# Deconvolution Problems for Structured Sparse Signal 

Han-Wen Kuo

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy under the Executive Committee of the Graduate School of Arts and Sciences
(C) 2021

Han-Wen Kuo
All Rights Reserved

ABSTRACT<br>Deconvolution Problems for Structured Sparse Signal<br>Han-Wen Kuo

This dissertation studies deconvolution problems of structured sparse signals appears in nature, science and engineering. We discuss about the intrinsic solution to the problem of short-and-sparse deconvolution, how these solutions structured the optimization problem, and how do we design an efficient and practical algorithm base on aforementioned analytical findings. To fully utilized the information of structured sparse signals efficiently, we also propose a sensing method while the sampling acquisition is expansive, study its sample limit and algorithms for signal recovery with limited samples.

## Table of Contents

List of Figures ..... v
Acknowledgements ..... xii
1 Introduction ..... 1
1.1 Signals and data of short-and-sparse convolution model ..... 1
1.2 Mathematical formulation of short-and-sparse convolution ..... 3
1.3 Classifications of Short-and-Sparse Deconvolution problems ..... 5
1.4 Selected Topics Short-and-Sparse Deconvolution Problems ..... 7
1.4.1 Scanning electrochemical microscopy with line probe ..... 7
1.4.2 Deconvolution of both short-and-sparse ..... 7
1.5 Contribution of this thesis ..... 8
1.6 Outline ..... 9
2 Compressed Sensing Microscopy with Scanning Line Probe ..... 10
2.1 Introduction ..... 11
2.1.1 Contribution ..... 13
2.1.2 Related work ..... 13
2.2 Line scans measurement model ..... 14
2.2.1 Line projection ..... 14
2.2.2 Line scans ..... 15
2.3 Setup of SECM microscope ..... 16
2.3.1 $\quad$ Fabrication of Redox line probe ..... 16
2.3.2 Execution of a single CLP line scan ..... 17
2.3.3 Sample repositioning between successive scans ..... 17
2.4 Promises and problems of line scans ..... 18
2.4.1 Compressed Sensing of line projections for highly localized image ..... 20
2.4.2 Reconstructability from line projections of localized image in practice ..... 21
2.4.3 Obstacles of image reconstruction from line scans ..... 25
2.5 Reconstruction from line scans ..... 26
2.5.1 Sparse recovery with Lasso from line projections ..... 27
2.5.2 Computation of line projection ..... 29
2.5.3 Coping with nonidealities of practical line scans with reconstruction algorithm ..... 32
2.5.4 Image reconstruction algorithm from line scans ..... 35
2.6 Real data experiments ..... 37
2.7 Summary \& Discussion. ..... 39
3 Short-and-Sparse Deconvolution ..... 41
3.1 Introduction ..... 41
3.2 Formulation and Assumptions ..... 43
3.2.1 Nonconvex SaS over the Sphere ..... 43
3.2.2 Analysis Setting and Assumptions ..... 45
3.3 Main Results: Geometry and Algorithms ..... 47
3.3.1 Geometry of the Objective $\varphi_{\rho}$ ..... 47
3.3.2 Provable Algorithm for SaS Deconvolution ..... 51
3.3.3 Relationship to the Literature ..... 55
3.3.4 Notations ..... 57
3.4 Geometry of $\varphi_{\rho}$ in Shift Space ..... 58
3.4.1 Shifts and Correlations ..... 59
3.4.2 Shifts and the Calculus of $\varphi_{\ell^{1}}$ ..... 59
3.4.3 Any Local Minimizer is a Near Shift ..... 64
3.5 Provable Algorithm ..... 66
3.5.1 Minimization ..... 66
3.5.2 Local Refinement ..... 69
3.6 From Analysis to Practical Algorithm ..... 72
3.7 Experiments ..... 73
3.8 Summary \& Discussion ..... 74
4 Discussion ..... 77
Bibliography ..... 79
A Appendix: Compressed Sensing Microscopy with Scanning Line Probe ..... 88
A. 1 Analytic Derivations ..... 88
A.1.1 Proof of Theorem 2.4.1 ..... 88
A.1.2 Proof of Theorem 2.4.2 ..... 89
A.1.3 Proof of Theorem 2.4.3 ..... 91
A.1.4 Proof of Theorem 2.4.4 ..... 92
A.1.5 Proof of Theorem 2.5.1 ..... 93
B Appendix: Short-and-Sparse Deconvolution ..... 95
B. 1 Basic bounds for Bernoulli-Gaussian vectors ..... 95
B. 2 Vectors in shift space ..... 100
B. 3 Euclidean gradient as soft-thresholding in shift space ..... 107
B. 4 Euclidean Hessian as logic function in shift space ..... 115
B. 5 Geometric relation between $\rho$ and $\ell^{1}$-norm ..... 119
B. 6 Analysis of geometry ..... 125
B.6.1 Negative curvature ..... 127
B.6.2 Large gradient ..... 131
B.6.3 Convex near solutions ..... 136
B.6.4 Retraction toward subspace ..... 140
B.6.5 Proof of Theorem 3.4.1 ..... 143
B.7 Analysis of algorithm - minimization within widened subspace ..... 145
B.7.1 Initialization near subspace ..... 145
B.7.2 Minimization near subspace (Proof of Theorem 3.5.1) ..... 148
B. 8 Analysis of algorithm - local refinement ..... 153
B.8.1 $\quad$ Reweighted Lasso finds the large entries of $\boldsymbol{x}_{0}$ ..... 155
B.8.2 Least squares solution $\boldsymbol{a}^{(k)}$ contracts ..... 160
B.8.3 Linear convergence of alternating minimization (Proof of Theorem 3.5.2|) ..... 167
B.8.4 Supporting lemmas for refinement ..... 168
B. 9 Finite sample approximation ..... 172
B.9.1 Proof of/Theorem B.3.4 ..... 172
B.9.2 Proof of Theorem B.4.3 ..... 175
B.9.3 Proof of/Theorem B.5.5 ..... 177
B. 10 Tools ..... 180

