-\boldsymbol{a}_0\|_2\,, \qquad\qquad (\text{H.119})
\end{aligned}$$

finishing the proof. ■

Again, using a variant of the argument for Lemma H.6, we have the following:

**Lemma H.9** (Off-diagonal of submatrix of $\boldsymbol{C}_{\boldsymbol{a}}^{*}\boldsymbol{C}_{\boldsymbol{a}_0}$)**.** *Suppose that $\boldsymbol{a}_0$ is $\mu$-shift coherent and $\|\boldsymbol{a}-\boldsymbol{a}_0\|_2 \le \widetilde{\mu}$. For any $J\subset I$, if*

$$\kappa_J \;=\; \max_{\ell}\left|J\cap\{\ell,\ell+1,\dots,\ell+p-1\}\right| \qquad (\text{H.120})$$

$$\kappa_{I\setminus J} \;=\; \max_{\ell}\left|(I\setminus J)\cap\{\ell,\ell+1,\dots,\ell+p-1\}\right| \qquad (\text{H.121})$$

*Then*

$$\left\|[\boldsymbol{C}_{\boldsymbol{a}}^{*}\boldsymbol{C}_{\boldsymbol{a}_0}]_{J,I\setminus J}\right\|_2 \;\le\; 6\sqrt{\kappa_J\kappa_{I\setminus J}}\,\widetilde{\mu}. \qquad (\text{H.122})$$

**Proof.** Take $r=\lceil n/p\rceil$ and for $\ell=0,\dots,r-1$, write

$$J_\ell = J\cap([p]+p\ell), \qquad L_\ell = (I\setminus J)\cap([p]+p\ell),$$

Take $\boldsymbol{v}\in\mathbb{R}^{|I\setminus J|}$ arbitrary and notice that

$$\begin{aligned}
\left\|[\boldsymbol{C}_{\boldsymbol{a}}^{*}\boldsymbol{C}_{\boldsymbol{a}_0}]_{J,I\setminus J}\,\boldsymbol{v}\right\|_2^2 &= \sum_{\ell=0}^{r-1}\left\|[\boldsymbol{C}_{\boldsymbol{a}}^{*}\boldsymbol{C}_{\boldsymbol{a}_0}]_{J_\ell,I\setminus J}\,\boldsymbol{v}\right\|_2^2 \\
&= \sum_{\ell=0}^{r-1}\left\|[\boldsymbol{C}_{\boldsymbol{a}}^{*}\boldsymbol{C}_{\boldsymbol{a}_0}]_{J_\ell,L_{\ell-1}\cup L_\ell\cup L_{\ell+1}}\,\boldsymbol{v}_{L_{\ell-1}\cup L_\ell\cup L_{\ell+1}}\right\|_2^2 \\
&\le 4\widetilde{\mu}^2\times\kappa_J\times 3\kappa_{I\setminus J}\times\sum_{\ell=0}^{r-1}\left\|\boldsymbol{v}_{L_{\ell-1}\cup L_\ell\cup L_{\ell+1}}\right\|_2^2 \\
&\le 4\widetilde{\mu}^2\times\kappa_J\times 3\kappa_{I\setminus J}\times 3\|\boldsymbol{v}\|_2^2, \qquad\qquad (\text{H.123})
\end{aligned}$$

giving the result. ■

**Lemma H.10** (Perturbation of vector over sphere)**.** *If both $\boldsymbol{a},\boldsymbol{a}_0$ are unit vectors in inner product space, then*

$$\left|\langle\boldsymbol{a},\boldsymbol{a}-\boldsymbol{a}_0\rangle\right| \;\le\; \tfrac{1}{2}\,\|\boldsymbol{a}-\boldsymbol{a}_0\|_2^2\,. \qquad (\text{H.124})$$

**Proof.** Via simple norm inequalities:

$$\tfrac{1}{2}\left\|\boldsymbol{a}-\boldsymbol{a}_0\right\|_2^2 \;=\; 1-\langle \boldsymbol{a},\boldsymbol{a}_0\rangle \;=\; 1-\langle \boldsymbol{a},\boldsymbol{a}_0-\boldsymbol{a}+\boldsymbol{a}\rangle \;=\; \langle \boldsymbol{a},\boldsymbol{a}-\boldsymbol{a}_0\rangle \;>\; 0 \tag{H.125}$$

■

**Lemma H.11** (Convolution of short and sparse). *Suppose $\boldsymbol{\delta}\in\mathbb{R}^p$, and $\boldsymbol{v}\in\mathbb{R}^n$ where $\operatorname{supp}(\boldsymbol{v})=I$ satisfies*

$$\max_{\ell\in[n]} |\, I\cap([p]+\ell)\,| \;\le\; \kappa \tag{H.126}$$

*then*

$$\left\|\boldsymbol{\delta}*\boldsymbol{v}\right\|_2 \;\le\; \sqrt{2\kappa}\left\|\boldsymbol{\delta}\right\|_2\left\|\boldsymbol{v}\right\|_2 \tag{H.127}$$

**Proof.** Since every $p$-contiguous segment of $I$ has at most $\kappa$ elements, by splitting $I=I_1\uplus I_2\uplus,\ldots,\uplus I_\kappa\uplus R$ such that each sets $I_i$ are $p$-separated:

$$\begin{aligned}
I_1 &= \{i_1,i_{\kappa+1},i_{2\kappa+1},\ldots\}\cap\{0,\ldots,n-p-1\}\,,\\
I_2 &= \{i_2,i_{\kappa+2},i_{2\kappa+2},\ldots\}\cap\{0,\ldots,n-p-1\}\,,\\
&\;\vdots\\
I_\kappa &= \{i_\kappa,i_{2\kappa},i_{3\kappa},\ldots\}\cap\{0,\ldots,n-p-1\}\,, &\text{(H.128)}\\
R &= I\cap\{n-p,\ldots,n-1\}\,. &\text{(H.129)}
\end{aligned}$$