## List of Figures

1.1 Signals and data under short-and-sparse convolutional model. In calcium imaging (top), the neuron fires at different time, leading the calcium(II) ion bound indicator to emit short pulses of fluorescence signal. For natural images (mid), the image is taken when the camera is shaken, causing a blurring artifact for the final image. And for transmission electronic microscopy (bot), an image of NaFeCoAs is taken with the brighter part are Co defects. All of these signals can be effectively modeled as convolution between short and sparse components.
2.1 Scanning electrochemical microscope with continuous line probe |OKL+18]. Left: the lab made SECM device with line probe, mounted on an automated probe arms with a rotating sample stage. Right: closeup side view of the line probe near the sample surface.11
2.2 Scanning procedure of SECM with continuous line electrode probe. The user begins with mounting the sample on a rotational stage of microscope and chooses $m$ scanning angles. The microscope then carries on sweeping the line probe across the sample, and measures the accumulated current generated between the interreaction of probe and the sample at equispaced intervals of moving distance. After a sweep ends, the sample is rotated to another scanning angle and the scanning sweep procedure repeats, until all $m$ line scans are finished. Collecting all scan lines, and providing the information of the scanning angles, the microscope system parameters (such as the point spread function) and the sparse representing basis of image, the final sample image is produced via computation with sparse reconstruction algorithm. 12
2.3 Mathematical expression of a single measurement from the line probe. When the stage rotate by $\theta$ clockwise, the relative rotation of probe to sample is counterclockwise by $\theta$. The grey line in the figure represents the rotated line probe, orienting in direction $\boldsymbol{u}_{\theta}=(\cos \theta, \sin \theta)$, and is sweeping in direction $\boldsymbol{u}_{\theta}^{\perp}=(\sin \theta,-\cos \theta)$. When it comes across the point $\boldsymbol{w}_{i}$ where $t=\left\langle\boldsymbol{u}_{\theta}^{\perp}, \boldsymbol{w}_{i}\right\rangle$, it integrates over the contact region $\ell_{\theta, t}$ between the probe and substate and produces a measurement $\boldsymbol{R}_{\theta}(t)$.
2.4 Side view schematic of SECM line probe [OKL ${ }^{+}$18|. Cross-sectional side-view of a CLP in contact with a sample surface. The angle of the CLP with respect to the sample surface $\theta_{\text {CLP }}$ and thickness of the bottom insulating layer of the probe $t_{I}$ determine the mean separation distance between the active sensing element and the sample surface $\left(d_{m}\right)$.
2.5 Schematic top views for rotation and translation scheme for CLP sample stage [DKS ${ }^{+}$19]. We illustrate example contains 3 electroactive disks. The scanning area of the previous scan $A_{N}$ is shaded in green, while the axis of rotation is located at ( $x_{r}, y_{r}$ ) and marked with a dot (left). After the stage is rotated with respect to $\left(x_{r}, y_{r}\right)$ by an angle $\theta_{s}$ (mid), then translate of the sample stage using the $X$ - and $Y$-motors to a new position corresponding to the start location for the next scan (right). The new scan area, $A_{N+1}$, for the next scan is shaded gray, while the black circle marks the common area of analysis for both scan angles, which contains all of the electroactive objects of interest. 18
2.6 Proof sketch for sufficiency of image recovery from three line projections. Given a sample with separated tiny discs $\boldsymbol{w}_{1}, \ldots \boldsymbol{w}_{k}$ (black dots), randomly choosing three lines projection forms lines $\widetilde{\boldsymbol{R}}_{1}, \widetilde{\boldsymbol{R}}_{2}, \widetilde{\boldsymbol{R}}_{3}$, in which all the discs after line projection (red dots) are well-separated. From each of these lines, we construct the dual $\boldsymbol{Q}$ as center of red dots, and a back projection image form the dual (dash lines), forming the set $\mathrm{\cup}_{j=1}^{k} \ell_{\theta_{i}, t_{j}}$. Intersection of three such line sets is exactly the set of ground truth disc centers. 21
2.7 Least eigenvalue of $\widetilde{G}$ with Gaussian motifs on hexagonal lattice. We show an example image of local features which are placed on the lattice locations (left), and calculate the least eigenvalue with varying number of motifs and distance-to-diameter ratio (right). When the motifs are highly overlapping $d / 2 r=0.5$, then $\widetilde{\boldsymbol{G}}$ is almost rank-deficient; when $d / 2 r \geq 1$, then $\widetilde{G}$ is stably full rank regardless of number of motifs. The result remains almost identical when the lattice is of other form such as rectangular grid, we therefore consider the distance-todiameter ratio is the determining factor for injectivity of line projections even in more general settings.
2.8 The point spread function of line probe. The PSF of line probe is skewed in the probe sweeping direction. We show an estimated PSF with close form used for reconstruction (left); and the software (LabVIEW) simulated PSF whose shape and intensity changes as the contacting angle varies (right). 25
2.9 Signal model of superposing electroactive species at different location. Left: An optical microscope view of a disc Right: the heatmap image of the substate $\boldsymbol{Y}$ is convolution between electroactive species $\boldsymbol{D}$ and its activation map $\boldsymbol{X}_{0}$.
2.10 Phase transition [OKL ${ }^{+}$18] of fixed image size (top) and fixed density (bot) on support recovery with Lasso. In each experiments, $d / 2 r \geq 1$ is ensured. In either cases, the phase transitions (left) show the number of samples required is almost linearly proportional to the number of discs for exact reconstruction. And the the advancement of scanning efficiency (right) is presented in comparison with the point probe scans. For the fixed size case, we let (image area)/(disc area) $\approx 1200$; for the fixed density case, we let density $\approx(1 / 6) \cdot(\max$ density). 28
2.11 Back projection image from the scan lines. We demonstrate a simple example (left) where four discs are line projected with angles $\left\{0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}\right\}$ then undergo convolution with the simulated PSF (mid). Here, the arrows indicates the probe sweeping direction. The back projection image (right) is the superposition of back projection image of each line; and the back projection of a single line $\boldsymbol{R}_{\theta}$ assigns value $\boldsymbol{R}_{\theta}(t)$ along the sweeping directions (arrows) onto the support $\ell_{\theta, t}$ for every $t$.
2.12 SECM image reconstruction with pure Lasso and reweighted Lasso. We apply three algorithm to reconstruct the image (left) with 6 line scans with simulated PSF in|Figure 2.8. The reconstruction from Lasso with large $\lambda$ (mid left) has unbalanced magnitude due to the coherence of line scans, and from Lasso with small $\lambda$ (mid right) gives blurry image by weakened sparsity regularizer. Reweighing Lasso can adjust the sparse regularizer in each iteration and consistently gives good result.
2.13 SECM image reconstruction with reweighed Lasso and reweighed calibrating Lasso. We simulate a line scan with uneven magnitude (left), and reconstruct the image (mid left) with two algorithm. The algorithm with reweighting only (mid right) cannot identify the correct support; where the reweighting plus calibration (right) method well approximates the image.
2.14 Performance of reweighting method versus Lasso. We use 8 line scans when the disc number is below 16, and 16 line scans when disc number is above for reconstruction. The experiments show reweighting method outperforms vanilla Lasso with various penalty variable $\lambda$ setting w.r.t. normalized (to 1) magnitude difference between the ground truth images and reconstructed images.
2.15 Real signal experiments on three platinum discs |DKS ${ }^{+}$19|. We show the reconstruction result of a three disc sample (up-left), which is scanned with line probe in seven different directions (up-right). The arrow in optical image represents the line probe sweeping direction, while as $\theta_{s}$ stands for clockwise rotation of the sample. The black circle indicates the correct disc location in each images. Compare to the point probe, in which the shifts of disc location are resulted from the skew of PSF (down-left), our line scan reconstruction accurately recovers the exact location (down-right). For both of the reconstructed images, the resolution is $10 \mu \mathrm{~m}$ per pixel. 38
2.16 Real signal experiments of $\mathbf{8 , 1 0}$ platinum discs. Showing the optimal image of the $\mathbf{8}$ discs (up) and 10 discs (down) sample, and their corresponding line scans, reconstructed image and reconstructed disc location map. In optical image, the arrows represent the line probe sweeping direction, while as $\theta_{s}$ stands for clockwise rotation of the sample. In both examples, our algorithm is able to successfully obtain these images of the discs, with most of the disc locations can be approximately represented by an one-sparse vector. Here, the image resolution is $20 \mu \mathrm{~m}$ per pixel.
3.1 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ and $s_{-\ell}\left[\boldsymbol{x}_{0}\right]$, 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.43
3.2 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 $\mid$ Theorem 3.3.1. which, loosely speaking, asserts that when $\theta \lesssim 1 /\left(p_{0} \sqrt{\mu}+\sqrt{p_{0}}\right)$, our method succeeds. 45
3.3 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. 47
3.4 Geometry of $\varphi_{\rho}$ near the span $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}\right\}}$ of two shifts of $\boldsymbol{a}_{0}$. Left: each pair of shifts $s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right]$, $s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]$ defines a linear subspace $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}\right\}}$ of $\mathbb{R}^{p}$. Center/right: every local minimum of $\varphi_{\rho}$ near $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}\right\}}$ (red line) is close to either $s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right]$ or $s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]$; there is a negative curvature in the middle of $s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right], s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]$, and $\varphi_{\rho}$ is convex in direction away from $\mathcal{S}_{\ell_{1}, \ell_{2}}$. 48
3.5 Geometry of $\varphi_{\rho}$ over the span $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}, \ell_{3}\right\}}$ of three shifts of $\boldsymbol{a}_{0}$. The subspace $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}, \ell_{3}\right\}}$ is threedimensional; 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 $s_{\ell_{i}}\left[\boldsymbol{a}_{0}\right]$, and are the only minimizers near $\mathcal{S}_{\ell_{1}, \ell_{2}, \ell_{3}}$.
3.6 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. 50
3.7 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ 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 $a_{0}$ (right).
3.8 Growth of $\varphi_{\rho}$ away from $\mathcal{S}_{\tau}$. Because $\varphi_{\rho}$ grows away from $\mathcal{S}_{\boldsymbol{\tau}}$, small-stepping descent methods stay near $\mathcal{S}_{\tau}$.53
3.9 Local minimization and refinement. Left: data-driven initialization $\boldsymbol{a}^{(0)}$ consisting of a nearsuperposition 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 $\boldsymbol{a}_{0}$. 54
3.10 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$, 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 $\boldsymbol{\beta}\left(\boldsymbol{a}^{+}\right)$after, one projected gradient step $\boldsymbol{a}^{+}=\boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\boldsymbol{a}-t \cdot \operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})\right]$. Notice that the small entries of $\boldsymbol{\beta}$ are shrunk towards zero. Right: the gradient $\operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})$ is large whenever it is easy to sparsify $\boldsymbol{\beta}$; in particular, when the largest entry $\boldsymbol{\beta}_{(0)} \gg \boldsymbol{\beta}_{(1)} \gg 0$.
3.11 Hessian Breaks Symmetry. Left: contribution of $-s_{i}\left[\boldsymbol{a}_{0}\right] s_{i}\left[\boldsymbol{a}_{0}\right]^{*}$ to the Euclidean hessian. If $\left|\boldsymbol{\beta}_{i}\right| \gg \lambda$ the Euclidean hessian exhibits a strong negative component in the $s_{i}\left[\boldsymbol{a}_{0}\right]$ direction. The Riemmanian hessian exhibits negative curvature in directions spanned by $s_{i}\left[\boldsymbol{a}_{0}\right]$ with corresponding $\left|\boldsymbol{\beta}_{i}\right| \gg \lambda$ and positive curvature in directions spanned by $s_{i}\left[\boldsymbol{a}_{0}\right]$ with $\left|\boldsymbol{\beta}_{i}\right| \ll$ $\lambda$. Middle: this creates negative curvature along the subspace $\mathcal{S}_{\tau}$ 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.
3.12 Success probability of SaS deconvolution under generic $a_{0}, x_{0}$ 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 $\mid$ Algorithm 7 works with high probability when $\theta \lesssim \frac{1}{\sqrt{p_{0}}}$ in generic case. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 75
B. $1 \quad$ 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 $2 p$-separated. Left: the yellow line is the function $\boldsymbol{\beta}_{i} \rightarrow \boldsymbol{\beta}_{i}\left(1-\operatorname{erf}_{\boldsymbol{\beta}_{i}}(\lambda, 0)\right)$ derived from (B.53), and the blue/red lines are its upper/lower bound (B.55) utilized in the analysis respectively. Right: functions of $\boldsymbol{\beta}_{i} \rightarrow \boldsymbol{\beta}_{i}\left(1-\operatorname{erf}_{\boldsymbol{\beta}_{i}}(\lambda, 0)\right)$ with different $\lambda$, the section of function of $\boldsymbol{\beta}_{i}>\nu_{2} \lambda$ are close to linear. 109
B. 2 A numerical example for $\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}$. We provide a figure to illustrate the expectation of $-\frac{1}{n \theta}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}$ when entries of $\boldsymbol{x}_{0}$ are $2 p$-separated, as a function plot of $\boldsymbol{\beta}_{i} \rightarrow 1-$ $\operatorname{erf}_{\boldsymbol{\beta}_{i}}(\lambda, 0)+f_{\boldsymbol{\beta}_{i}}(\lambda, 0)$ from $(\overline{\mathrm{B}} .86)$ with different $\lambda$. When $\left|\boldsymbol{\beta}_{i}\right| \approx \nu_{2} \lambda$ where $\nu_{2}=\sqrt{2 / \pi}$, then the its function value is close to 0.5 . If $\left|\boldsymbol{\beta}_{i}\right|$ is much larger then $\lambda$ its value grow to 1 , implies there is a negative curvature at $s_{i}\left[\boldsymbol{a}_{0}\right]$ direction. Similarly if $\left|\boldsymbol{\beta}_{\boldsymbol{i}}\right|$ is much smaller then $\lambda$ the function value is 0 thus the curvature is positive in $s_{i}\left[\boldsymbol{a}_{0}\right]$ direction.
B. 3 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 $s_{i}\left[\boldsymbol{a}_{0}\right]$ and $s_{j}\left[\boldsymbol{a}_{0}\right]$. When $\boldsymbol{a}$ lies near middle of two shifts (light green region) such that $\left|\boldsymbol{\beta}_{i}\right| \approx\left|\boldsymbol{\beta}_{j}\right|$, then there exists a negative curvature direction in subspace $\mathcal{S}_{\{i, j\}}$. When $a$ leans closer to one of the shifts $s_{i}\left[\boldsymbol{a}_{0}\right]$ (blue green region), its negative gradient direction points at that nearest shift. When $\boldsymbol{a}$ is in the neighborhood of the shift $s_{i}\left[\boldsymbol{a}_{0}\right]$ (dark green region) such that $\left|\boldsymbol{\beta}_{i}\right| \ll \lambda$, it will be strongly convex at $a$, and the unique minimizer within the convex region will be close to $s_{i}\left[\boldsymbol{a}_{0}\right]$. Finally, the negative gradient will be pointing back toward the subspace $\mathcal{S}_{\{i, j\}}$ if near boundary (grey region).
B. 4 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 $\boldsymbol{v}^{*} \operatorname{Hess}[\varphi](\boldsymbol{a}) \boldsymbol{v}$ is negative. In large gradient region, the norm of Riemannian gradient $\|\operatorname{grad}[\varphi](\boldsymbol{a})\|_{2}$ will be strictly greater then 0 and pointing at the nearest shift. Finally there is a convex region near all shifts such that Hess $[\varphi](\boldsymbol{a})$ is positive semidefinite.