Then the p-separating property gives $\left\|\boldsymbol{\delta}*\boldsymbol{P}_{I_i}\boldsymbol{v}\right\|_2=\left\|\boldsymbol{\delta}\right\|_2\left\|\boldsymbol{P}_{I_i}\boldsymbol{v}\right\|_2$. Hence:

$$\begin{aligned}
\left\|\boldsymbol{\delta}*\boldsymbol{P}_I\boldsymbol{v}\right\|_2 &= \left\|\sum_{i\in\kappa}\boldsymbol{\delta}*\boldsymbol{P}_{I_i}\boldsymbol{v}+\boldsymbol{\delta}*\boldsymbol{P}_R\boldsymbol{v}\right\|_2 \;\le\; \sum_{i\in\kappa}\left\|\boldsymbol{\delta}*\boldsymbol{P}_{I_i}\boldsymbol{v}\right\|_2+\left\|\boldsymbol{\delta}*\boldsymbol{P}_R\boldsymbol{v}\right\|\\
&= \left\|\boldsymbol{\delta}\right\|_2\sum_{i\in\kappa}\left\|\boldsymbol{v}_{I_i}\right\|_2 \;+\; \left\|\boldsymbol{\delta}\right\|_2\left\|\boldsymbol{P}_R\boldsymbol{v}\right\|_1\\
&\le \sqrt{\kappa}\left\|\boldsymbol{v}_{I_1,\uplus,\ldots,\uplus I_\kappa}\right\|_2\left\|\boldsymbol{\delta}\right\|_2 \;+\; \sqrt{\kappa}\left\|\boldsymbol{v}_R\right\|_2\left\|\boldsymbol{\delta}\right\|_2\\
&\le \sqrt{2\kappa}\left\|\boldsymbol{v}\right\|_2\left\|\boldsymbol{\delta}\right\|_2\,, &\text{(H.130)}
\end{aligned}$$

where the last two inequalities were coming from Cauchy-Schwartz. ■

# I Finite sample approximation

In this section we collect several major components of proof about large sample deviation. In particular, the concentration for shift space gradient $\boldsymbol{\chi}(\boldsymbol{\beta})_i$, shift space Hessian diagonals $\left\|\boldsymbol{P}_{I(\boldsymbol{a})}s_{-i}[\boldsymbol{x}_0]\right\|_2$, and the set of gradients discontinuity entries $|J_B(\boldsymbol{a})|$.

## I.1 Proof of Corollary C.4

**Proof.** 1. (<u>$\varepsilon$-net</u>) Write $\boldsymbol{x}$ as $\boldsymbol{x}_0$ and $\|\boldsymbol{\beta}\|_2 = \eta$ through out this proof, firstly from Definition B.1 for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$, we know $\eta \le 1 + c_\mu + \frac{c_\mu}{\sqrt{\theta k}\log\theta^{-1}} \le \sqrt{p}$. Define $\varepsilon = \frac{c_2}{2n^{3/2}p^{3/2}}$ and consider the $\varepsilon$-net $\mathcal{N}_\varepsilon$ for sphere of radius $\eta$. From Lemma J.5 we know for any $c_2 < 1$:

$$|\mathcal{N}_\varepsilon| \le \left(\frac{3\eta}{\varepsilon}\right)^{2p} \le \left(\frac{3n^{3/2}p^2}{c_2}\right)^{2p} \le \left(\frac{3np^2}{c_2}\right)^{3p} \tag{I.1}$$

for each $i \in [n]$ define such net as $\mathcal{N}_{\varepsilon,i}$, and define an event such that all center of subsets in $\mathcal{N}_{\varepsilon,i}$ are being well-behaved:

$$\mathcal{E}_{\mathrm{Net}} := \left\{\forall\, i \in [n], \quad \boldsymbol{\sigma}_i n^{-1}\boldsymbol{\chi}[\boldsymbol{\beta}_\varepsilon]_i - \boldsymbol{\sigma}_i n^{-1}\overline{\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}_\varepsilon]_i} < \frac{c_1\theta}{p^{3/2}} \quad \forall\, \boldsymbol{\beta}_\varepsilon \in \mathcal{N}_{\varepsilon,i},\right\} \tag{I.2}$$

2. (<u>Lipschitz constant</u>) The Lipschitz constant $L$ of $\boldsymbol{\chi}[\cdot]_i$ w.r.t $\boldsymbol{\beta}$ is bounded in terms of $\boldsymbol{x}$ regardless of entry $i$:

$$\begin{aligned}
|\boldsymbol{\chi}[\boldsymbol{\beta}]_i - \boldsymbol{\chi}[\boldsymbol{\beta}']_i| &\le \left|\boldsymbol{e}_i^*\check{\boldsymbol{C}}_{\boldsymbol{x}}\mathcal{S}_\lambda\left[\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\beta}\right] - \boldsymbol{e}_i^*\check{\boldsymbol{C}}_{\boldsymbol{x}}\mathcal{S}_\lambda\left[\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\beta}'\right]\right| \le \|\boldsymbol{x}\|_2\left\|\mathcal{S}_\lambda\left[\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\beta}\right] - \mathcal{S}_\lambda\left[\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\beta}'\right]\right\|_2 \\
&\le \|\boldsymbol{x}\|_2\sqrt{\sum_{j\in[n]}\left|\mathcal{S}_\lambda\left[\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\beta}\right]_j - \mathcal{S}_\lambda\left[\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\beta}'\right]_j\right|^2} \le \|\boldsymbol{x}\|_2\left\|\check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\beta} - \check{\boldsymbol{C}}_{\boldsymbol{x}}\boldsymbol{\beta}'\right\|_2 \\
&\le \|\boldsymbol{x}\|_2\cdot\|\boldsymbol{x}\|_1\cdot\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 =: L\,\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2
\end{aligned} \tag{I.3}$$

Define the event that $\boldsymbol{\chi}[\boldsymbol{\beta}]_i$ that has small Lipschitz constant as

$$\mathcal{E}_{\mathrm{Lip}} := \left\{L < 2n^{3/2}\theta\right\} \tag{I.4}$$

on the event $\mathcal{E}_{\mathrm{Lip}}$, for every points in $\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ and $i \in [n]$, there exists some $\boldsymbol{\beta}_\varepsilon \in \mathcal{N}_{\varepsilon,i}$ such that

$$\left|\left(\boldsymbol{\sigma}_i n^{-1}\boldsymbol{\chi}[\boldsymbol{\beta}]_i - \boldsymbol{\sigma}_i n^{-1}\overline{\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i}\right) - \left(\boldsymbol{\sigma}_i n^{-1}\boldsymbol{\chi}[\boldsymbol{\beta}_\varepsilon]_i - \boldsymbol{\sigma}_i n^{-1}\overline{\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}_\varepsilon]_i}\right)\right| \le 2L\varepsilon \le \frac{c_2\theta}{p^{3/2}} \tag{I.5}$$

On event $\mathcal{E}_{\mathrm{Lip}}\cap\mathcal{E}_{\mathrm{Net}}$, (I.2), (I.5) implies $\boldsymbol{\chi}[\boldsymbol{\beta}]$ is well concentrated entrywise and anywhere in $\cup_{|\boldsymbol{\tau}|\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$:

$$\left|\boldsymbol{\sigma}_i n^{-1}\boldsymbol{\chi}[\boldsymbol{\beta}]_i - \boldsymbol{\sigma}_i n^{-1}\overline{\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}]_i}\right| \le \frac{(c_1+c_2)\theta}{p^{3/2}}, \quad \forall\, \boldsymbol{a} \in \cup_{k\le k}\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu)),\ \forall\, i \in [n] \tag{I.6}$$

as desired, where, using Lemma A.2,

$$\mathbb{P}\left[\mathcal{E}_{\mathrm{Lip}}^c\right] \le \mathbb{P}\left[\|\boldsymbol{x}\|_2^2 > 2n\theta\right] \le 1/n; \tag{I.7}$$

and using union bound,

$$\mathbb{P}\left[\mathcal{E}_{\mathrm{Net}}^c\right] \le \mathbb{P}\left[\max_{\substack{\boldsymbol{a}_\varepsilon\in\mathcal{N}_{\varepsilon,i}\\ i\in[n]}} \boldsymbol{\sigma}_i n^{-1}\boldsymbol{\chi}[\boldsymbol{\beta}_\varepsilon]_i - \boldsymbol{\sigma}_i n^{-1}\overline{\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}_\varepsilon]_i} > \frac{c_1\theta}{p^{3/2}}\right]$$

$$\leq n\left|\mathcal{N}_\varepsilon\right|\mathbb{P}\left[\boldsymbol{\sigma}_0 n^{-1}\boldsymbol{\chi}[\boldsymbol{\beta}_\varepsilon]_0-\boldsymbol{\sigma}_0 n^{-1}\mathbb{E}\boldsymbol{\chi}[\boldsymbol{\beta}_\varepsilon]_0>\frac{c_1\theta}{p^{3/2}}\right]. \tag{I.8}$$

3. ( $\underline{\text{Bound } \mathbb{P}\left[\mathcal{E}_{\text{Net}}^c\right]}$) Wlog write $n=t\cdot(2p)$ for some integer $t$ and $2p\geq 4p_0-3$ and replace $\boldsymbol{x}_0$ with $\boldsymbol{x}$. Observe that $\boldsymbol{Z}_j(\boldsymbol{\beta})$ from (C.9) is independent of $\boldsymbol{Z}_{j+2p}(\boldsymbol{\beta})$ for all $j\in[n]$ while all $\boldsymbol{Z}_j$ are identical distributed. We write $\boldsymbol{\chi}[\boldsymbol{\beta}]_0$ as sum of iid r.v.s. as

$$\boldsymbol{\chi}[\boldsymbol{\beta}]_0=\sum_{j\in[n]}\boldsymbol{Z}_j(\boldsymbol{\beta})=\sum_{k\in[2p]}\left(\sum_{t=0}^{n/2p-1}\boldsymbol{Z}_{k+2tp}(\boldsymbol{\beta})\right)$$

wlog let $\boldsymbol{\sigma}_0=1$ and split the independent r.v.s, write $\mathbb{E}\boldsymbol{Z}_0=\mathbb{E}\boldsymbol{Z}$, bound the tail probability of $\boldsymbol{\chi}[\boldsymbol{\beta}]_0$ as

$$\mathbb{P}\left[n^{-1}\boldsymbol{\chi}[\boldsymbol{\beta}]_0>n^{-1}\overline{\mathbb{E}\boldsymbol{\chi}(\boldsymbol{\beta})}_0+\frac{c_1\theta}{p^{3/2}}\right]\leq 2p\cdot\mathbb{P}\left[\sum_{t=0}^{n/2p-1}\boldsymbol{Z}_{2tp}(\boldsymbol{\beta})>\frac{n}{2p}\mathbb{E}\boldsymbol{Z}(\boldsymbol{\beta})+\frac{c_1 n\theta}{2p^{5/2}}\right] \tag{I.9}$$

The moments of $\boldsymbol{Z}_0$ can be bounded by using $|\boldsymbol{Z}_0(\boldsymbol{\beta})|\leq|\boldsymbol{x}_0|\,|\boldsymbol{\beta}_0\boldsymbol{x}_0+\boldsymbol{s}_0|\leq\boldsymbol{\beta}_0\boldsymbol{x}_0^2+|\boldsymbol{x}_0|\,|\boldsymbol{s}_0|$ where $\boldsymbol{s}_0=\sum_{\ell\neq 0}\boldsymbol{x}_\ell\boldsymbol{\beta}_\ell$, write $\boldsymbol{x}=\boldsymbol{\omega}\circ\boldsymbol{g}\sim_{\text{i.i.d.}}\mathrm{BG}(\theta)$. For the 2-norm we know

$$\mathbb{E}\left|\boldsymbol{s}_0\right|^2=\mathbb{E}\left|\sum_\ell\boldsymbol{x}_\ell\boldsymbol{\beta}_\ell\right|^2\leq\theta\left\|\boldsymbol{\beta}\right\|_2^2\leq\theta\left(1+c_\mu+\frac{c_\mu}{\theta k^2}\right)\leq\frac{1}{2} \tag{I.10}$$

As for the $q$-norm, use the moment generating function bound, such that for all $t\geq 0$:

$$\begin{aligned}\mathbb{E}\left|\boldsymbol{s}_0\right|^q&\leq q!t^{-q}\mathbb{E}\exp\left[t\left|\boldsymbol{s}_0\right|\right]\leq q!t^{-q}\prod_\ell\mathbb{E}_{\boldsymbol{\omega}_\ell,\boldsymbol{g}_\ell}\exp\left[t\boldsymbol{\omega}_\ell\left|\boldsymbol{g}_\ell\right|\left|\boldsymbol{\beta}_\ell\right|\right]\leq 2q!t^{-q}\prod_\ell\mathbb{E}_{\boldsymbol{\omega}_\ell}\exp\left[\boldsymbol{\omega}_\ell t^2\boldsymbol{\beta}_\ell^2/2\right]\\&\leq 2q!t^{-q}\prod_\ell\left(1-\theta+\theta\exp\left[t^2\boldsymbol{\beta}_\ell^2/2\right]\right)\end{aligned} \tag{I.11}$$

notice that the entrywise twice derivative of (I.11) w.r.t. $\boldsymbol{\beta}_\ell^2$'s are always positive, this function is convex for all $\beta_\ell^2$. Constrain on the polytope $\sum_\ell\boldsymbol{\beta}_\ell^2\leq\|\boldsymbol{\beta}\|_2^2$, the maximizer of (I.11) w.r.t. $\boldsymbol{\beta}_\ell^2$'s occurs and a vertex point where $\boldsymbol{\beta}_0^2=\|\boldsymbol{\beta}\|_2^2$. Thus

$$\text{(I.11)}\leq 2q!t^{-q}\left(1-\theta+\theta\exp\left[t^2\left\|\boldsymbol{\beta}\right\|_2^2/2\right]\right)\prod_{\ell\neq 0}(1-\theta+\theta e^0)\leq 2q!t^{-q}(1+\theta\exp[\|\boldsymbol{\beta}\|_2^2\,t^2/2]).$$