## Acknowledgments

My first and foremost thanks go undoubtedly to my advisor Professor John Wright. It has been nothing short of amazement to work with a truly intellectual inspiring advisor, as well as a man with unquenchable enthusiasm for research and teaching. The joy to work on theoretical projects with John is simply beyond description, as he always seems to have a mind of his own-one seems to be the mystical combination of unbounded imagination and rigorous theoretical discipline. People say the pursuit in theoretical study is a journey of cold and lonely, as many find it difficult to share their state of mind amid compiling mathematical analysis. To my surprise and delight, I have been well accompanied in the path of theory development, as John can always spot the core of subject no matter how chaotic discussion was beginning with. Always keen and precise, and even more so humorous and funny, has made the theoretical research as well as my journey through Ph.D. exceptionally enriching and delightful at times.

I want to express my gratitude toward Professor Daniel Esposito, who has introduced John and I a truly exciting and novel microscopy project. I feel it is such a rare and precious chance to acquire experiences bridging theory and practice in signal processing and to work with one with truly open mind that always engaged with new ideas and be keen on subject outside from whose area of expertise. There has been no short of advises in academia on staying away from interdisciplinary research, as the joint work across different field is always uncertain and risky; nevertheless the reward can tremendous, which I found true during our collaboration. As the project proceeds, I find it increasingly enjoyable after I have cold more and developed deeper understanding on the chemistry side (hope they feel so likewise); realizing how my domain knowledge can be benefit to others, and more importantly delivering it comprehensively in a longterm collaborative project.