Choose $t=\sqrt{q}/\left\|\boldsymbol{\beta}\right\|_2$, use $q!!>(q!/2)\cdot(e/q)^{q/2}$, we have

$$\mathbb{E}\left|\boldsymbol{s}_0\right|^q\leq 2q!q^{-q/2}\left\|\boldsymbol{\beta}\right\|_2^q\left(1+\theta\exp\left[q/2\right]\right)\leq 8\left\|\boldsymbol{\beta}\right\|_2^q\max\left\{e^{-q/2},\theta\right\}q!!. \tag{I.12}$$

Apply Jensen's inequality $\left(\sum_{i=1}^N\boldsymbol{z}_i\right)^q\leq N^{q-1}\sum_{i=1}^N\boldsymbol{z}_i^q$, use Gaussian moment Lemma J.2 , (I.10) and (I.12), obtain for $q\geq 3$,

$$\begin{aligned}\mathbb{E}Z(\boldsymbol{\beta})^2&\leq\mathbb{E}\left(\boldsymbol{\beta}_0\boldsymbol{x}_0^2+|\boldsymbol{x}_0|\,|\boldsymbol{s}_0|\right)^2\leq 2\mathbb{E}\left[\boldsymbol{\beta}_0^2\boldsymbol{x}_0^4+\boldsymbol{x}_0^2\boldsymbol{s}_0^2\right]\leq 6\theta+2\theta^2\left\|\boldsymbol{\beta}\right\|_2^2\leq 7\theta,\\\mathbb{E}\boldsymbol{Z}(\boldsymbol{\beta})^q&\leq\mathbb{E}\left(\boldsymbol{\beta}_0\boldsymbol{x}_0^2+|\boldsymbol{x}_0|\,|\boldsymbol{s}_0|\right)^q\leq 2^{q-1}\left(\mathbb{E}\boldsymbol{x}_0^{2q}+\mathbb{E}\left|\boldsymbol{x}_0\right|^q\mathbb{E}\left|\boldsymbol{s}_0\right|^q\right)\\&\leq\theta 2^{q-1}(2q-1)!!+\theta 2^{q-1}(q-1)!!\left(8\left\|\boldsymbol{\beta}\right\|_2^q\max\left\{e^{-q/2},\theta\right\}q!!\right)\\&\leq\theta 4^q q!+\theta 2^q\left\|\boldsymbol{\beta}\right\|_2^q q!.\end{aligned}$$

Thus, recall that $\|\boldsymbol{\beta}\|_2 = \eta$, use $(\sigma^2, R) = (8\theta\eta^2, 4\eta)$, from (I.8)-(I.9), apply Bernstein inequality Lemma J.4 with $n \geq Cp^5\theta^{-2}\log p$, and $c_1, c_2 \in [0,1]$ we have

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}_{\mathrm{Net}}^c\right] &\leq 2np\left|\mathcal{N}_\varepsilon\right| \cdot \mathbb{P}\left[\sum_{t=0}^{n/2p-1} \boldsymbol{Z}_{2tp}(\boldsymbol{\beta}) > \frac{n}{2p}\mathbb{E}\boldsymbol{Z}(\boldsymbol{\beta}) + \frac{c_1 n\theta}{2p^{5/2}}\right] \leq 2np\left(\frac{3np^2}{c_2}\right)^{3p} \exp\left(\frac{-\left(c_1 n\theta/2p^{5/2}\right)^2}{16n\theta\eta^2/2p + 8\eta c_1 n\theta/2p^{5/2}}\right) \\
&\leq \exp\left(4p\log\left(\frac{3np^2}{c_2}\right) - \frac{\left(c_1 n\theta/2p^{5/2}\right)^2}{16n\theta\eta^2/p}\right) \leq \exp\left(4p\log\left(\frac{3np^2}{c_2}\right) - \frac{c_1^2 n\theta^2}{64p^4}\right) \\
&\leq \exp\left(\frac{-c_1^2 n\theta^2}{100p^4}\right) \leq \frac{1}{n}
\end{aligned} \tag{I.13}$$

when $\frac{C}{\log C} > \frac{10^5}{c_1^2 c_2}$. The proof of lower bound and negative $\beta_0$ is derived in the same manner. ■

## I.2 Proof of Corollary D.3

**Proof.** Write $\boldsymbol{x}$ as $\boldsymbol{x}_0$ though our this proof. Write $\boldsymbol{\beta}_i\boldsymbol{x}_j + \boldsymbol{s}_j = \sum_{\ell\in[\pm p]}\boldsymbol{\beta}_\ell\boldsymbol{x}_{\ell-i+j} = \left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[\pm p]-i+j}\right\rangle$, and the support w.r.t. some $\boldsymbol{a}$ as $I(\boldsymbol{\beta})$. Define the random variable $\boldsymbol{Z}_{ij}(\boldsymbol{\beta})$ as

$$\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})}s_{-i}[\boldsymbol{x}]\right\|_2^2 = \sum_{j\in[n]} \boldsymbol{x}_j^2 \mathbf{1}_{\left\{\left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[\pm p]-i+j}\right\rangle\right| > \lambda\right\}} =: \sum_{j\in[n]} \boldsymbol{Z}_{ij}(\boldsymbol{\beta}) \tag{I.14}$$

and define $\left\{\overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta})\right\}_{j\in[n]}$ that are independent r.v.s. and as a upper bounding function of $\boldsymbol{Z}_{ij}(\boldsymbol{\beta})$ as

$$\overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}) := \begin{cases} \boldsymbol{x}_j^2, & \left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[\pm p]-i+j}\right\rangle\right| > \lambda \\ 0, & \left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[\pm p]-i+j}\right\rangle\right| < \lambda/2\,, \\ \frac{\boldsymbol{x}_j^2}{\lambda/2}\left(\left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[\pm p]-i+j}\right\rangle\right| - \lambda/2\right), & \text{otherwise} \end{cases} \tag{I.15}$$