I also want to express my thanks to other members of thesis committee Professor Christine Hendon, Carlos Fernandez-Granda and Predrag Jelenkovic for sparing your already busy time and being patience for reviewing my thesis and attending the defense. Many of your past works and services in this area has inspired the idea in my thesis, as such this is to my greatest honor to have you serve in the board.

I am exceptionally fortunate to have had great colleagues and friends throughout the years of study. I would like to thank my fellow group members and collaborators, including Yuqian Zhang, Anna Ellisabeth Dorfi, Yenson Lau, Qing Qu, Robert Colgan, Sam Buchanan, Glen O'Neil, Cun Mu and everyone I engaged with on discussing my research topics. I want to give my thanks especially to Yuqian, Yenson and Anna. Yuqian and Yenson have worked tirelessly with me on the short-and-sparse deconvolution project, for which we had uncountable discussions as well as sharing of opinions. This tremendously help me on the research, and their idea is scattered across many places in this thesis. Eventually we are working on the deconvolution project together by sharing the work to every members' strength, and the experience of it is far beyond then mere research itself. Anna has been working mainly on the building the novel microscopic device while I have been in a supporting rule. Nevertheless, it is one of my unforgettable experience to provide helps and ideas, shaping the direction on how to build the device, working on the crosstalk between the hardware, software, the microscopic data, and eventually seeing the project bears fruit in first hand.

The department of electrical and engineering and data science institute also provides me with rich resource and even homy environment during my time of study, to the extent that I feel sometimes I'm taking advantage of them. All the staff members from department are always inviting for students and has provided me with great help. This has made my life in Columbia easy and comfy, and their effort certainly deserve more then a paragraph in acknowledgement in someone's thesis.

I must express my dearest gratitude all of my family members, who has providing me with continuing support and encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Finally, I want to quote one of my favorite words from T. S. Elliot in his critical essay on literacy tradition, as my final thanks to all contributor for this work: "The emotion of art is impersonal. And the poet cannot reach this impersonality without surrendering himself wholly to the work to be done. And he is not likely to know what is to be does unless he lives in what is not merely the present, but the present moment of the past, unless he is conscious, not of what is dead, but of what is already living"

Apr 30, 2021

## Chapter 1

## Introduction

### 1.1 Signals and data of short-and-sparse convolution model

Many signals and datasets in a wide range of areas, from natural science such as biomedical and astronomy [Lew98, Sah07], engineering science like microscopy and sound engineering [SCI75, Can76, CLC ${ }^{+}$17], to modern data in information era coming from time-series and natural images [Don81, WT14, CVR14], can be effectively modeled as superpositions of multiple translated copies of a basic pattern. Signals and data of this nature are mathematically described as the convolution of two components; where one component represents the repeating pattern whose size or length is comparably smaller than that of the signal, and the other component is regarded as the sparsely populated activation map which often stands for timestamp or location in the temporal or spatial axis. These type of signals and data are being addressed to be under short-and-sparse (SaS) convolutional model, and applications associated with information finding from these instances often can be effectively characterized as to identify either one or sometimes both of the convolving components.

In neural science, neuron produces a short pulse of action potential when excited, which is often referred to as the spikes, since the pulse often exhibits a characteristic rise and decay in potential, whereas these electronic spikes or its associated emission of chemical elements can be detected via electrodes or fluorescence indicators [BCOC14, $\left.\mathrm{ODB}^{+} 15\right]$. Neurons communicate to each other via these spikes, with its messages believed to be encoded by the firing rate and the temporal pattern $\left[\mathrm{SKS}^{+} 09\right]$. In order to resolve these cellular messages, scientists employs spike detection or spike sorting techniques for analysis of electronic signal activated by one or more neurons in brain. Here, the spike detection refers to the process of finding the


Figure 1.1: Signals and data under short-and-sparse convolutional model. In calcium imaging (top), the neuron fires at different time, leading the calcium(II) ion bound indicator to emit short pulses of fluorescence signal. For natural images (mid), the image is taken when the camera is shaken, causing a blurring artifact for the final image. And for transmission electronic microscopy (bot), an image of NaFeCoAs is taken with the brighter part are Co defects. All of these signals can be effectively modeled as convolution between short and sparse components.
frequency or temporal pattern by identifying the timestamps when the neuron is excited, while spike sorting is a methodology to sort out different neurons base on the variety of spikes shape recorded by the electrode. Since the spike patterns of each neurons are unique and reproducible overtime [Lew98], the detected signal follows short-and-sparse convolutional model, and the spike detection and sorting problem can be effectively cast as to deconvolve both of the convolving components.

In microscopic imaging of material and biomedical science, researchers are interested in visualizing microscale to nanoscale structure from crystalline solids such as semiconductors and superconductors, to the cellar structure of biological sample including proteins and viruses [BJT ${ }^{+} 07$, Rei13]. The superconductivity and semiconductivity of metal, metalloid, alloy and compound materials can be determined by its repeating impurities or doping patterns in the lattice structure [ZZW13, VJJW $\left.{ }^{+16}, \mathrm{CSL}^{+} 20\right]$, whose structure is usually investigated via studying its transmission electron microscopic (TEM) images [HHMN13]; similarly for many biological specimens formed by repeating pattern of molecules or cells, single-particle analysis is performed to enhance the resolution of these patterns under TEM imaging [LBC99, SNRS ${ }^{+}$08]. In these problems the microscopic image can also be regarded as the convolution between the repeating pattern and its occurrence location, so its pattern recognition problem is de facto a deconvolution problem.