Similar to proof of Corollary C.4. Let $\|\boldsymbol{\beta}\|_2 \leq \eta \leq \sqrt{p}$. Define $\varepsilon = \frac{c_2'\lambda}{24np\sqrt{p\theta\log n\log\theta^{-1}}}$ for some $c_2' > 0$ and consider the $\varepsilon$-net $\mathcal{N}_\varepsilon$ for sphere of radius $\eta$. From Lemma J.5 we know

$$|\mathcal{N}_\varepsilon| \leq \left(\frac{3\eta}{\varepsilon}\right)^{2p} \leq \left(\frac{72}{c_2'c_\lambda}np^2\sqrt{\theta\,|\boldsymbol{\tau}|\log n\log\theta^{-1}}\right)^{2p} \leq \left(\frac{72}{c_2'c_\lambda}np^2\log n\right)^{2p}, \tag{I.16}$$

for each $i \in [n]$ define such net as $\mathcal{N}_{\varepsilon,i}$, and define an event such that all center of subsets in $\mathcal{N}_{\varepsilon,i}$ are being well-behaved:

$$\mathcal{E}_{\mathrm{Net}} := \left\{\forall\, i \in [n], \quad \left|n^{-1}\sum_{j\in[n]}\overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}_\varepsilon) - \mathbb{E}\overline{\boldsymbol{Z}}_i(\boldsymbol{\beta}_\varepsilon)\right| \leq \frac{c_1'\theta}{p} \quad \forall\,\boldsymbol{\beta}_\varepsilon \in \mathcal{N}_{\varepsilon,i}\right\}, \tag{I.17}$$

Also, $\sum_j \overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta})$ is a Lipchitz function over $\boldsymbol{\beta}$ for every $i \in [n]$ as

$$\begin{aligned}
\left|\sum_{j\in[n]}\overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}) - \sum_{j\in[n]}\overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}')\right| &\leq \sum_{j\in[n]}\frac{\boldsymbol{x}_j^2}{\lambda/2}\left|\left\langle\boldsymbol{\beta}-\boldsymbol{\beta}', \boldsymbol{x}_{[\pm p]-i+j}\right\rangle\right| \leq \sum_{j\in[n]}\frac{\boldsymbol{x}_j^2\left\|\boldsymbol{x}_{[\pm p]-i+j}\right\|_2}{\lambda/2}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}'\right\|_2, \\
&\leq \frac{1}{\lambda/2}\|\boldsymbol{x}\|_2^2 \cdot \max_{j\in[n]}\left\|\boldsymbol{x}_{[\pm p]+j}\right\|_2 \cdot \left\|\boldsymbol{\beta}-\boldsymbol{\beta}'\right\|_2 := L\left\|\boldsymbol{\beta}-\boldsymbol{\beta}'\right\|_2,
\end{aligned} \tag{I.18}$$

and define event $\mathcal{E}_{\mathrm{Lip}}$ such that the Lipchitz constant is bounded as

$$\mathcal{E}_{\mathrm{Lip}} := \left\{L \leq 12n\theta\sqrt{p\theta\log n\log\theta^{-1}}\lambda^{-1}\right\}, \tag{I.19}$$

then on event $\mathcal{E}_{\text{Lip}}$, for any points $\boldsymbol{\beta}$ in $\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ and $i \in [n]$, there exists some $\boldsymbol{\beta}_\varepsilon$ in $\mathcal{N}_{\varepsilon,i}$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_\varepsilon\|_2 \le \varepsilon$, and thus

$$\left|\left(n^{-1}\sum_{j\in[n]} \overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}) - \mathbb{E}\overline{\boldsymbol{Z}}_i(\boldsymbol{\beta})\right) - \left(n^{-1}\sum_{j\in[n]} \overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}_\varepsilon) - \mathbb{E}\overline{\boldsymbol{Z}}_i(\boldsymbol{\beta}_\varepsilon)\right)\right| \le 2L\varepsilon \le \frac{c_2'\theta}{p}. \tag{I.20}$$

On event $\mathcal{E}_{\text{Lip}} \cap \mathcal{E}_{\text{Net}}$, from (I.17), (I.20), we can conclude that for all $\boldsymbol{\beta} \in \mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ and $i \in [n]$ that:

$$n^{-1}\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}[\boldsymbol{x}_0]\right\|_2^2 - n^{-1}\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}[\boldsymbol{x}_0]\right\|_2^2 \le n^{-1}\sum_{j\in[n]} \overline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}) - \mathbb{E}\overline{\boldsymbol{Z}}_i(\boldsymbol{\beta}) \le \frac{(c_1' + c_2')\theta}{p} \tag{I.21}$$

as desired, where the error probability of $\mathcal{E}_{\text{Lip}}^c$ is bounded using Lemma A.2 and Lemma A.3, which give

$$\mathbb{P}\left[\mathcal{E}_{\text{Lip}}^c\right] \le \mathbb{P}\left[\|\boldsymbol{x}\|_2^2 > 2n\theta\right] + \mathbb{P}\left[\max_{j\in[n]} \left\|\boldsymbol{x}_{[\pm p]+j}\right\|_2 > 3\sqrt{p\theta\log n\log\theta^{-1}}\right] \le 3/n, \tag{I.22}$$

when $n > 10^3\theta^{-1}$. As for $\mathcal{E}_{\text{Net}}^c$ use union bound and split the r.v.s since $\boldsymbol{Z}_j, \boldsymbol{Z}_{j+2p}$ are independent for all $j$:

$$\mathbb{P}\left[\mathcal{E}_{\text{Net}}^c\right] \le 2np \cdot |\mathcal{N}_\varepsilon| \cdot \mathbb{P}\left[\left|\sum_k^{n/2p} \overline{\boldsymbol{Z}}_{i,2kj}(\boldsymbol{\beta}) - \frac{n}{2p}\mathbb{E}\overline{\boldsymbol{Z}}_i(\boldsymbol{\beta})\right| \ge \frac{c_1' n\theta}{2p^2}\right].$$

Now we calculate the variance and $L^q$-norm of $\sum_k \overline{\boldsymbol{Z}}_{i,2kj}$ for $q \ge 3$:

$$\begin{cases} \mathbb{E}\overline{\boldsymbol{Z}}_{i,j}^2 \le \mathbb{E}\boldsymbol{x}_j^4 \le 3\theta \\ \mathbb{E}\overline{\boldsymbol{Z}}_{i,j}^q \le \mathbb{E}\boldsymbol{x}_j^{2q} \le \theta(2q-1)!! \le \frac{1}{2}\cdot(3\theta)\cdot 2^{q-2}q! \end{cases} \tag{I.23}$$

and apply Bernstein inequality with $(\sigma^2, R) = (3\theta, 2)$, then use $n \ge Cp^4\theta^{-1}\log p$ and $c_1', c_2' < 1$ to obtain

$$\begin{aligned} 2np\,|\mathcal{N}_\varepsilon|\,\mathbb{P}\left[\left|\sum_k^{n/2p} \overline{\boldsymbol{Z}}_{i,2kj}(\boldsymbol{\beta}) - \frac{n}{2p^2}\mathbb{E}\overline{\boldsymbol{Z}}_i\right| \ge \frac{c_1' n\theta}{2p^2}\right] &\le \exp\left[\log(2np) + 2p\log\left(\frac{72}{c_2' c_\lambda}np^2\log n\right) - \frac{(c_1' n\theta/2p^2)^2}{6n\theta/2p + 4c_1' n\theta/2p^2}\right] \\ &\le \exp\left[3p\log\left(\frac{72}{c_2' c_\lambda}np^2\log n\right) - \frac{c_1'^2 n\theta}{24p^3}\right] \\ &\le \exp[-c_1'^2 n\theta/(50p^3)] \le 1/n, \end{aligned} \tag{I.24}$$