Another common example is deblurring or super-resolution problem in image processing. Image deblurring is a process to remove the blurring artifact of an image taken with moving, defocused camera or
other similar phenomenons [HNO06], which has become a a basic but increasingly important application due to proliferation of digital camera and amateur photograpy. The blurring process can be regarded as the convolution between a blurring kernel and the original image, which in cases of natural image scenes are often assumed to have only a few of sharp edges. To identify the blurring kernel and thus to subsequently recover the image, people exploit this particular phenomenon of sharp edges by deblurring via the use of the gradient images, which is known to be sparse [BT09]. As such, the gradient of the blurred image is the convolution of blurring kernel and the original sparse gradient image, and the blurring kernel can be identified via deconvolution methods, which can used to produce the high-resolution image.

In Figure 1.1 we illustrates several concrete examples for short-and-sparse convolution model in practice. The first row shows an instance of luminous intensity of the calcium(II) ion bound indicator from the firing of a single neuron captured by a fluorescence microscope, which is effectively the convolution between the neuron firing pulse and its specific temporal pattern. In the second row, we show a blurred image capture by a shaking camera, virtually an outcome of convolution between the blurring kernel and the original sharp image. At last we demonstrate a TEM microscopic image of a superconductor NaFeCoAs with Co defects which exhibits repeating dumbbell pattern in the lattice. These surveyed examples can all be characterized as the convolution between a short (or small) pattern with a sparse activation map signal; and more importantly, deconvolution of these two convolving components more than often plays an essential role in their relating applications respectively.

In this introduction, we well first mathematically define the short and sparse convolutional model in Section 1.2, and base on the proposed signal model, discuss about some related application classes base on SaS signals in Section 1.3. Then in Section 1.4.1 and Section 1.4.2 we will present two examples for such applications, which will be elucidated in details in later chapters. Finally in Section 1.5 we will briefly highlight the major contributions of this thesis.

### 1.2 Mathematical formulation of short-and-sparse convolution

In continuum, a one-dimensional integrable signal $\boldsymbol{y}$ has intensity at $t$ where $t \in \mathbb{R}$ depending whether these integrable patterns $\boldsymbol{a}_{0}$ are activating at $t$. When the intensity is linear superpositions of the translated patterns base on an activation map $\boldsymbol{x}_{0}=\sum_{i=1}^{k} \alpha_{i} \boldsymbol{\delta}_{t_{i}}{ }^{1}$ as finite sum of Dirac measures, the intensity $\boldsymbol{y}(t)$ is

[^0]obtained as
\[

$$
\begin{align*}
\boldsymbol{y}(t) & =\left(\boldsymbol{a}_{0} * \boldsymbol{x}_{0}\right)(t)+\boldsymbol{n}(t)=\int_{-\infty}^{\infty} \boldsymbol{a}_{0}(t-s) \boldsymbol{x}_{0}(d s)+\boldsymbol{n}(t) \\
& =\sum_{i=1}^{k} \alpha_{i} \boldsymbol{a}_{0}\left(t-t_{i}\right)+\boldsymbol{n}(t) . \tag{1.1}
\end{align*}
$$
\]

where $(*)$ is known as the convolution operator, and additive noise $\boldsymbol{n}(t)$ model subject to different signal assumptions. Similarly, in two dimension, the intensity $\boldsymbol{Y}(x, y)$ can be calculated via

$$
\begin{align*}
\boldsymbol{Y}(x, y) & =\left(\boldsymbol{A}_{0} * \boldsymbol{X}_{0}\right)(x, y)+\boldsymbol{n}(x, y)=\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \boldsymbol{A}_{0}\left(x-x^{\prime}, y-y^{\prime}\right) \boldsymbol{X}_{0}\left(d x^{\prime}, d y^{\prime}\right)+\boldsymbol{n}(x, y) \\
& =\sum_{i=1}^{k} \alpha_{i} \boldsymbol{A}_{0}\left(x-x_{i}, y-y_{i}\right)+\boldsymbol{n}(x, y) \tag{1.2}
\end{align*}
$$

When a sampled or discretized data is considered, assuming the activation map is a sparse vector $\boldsymbol{x}_{0}=$ $\sum_{i=1}^{k} \alpha_{i} \boldsymbol{e}_{t_{i}}$, similarly we get

$$
\begin{equation*}
\boldsymbol{y}_{t}=\left(\boldsymbol{a}_{0} * \boldsymbol{x}_{0}\right)_{t}+\boldsymbol{n}_{t}=\sum_{j=-\infty}^{\infty} \boldsymbol{a}_{0 t-j} \boldsymbol{x}_{0 j}+\boldsymbol{n}_{t}=\sum_{i=1}^{k} \alpha_{i} \boldsymbol{a}_{0 t-t_{i}}+\boldsymbol{n}_{t} . \tag{1.3}
\end{equation*}
$$

The general objective of the short-and-sparse deconvolution (SaSD) problem can be realized as to resolve key properties of either or both of $\boldsymbol{a}_{0}$ and $\boldsymbol{x}_{0}$ given the signal $\boldsymbol{y}$ with some prior knowledges of $\boldsymbol{a}_{0}, \boldsymbol{x}_{0}$ and the noise model. Without any priors of either of the convolving component, even under noiseless and discrete cases, and knowing the $\boldsymbol{a}_{0}$ and $\boldsymbol{x}_{0}$ are short and sparse, the SaSD problem will be ill-posed. To be more specific, it is known that the "pattern" of solutions for (1.3) is non-unique [CVR14], and moreover in many cases the existence of solutions cannot be verified [BB98].

Fortunately in many applications, an educated guess or assumption on properties of either $\boldsymbol{a}_{0}$ or $\boldsymbol{x}_{0}$ could often be assured, thereby in general the SaSD problems come with various type of constraints inspired by prior knowledges on various associated problems. In the spike detection/sorting problem and single-particle analysis in neuroscience and biology, the pattern of the short $a_{0}$ can be reasonably assumed. The electronic signal shape of excited neurons spikes can often characterized by a sharp depolarization (rise), repolarization (decay), and refractory (recovery) phase $\left[\mathrm{PAF}^{+} 04\right]$ with various possible increase/decrease rates in each phases for different spike types [AG77], which provides a good assumption on modeling of the short pattern $\boldsymbol{a}_{0}$. For certain case in single-particle analysis, the shape of the target specimen can be reasonably estimated and used for the initial condition for the analysis algorithms |Jen10. Also in image deblurring scenarios, it is known that the aperture defocus problem is well related to the imaging physics, thereby adequate modeling
of blurring kernel can be applied to the deconvolution problems [ZN09].