where the last two inequalities holds when $\frac{C}{\log C} \ge \frac{10^5}{c_1'^2 c_2' c_\lambda}$. The other side of inequality of (D.9) can be derived by defining $\underline{\boldsymbol{Z}}_{ij}$ as

$$\underline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}) := \begin{cases} \boldsymbol{x}_j^2, & \left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[\pm p]-i+j}\right\rangle\right| > 3\lambda/2 \\ 0, & \left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[\pm p]-i+j}\right\rangle\right| < \lambda \\ \frac{\boldsymbol{x}_j^2}{\lambda/2}\left(\left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[\pm p]-i+j}\right\rangle\right| - \lambda\right), & \text{otherwise} \end{cases}, \tag{I.25}$$

and define $\mathcal{E}_{\text{Net}}$, $\mathcal{E}_{\text{Lip}}$ similarly, such that on intersection of these events,

$$n^{-1}\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}[\boldsymbol{x}]\right\|_2^2 - n^{-1}\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}[\boldsymbol{x}]\right\|_2^2 \ge n^{-1}\sum_{j\in[n]} \underline{\boldsymbol{Z}}_{ij}(\boldsymbol{\beta}) - \mathbb{E}\underline{\boldsymbol{Z}}_i(\boldsymbol{\beta}) \ge \frac{(c_1' + c_2')\theta}{p} \tag{I.26}$$

as desired. ■

## I.3 Proof of Lemma E.5

**Proof.** 1. (Expectation upper bound) We will write $\boldsymbol{x}$ as $\boldsymbol{x}_0$. Similar to proof of Corollary C.4 let $\|\boldsymbol{\beta}\|_2 \leq \eta \leq \sqrt{p}$. For each $i \in [n]$, define the random variable

$$\boldsymbol{X}_i(\boldsymbol{\beta}) = \mathbf{1}_{\{|\langle s_i[\boldsymbol{x}],\boldsymbol{\beta}\rangle - \lambda| \leq B\}} + \mathbf{1}_{\{|\langle s_i[\boldsymbol{x}],\boldsymbol{\beta}\rangle + \lambda| \leq B\}}, \tag{I.27}$$

then number of indices for vector $\boldsymbol{x} * \check{\boldsymbol{\beta}}$ that are within $B$ of $\pm\lambda$ is a random variable $\sum_{i\in[n]} \boldsymbol{X}_i(\boldsymbol{\beta})$. For each of the $\boldsymbol{X}_i(\boldsymbol{\beta})$'s consider an upper bound $\overline{\boldsymbol{X}}_i(\boldsymbol{\beta})$ defined as

$$\overline{\boldsymbol{X}}_i(\boldsymbol{\beta}) = \begin{cases} \frac{1}{M}\left(|\langle s_i[\boldsymbol{x}],\boldsymbol{\beta}\rangle| - (\lambda - B - M)\right) & |\langle s_i[\boldsymbol{x}],\boldsymbol{\beta}\rangle| \in [\lambda - B - M, \lambda - B] \\ 1 & |\langle s_i[\boldsymbol{x}],\boldsymbol{\beta}\rangle| \in [\lambda - B, \lambda + B] \\ \frac{1}{M}\left((\lambda + B + M) - |\langle s_i[\boldsymbol{x}],\boldsymbol{\beta}\rangle|\right) & |\langle s_i[\boldsymbol{x}],\boldsymbol{\beta}\rangle| \in [\lambda + B, \lambda + B + M] \\ 0 & \text{else} \end{cases} \tag{I.28}$$

where $B < M = c\lambda\theta^2/\left(p\log n\right) \leq \lambda/4$ for some constant $0 < c < 1$.

Notice that $\boldsymbol{x} \sim_{\text{i.i.d.}} \mathrm{BG}(\theta)$ is equal in distribution to $\boldsymbol{P}_{I(\boldsymbol{a})}\boldsymbol{g}$, where $\boldsymbol{g} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$, and $I(\boldsymbol{a}) \subseteq [n]$ is an independent Bernoulli subset. Conditioned on $I(\boldsymbol{a})$, $\langle \boldsymbol{x},\boldsymbol{\beta}\rangle = \left\langle \boldsymbol{g}, \boldsymbol{P}_{I(\boldsymbol{a})}\boldsymbol{\beta}\right\rangle \sim \mathcal{N}(0, \left\|\boldsymbol{P}_{I(\boldsymbol{a})}\boldsymbol{\beta}\right\|_2^2)$. For all realizations of $I(\boldsymbol{a})$, the variance $\left\|\boldsymbol{P}_{I(\boldsymbol{a})}\boldsymbol{\beta}\right\|_2^2$ is bounded by $\left\|\boldsymbol{P}_{I(\boldsymbol{a})}\boldsymbol{\beta}\right\|_2^2 \leq \|\boldsymbol{\beta}\|_2^2 \leq p$. Using these observations, and letting $f_\sigma(t) = \left(\sqrt{2\pi}\sigma\right)^{-1}\exp\left(-t^2/2\sigma^2\right)$ denote the pdf of an $\mathcal{N}(0,\sigma^2)$ random variable, the expectation of $\sum_i \overline{\boldsymbol{X}}_i(\boldsymbol{\beta})$ can be upper bounded as

$$\begin{aligned} \sum_{i\in[n]} \mathbb{E}\left[\overline{\boldsymbol{X}}_i(\boldsymbol{\beta})\right] &\leq (2n)\cdot\mathbb{P}\left[\langle \boldsymbol{x},\boldsymbol{\beta}\rangle \in [\lambda - B - M, \lambda + B + M]\right] \\ &\leq (2n)\cdot 2(B+M) \sup_{\sigma^2\in(0,p]} \max_{t\in[\lambda-B-M,\lambda+B+M]} f_\sigma(t) \\ &\leq 4n(B+M) \sup_{\sigma^2\in(0,p]} f_\sigma\left(\lambda - B - M\right) \\ &\leq 4n(B+M) \sup_{\sigma^2\in(0,p]} f_\sigma\left(\lambda/2\right). \end{aligned} \tag{I.29}$$