### 1.3 Classifications of Short-and-Sparse Deconvolution problems

Depending on its associated applications, the SaSD problems can be classified into three categories base on the prior informations for the given signals:

Deconvolution with known short pattern. When the short pattern $\boldsymbol{a}_{0}$ in SaSD problem is known, it is effectively reduced to a sparse deconvolution problem, which can be regarded as a subproblem of sparse linear regression [BDE09] since the operation of convolution with a known $\boldsymbol{a}_{0}$ is linear. In the noiseless scenario, given $\boldsymbol{y}=\boldsymbol{a}_{0} * \boldsymbol{x}_{0}$ with $\boldsymbol{x}_{0}$ to be the only component being unknown, when $\boldsymbol{a}_{0}$ is an all-pass filter, then $\boldsymbol{x}_{0}$ can be retrieved via simply inverse the linear operation.

From convolution theory, the condition number of the linear operator $\left(\boldsymbol{a}_{0} * \cdot\right)$ is related to the Fourier spectrum of the kernel $\boldsymbol{a}_{0}$. In many real-world applications where $\boldsymbol{a}_{0}$ resembles a bell-shape function with slow rise and decay, which is a common response of physical systems or shape of particles appears in many detection problems. It's Fourier transform is close to the bell shape, meaning that $\boldsymbol{a}_{0}$ is low-pass, therefore in practice, direct linear inversion via $\boldsymbol{a}_{0}$ from given $\boldsymbol{y}$ will be highly sensitive to noise.

A common way to address the loss of information through low-pass filter in practical applications is to exploit the a priori knowledge of the target signal $\boldsymbol{x}_{0}$ we try to recover, in which a vast body of works has devoted to find various signal conditions sufficient conditions for the signal given cut-off frequency of $\boldsymbol{a}_{0}$ to be appropriately small. Characteristic examples includes when $\boldsymbol{x}_{0}$ is known to be consist of shiftinvariant components, then we can reasonably expect the signal recovery is achievable if $\boldsymbol{x}_{0}$ is appropriately preprocessed [EP10]; also when $\boldsymbol{x}_{0}$ is spatially sparse and sufficiently separated, this signal can then be efficiently recovered via $\ell_{1}$-minimization techniques [CFG14a].

The SaSD problem when $\boldsymbol{a}_{0}$ is known with sparse $\boldsymbol{x}_{0}$ can also being view through lens in sparse recovery perspective. If $a_{0}$ is randomly generated as entrywise i.i.d. Gaussian variables, then the convolution is invertible as long as $\boldsymbol{x}_{0}$ is sufficiently sparse [KMR14.

Deconvolution with partial knowledge of the short pattern As we noted, in common practices SaSD often arise from are the signal recovery problem with unknown linear system response and the detection problem of a structured signal. In these scenarios, an educated guess can usually be made base on either the understanding of the physical properties from the linear system, or the signature characteristic of the
structured signal. Such a priori knowledge of the short pattern $a_{0}$ is mathematically modeled for which recovery of $\boldsymbol{a}_{0}$ and $\boldsymbol{x}_{0}$ are utilized.

Perhaps the most common approach to incorporate the prior knowledge of the short pattern $\boldsymbol{a}_{0}$ into the SaSD problem is the parametric model in which the generating $a_{0}$ lies within a set of function which can be characterized by a few parameters, which can be regarded classical in image deblurring, where $\boldsymbol{a}_{0}$ is modeled base on the physical properties of imaging devices [CE17. This approach is especially popular on areas of scientific imaging, such as telescopic imaging [GSM06, DMWD08], microscopic imaging [PZBF+07], and medical imaging [MIPAV:https://mipav.cit.nih.gov/|. Common blurring kernel in these cases, for instance, can be characterized into linear motion blur ( $a_{0}(x)=\frac{x}{2 L}, x \in[-L, L]$ ); Gaussian or exponential blur $\left(\boldsymbol{a}_{0}(x)=K e^{-\frac{|x|}{s}}\right)$, and many other more complicated blur models depending on the system properties.

In the area of signal detection, the parametric method is likewise popular in spike sorting problems [LS00, Lew98, FZP17], in which the shape neuron spikes is estimated by various parametric model such as Gaussian interpolation or autoregressive model depending on the signal sampling modality of the neurons.

Deconvolution with unknown short pattern Without prior knowledge of the short pattern $\boldsymbol{a}_{0}$, the blind deconvolution becomes a subclass of bilinear inverse problem, which is known to be a problem without unique solution in general. An apparent example to the non-uniqueness is the scaling ambiguity, where given $\boldsymbol{y}=\boldsymbol{a}_{0} * \boldsymbol{x}_{0}$, then the convolution of another signal pair $\left(\frac{1}{\alpha} \boldsymbol{a}_{0}\right) *\left(\alpha \boldsymbol{x}_{0}\right)$ also generates $\boldsymbol{y}$. To make the problem more complex, even if we disregard the simple scaling ambiguity, without further constraining either of the convoluting components $\boldsymbol{a}_{0}$ and $x_{0}$ the problem is unsolvable, a good example can be found in [CM14.

As such, there are various way to constraint both $a_{0}$ and $x_{0}$ to which an efficient algorithm from deconvolution can be applied. Recently the popular topic for SaSD problem is to regard the short $a_{0}$ to be reside in a lower dimensional random subspace and $\boldsymbol{x}_{0}$ to be sparse [ARR14, KK17], which can be readily recovered via $\ell_{1}$-minimization technique. The randomness of $a_{0}$ ensures the deconvolution problem to have no further unidentifiably problem beyond scaling ambiguity hence exists algorithms to exactly solve SaSD, which has applications in communication where the subspace of $\boldsymbol{a}_{0}$ can be designed.