Notice that

$$\frac{d}{d\sigma}f_\sigma\left(\frac{\lambda}{2}\right) = \frac{d}{d\sigma}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\lambda^2}{8\sigma^2}\right) = \frac{\lambda^2 - 4\sigma^2}{4\sqrt{2\pi}\sigma^4}\exp\left(-\frac{\lambda^2}{8\sigma^2}\right),$$

and hence $f_\sigma(\lambda/2)$ is maximized at either $\sigma^2 = 0$, $\sigma^2 = p$ or $\sigma^2 = \lambda^2/4$. Comparing values at these points, we obtain that

$$\sup_{\sigma^2\in(0,p]} f_\sigma(\lambda/2) \quad\leq\quad f_{\lambda/2}(\lambda/2) \quad\leq\quad \frac{1}{\sqrt{2\pi}(\lambda/2)}\exp\left(-\frac{1}{2}\right) \quad\leq\quad \frac{1}{2\lambda}, \tag{I.30}$$

whence, by letting $B \leq c\lambda\theta^2/\left(p\log n\right)$, the upper bound of expectation become:

$$\sum_{i\in[n]} \mathbb{E}\left[\overline{\boldsymbol{X}}_i(\boldsymbol{\beta})\right] \leq \frac{4n}{2\lambda}(B+M) \leq \frac{4cn\theta^2}{p\log n} =: n\mathbb{E}\overline{\boldsymbol{X}(\boldsymbol{\beta})}. \tag{I.31}$$

2. ($\varepsilon$-net) Define $\varepsilon = \frac{c^2\lambda\theta^{3.5}}{3p^{2.5}\log^{2.5} n \log^{0.5}\theta^{-1}}$. Write $\lambda = c_\lambda/\sqrt{|\boldsymbol{\tau}|}$ and consider the $\varepsilon$-net $\mathcal{N}_\varepsilon$ for sphere of radius $\eta \leq \sqrt{p}$. From Lemma J.5 we know

$$|\mathcal{N}_\varepsilon| \leq \left(\frac{3\eta}{\varepsilon}\right)^{2p} \leq \left(\frac{81\,|\boldsymbol{\tau}|\,p^6\log^5 n\log\theta^{-1}}{c^4c_\lambda^2\theta^7}\right)^p \leq \left(\frac{2p\log n}{c\cdot c_\lambda}\right)^{13p} \tag{I.32}$$

and define an event such that all center of subsets in $\mathcal{N}_\varepsilon$ are being well-behaved:

$$\mathcal{E}_{\text{Net}} := \left\{ \sum_{i\in[n]} \overline{\boldsymbol{X}}_i(\boldsymbol{\beta}_\varepsilon) - n\mathbb{E}\overline{\boldsymbol{X}}(\boldsymbol{\beta}_\varepsilon) < \frac{18cn\theta^2}{p\log n} \quad \forall \boldsymbol{\beta}_\varepsilon \in \mathcal{N}_\varepsilon, \right\} \tag{I.33}$$

3. (Lipschitz constant) Furthermore, the function $\sum_i^n \overline{\boldsymbol{X}}_i(\boldsymbol{\beta})$ is Lipchitz over $\boldsymbol{\beta}$ such that

$$\left| \sum_{i\in[n]} \overline{\boldsymbol{X}}_i(\boldsymbol{\beta}) - \sum_{i\in[n]} \overline{\boldsymbol{X}}_i(\boldsymbol{\beta}') \right| \le \sum_{i\in[n]}^{n} \frac{1}{M} \left| \langle s_i[\boldsymbol{x}], \boldsymbol{\beta} - \boldsymbol{\beta}' \rangle \right| \le \frac{n}{M} \max_{i\in[n]} \left\| \boldsymbol{P}_{[\pm p]+i}\boldsymbol{x} \right\|_2 \left\| \boldsymbol{\beta} - \boldsymbol{\beta}' \right\|_2 =: L \left\| \boldsymbol{\beta} - \boldsymbol{\beta}' \right\|_2$$

define the set $\mathcal{N}_\varepsilon$ where Lipschitz constant is well bounded:

$$\mathcal{E}_{\text{Lip}} := \left\{ L \le \frac{3n\sqrt{p\theta\log n \log\theta^{-1}}}{M} \right\},$$

then on event $\mathcal{E}_{\text{Lip}}$, for every $\boldsymbol{\beta}$ in $\mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$, there exists some $\boldsymbol{\beta}_\varepsilon$ in $\mathcal{N}_{\varepsilon,i}$ with $\|\boldsymbol{\beta} - \boldsymbol{\beta}_\varepsilon\|_2 \le \varepsilon$, thus

$$\left| \left( \sum_{i\in[n]} \overline{\boldsymbol{X}}_i(\boldsymbol{\beta}) - n\mathbb{E}\overline{\boldsymbol{X}}(\boldsymbol{\beta}) \right) - \left( \sum_{i\in[n]} \overline{\boldsymbol{X}}_i(\boldsymbol{\beta}_\varepsilon) - n\mathbb{E}\overline{\boldsymbol{X}}(\boldsymbol{\beta}_\varepsilon) \right) \right| \le 2L\varepsilon \le \frac{2cn\theta^2}{p\log n}. \tag{I.34}$$

On event $\mathcal{E}_{\text{Lip}} \cap \mathcal{E}_{\text{Net}}$, from (I.31), (I.33) and (I.34), we can conclude that for every $\boldsymbol{\beta} \in \mathfrak{R}(\mathcal{S}_{\boldsymbol{\tau}}, \gamma(c_\mu))$ and $i \in [n]$,

$$\sum_{i\in[n]} \overline{\boldsymbol{X}}_i(\boldsymbol{\beta}) \le \frac{24cn\theta^2}{p\log n} \tag{I.35}$$

as desired, where the error probability of $\mathcal{E}^c_{\text{Lip}}$ is bounded using Lemma A.3, which gives

$$\mathbb{P}\left[\mathcal{E}^c_{\text{Lip}}\right] \le \mathbb{P}\left[ \max_{j\in[n]} \left\| \boldsymbol{x}_{[\pm p]+j} \right\|_2 > 3\sqrt{p\theta\log n\log\theta^{-1}} \right] \le 2/n, \tag{I.36}$$

4. (Bound $\mathbb{P}\left[\mathcal{E}^c_{\text{Net}}\right]$) Wlog let us assume that $2p$ divides $n$. By applying union bound and observing that $\overline{\boldsymbol{X}}_i(\boldsymbol{\beta})$ is independent of $\overline{\boldsymbol{X}}_{i+2p}(\boldsymbol{\beta})$ for any $i \in [n]$, we split $\sum_i \overline{\boldsymbol{X}}_i(\boldsymbol{\beta})$ into $n/2p$ independent sums of r.v.s, we have

$$\mathbb{P}\left[\mathcal{E}^c_{\text{Net}}\right] \le 2p\left|\mathcal{N}_\varepsilon\right| \cdot \mathbb{P}\left[ \sum_{j=0}^{n/2p-1} \left( \overline{\boldsymbol{X}}_{2pj}(\boldsymbol{\beta}) - \mathbb{E}\left[\overline{\boldsymbol{X}}(\boldsymbol{\beta})\right] \right) > \frac{9cn\theta^2}{p^2\log n} \right],$$

where each summand has bounded variance and $L^q$-norm derived similarly as its expectation such that

$$\mathbb{E}\overline{\boldsymbol{X}}_i(\boldsymbol{\beta})^q \le 2\cdot\mathbb{P}\left[\langle s_i[\boldsymbol{x}], \boldsymbol{\beta}\rangle \in [\lambda - B - M, \lambda + B + M]\right] \le 2\cdot\frac{1}{2\lambda}\cdot 2(B+M) \le \frac{4c\theta^2}{p\log n},$$

and apply Bernstein inequality Lemma J.4 with $(\sigma^2, R) = \left(4c\theta^2/\left(p\log n\right), 1\right)$, obtains

$$\mathbb{P}\left[ \sum_{j=0}^{n/2p-1} \left( \overline{\boldsymbol{X}}_{2pj}(\boldsymbol{\beta}) - \mathbb{E}\left[\overline{\boldsymbol{X}}(\boldsymbol{\beta})\right] \right) > \frac{9cn\theta^2}{p^2\log n} \right] \le \exp\left[ \frac{-(9cn\theta^2/p^2\log n)^2}{2cn\theta^2/p^2\log n + 2(9cn\theta^2/p^2\log n)} \right] \le \exp\left[ \frac{-4cn\theta^2}{p^2\log n} \right],$$

thus when $n = Cp^5\theta^{-2}\log p$:

$$\mathbb{P}\left[\mathcal{E}^c_{\text{Net}}\right] \le \exp\left[ \log(2p) + 13p\log\left(\frac{2p\log n}{c\cdot c_\lambda}\right) - \frac{4cn\theta^2}{p^2\log n} \right] \le 1/n \tag{I.37}$$

as long as $\frac{C}{\log C} > 10^5/\left(c^2\cdot c_\lambda\right)$.

■

# J Tools

**Lemma J.1** (Tail bound for Gaussian r.v.). *If $X \sim \mathcal{N}(0, \sigma^2)$, then its tail bound for $t > 0$ can be*

$$\mathbb{P}\left[X > t\right] \leq \frac{\sigma}{t\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{J.1}$$

**Lemma J.2** (Moments of the Gaussian random variables). *If $X \sim \mathcal{N}\left(0, \sigma^2\right)$, then ifor all integer $p \geq 1$,*

$$\mathbb{E}\left[|X|^p\right] \leq \sigma^p \left(p-1\right)!!. \tag{J.2}$$

**Lemma J.3** (Gaussian concentration inequality). *Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be a vector of $n$ independent standard normal variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be an $L$-Lipschitz function. Then for all $t > 0$,*

$$\mathbb{P}\left[|f(\boldsymbol{x}) - \mathbb{E}f(\boldsymbol{x})| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2L^2}\right). \tag{J.3}$$

**Lemma J.4** (Moment control Bernstein inequality for scalar r.v.s). ([FR13], Theorem 7.30) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent real-valued random variables. Suppose that there exist some positive number $R$ and $\sigma^2$ such that $\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\boldsymbol{X}_i^2\right] \leq \sigma^2$ and*

$$\tfrac{1}{n}\sum_{i=1}^n \mathbb{E}\left[|\boldsymbol{x}_k|^p\right] \leq \tfrac{1}{2}\sigma^2 R^{p-2} p!, \quad \text{for all integers } p \geq 3.$$

*Let $S \doteq \sum_{i=1}^n \boldsymbol{x}_i$, then for all $t > 0$, it holds that*

$$\mathbb{P}\left[|S - \mathbb{E}\left[S\right]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2n\sigma^2 + 2Rt}\right). \tag{J.4}$$

**Lemma J.5** ($\varepsilon$-net on sphere). [Ver10] *Let $(X, d)$ be a metric space and let $\varepsilon > 0$. A subset $\mathcal{N}_\varepsilon$ of $X$ is called an $\varepsilon$-net of $X$ if for every point $x \in X$ there exists some point $y \in \mathcal{N}_\varepsilon$ so that $d(x, y) \leq \varepsilon$. There exists an $\varepsilon$-net $\mathcal{N}_\varepsilon$ for the sphere $\mathbb{S}^{n-1}$ of size $|\mathcal{N}_\varepsilon| \leq (3/\varepsilon)^n$.*

**Lemma J.6** (Hanson-Wright). [RV+13] *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be independent, subgaussian random variables with sub-gaussian norm $\sup_{p\geq 1} p^{-1/2}\left(\mathbb{E}\,|x_i^p|\right)^{1/p} \leq \sigma$. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, then for every $t > 0$,*

$$\mathbb{P}\left[|\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} - \mathbb{E}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}| \geq t\right] \leq 2\exp\left(-c\min\left(\frac{t^2}{64\,\sigma^4\left\|\boldsymbol{A}\right\|_F^2}, \frac{t}{8\sqrt{2}\,\sigma^2\left\|\boldsymbol{A}\right\|_2}\right)\right). \tag{J.5}$$

**Lemma J.7** (Maximum of separable convex function). *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function of the form $f(x) = x - s(x)$ with $s : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying*

$$\frac{s(x)}{x} \leq \frac{s(y)}{y}, \quad \text{for all } x \geq y > 0.$$

*Then for $n \in \mathbb{N}$ and $0 < N \leq nL$,*

$$\max_{0\leq \boldsymbol{x}\leq L,\ \|\boldsymbol{x}\|_1\leq N} \sum_{i=1}^n f(\boldsymbol{x}_i) \leq N\left(1 - \frac{s(L)}{L}\right) \tag{J.6}$$

**Proof.** Since the feasible set is a convex polytope; the convex function $\sum_{i=1}^n f(\boldsymbol{x}_i)$ is maximized at a vertex, and that its vertices consist of $0$ and permutations of the vector $\big[\underbrace{L, \ldots, L}_{\lfloor N/L\rfloor}, r, 0, \ldots, 0\big]$, where $r = N - \lfloor N/L\rfloor L \leq L$. Then the function value at the maximizing vector $\boldsymbol{x}_*$ can be derived as:

$$\begin{aligned}\sum_{i=1}^n f(\boldsymbol{x}_{*i}) &= \left\lfloor \tfrac{N}{L}\right\rfloor f(L) + f(r) = \tfrac{N-r}{L}\left(L - s(L)\right) + \left(r - s(r)\right)\\ &= N\left(1 - \tfrac{s(L)}{L}\right) + r\left(\tfrac{s(L)}{L} - \tfrac{s(r)}{r}\right) \leq N\left(1 - \tfrac{s(L)}{L}\right)\end{aligned}$$

■