For SaSD in more general scenarios, however, such randomness assumption for the short signal is often too restrictive, hence a more general condition is preferred. In this thesis we will devote a section to discuss this problem, we will show when $\boldsymbol{a}_{0}$ is sufficiently non-smooth and $\boldsymbol{x}_{0}$ is sufficiently sparse, then SaSD can be solved with efficient algorithm.

### 1.4 Selected Topics Short-and-Sparse Deconvolution Problems

In this thesis we will focus on two problems of short and sparse deconvolution the author encountered when interacting with collaborators in different areas:

### 1.4.1 Scanning electrochemical microscopy with line probe

Scanning probe microscopy is a popular mean in nanotechnology for imaging the nanoscale phenomenon in fields of chemical, biological or material science when the the image resolution is required well beyond diffraction limit of traditional optical microscope. The conventional electrochemical microscope scans the subject with a point probe, and the image is scanned by adopting a point-by-point sampling method, measures the electrochemical reaction locally in each measurement. This can be tremendously time consuming especially when high-resolution imaging.

In many application, as aforementioned, the subject species is often well structured; in particular, we consider the case when the subject is consist of repeating small electrochemical reactive profile $\boldsymbol{A}$ at sparse locations with map $\boldsymbol{X}_{0}$. This special structure makes a more efficient scanning method beyond point probe possible.

Instead of the point probe, we study a novel non-localized probe, whose contacting surface between the probe and the subject is a straight line. We address it as the line probe. The overall scanning procedure can be described as follows: We first select a scanning angle $\theta$, orient the body of the probe along the direction $(\cos \theta, \sin \theta)$, and then sweeps the probe in the orthogonal direction $(\sin \theta,-\cos \theta)$. At each position $t(\sin s,-\cos s)$ along this sweeping direction, the probe produces a single measurement, which is generated along the enter length of the active sensing element of the line probe. This procedure is repeated many times with each time a different scanning angle is chosen, generates a couple of "line scans" instead of an image with conventional point probe.

Our study asks if scanning with line probe can effectively reconstruct $\boldsymbol{Y}=\boldsymbol{A} * \boldsymbol{X}_{0}$ with much fewer measurements versus the point probe, which effectively, is a sparse deconvolution problem with incomplete observation of $\boldsymbol{Y}$.

### 1.4.2 Deconvolution of both short-and-sparse

In practice, the actual motif $\boldsymbol{A}$ of sparse convoluting signal is commonly unknown, in which the deconvolution problem is often cast as identifying both $\boldsymbol{A}$ and $\boldsymbol{X}_{0}$ solely with $\boldsymbol{Y}$. We address it as the short-and-sparse
deconvolution. The main difficulty for deconvoluting short-and-sparse signal, is that there are multiple possible solution to the problem. Generally speaking, lacking of unique solution in an inverse problem inevitably makes the task highly convoluted, and even sometimes notoriously hard for designing efficient algorithm, let alone developing analytical understanding to the problem.

One good thing in short-and-sparse deconvolution, is that these multiple solutions can be completely characterized in two types: the scaling ambiguity and shift ambiguity:

- Scaling ambiguity: Let $\alpha>0$, then the convolution of the reversely scaled pairs ( $\alpha \boldsymbol{A}, \frac{1}{\alpha} \boldsymbol{X}_{0}$ ) generates the same observation, since $(\alpha \boldsymbol{A}) *\left(\frac{1}{\alpha} \boldsymbol{X}_{0}\right)=\boldsymbol{A} * \boldsymbol{X}_{0}$.
- Shift ambiguity: Let $\tau$ be integers, then the convolution of reversely shifted pairs ( $\left.s_{\tau}[\boldsymbol{A}], s_{-\tau}\left[\boldsymbol{X}_{0}\right]\right)$ also generates the same observation, by seeing $\left(s_{\tau}[\boldsymbol{A}]\right) *\left(s_{-\tau}\left[\boldsymbol{X}_{0}\right]\right)=\boldsymbol{A} * \boldsymbol{e}_{\tau} * \boldsymbol{e}_{-\tau} * \boldsymbol{X}_{0}=\boldsymbol{A} * \boldsymbol{X}_{0}$.

The existence of multiple discrete and separated solutions, implies when solving short-and-sparse deconvolution by means of optimization, the problem will inherently become highly non-convex. This non-convex objective landscape can be hard to fathom, reflecting the complexity of solving the inverse problem with multiple solutions. Nevertheless, by understanding the structure of the solutions (i.e the scaled/shifted copies of $\left(\boldsymbol{A}, \boldsymbol{X}_{0}\right)$ ), one can possibly grasp the geometric structure in the associated optimization problem, and even to design a more stable and efficient algorithm toward solving short-and-sparse deconvolution.

### 1.5 Contribution of this thesis

The thesis is mostly contributing to corroborate our theoretical understanding toward the two aforementioned problems. In particular, our main result can be summarized as follows:

In the scanning line probe microscopy, we show:

- When the profile $\boldsymbol{A}$ is infinitesimally small and well separated, merely three sweep of line scans are suffice to reconstruct $\boldsymbol{Y}=\boldsymbol{A} * \boldsymbol{X}_{0}$.
- When $\boldsymbol{A}$ is distributed uniformly and both number and radius of $\boldsymbol{A}$ is sufficiently (not trivially) small, finite number of line scan sweeps is an injective measurement restricted on the sparse location of $\boldsymbol{X}_{0}$.

In the short-and-sparse deconvolution, we show:

- When $\boldsymbol{A}$ is shift-incoherent or $\boldsymbol{X}_{0}$ is sufficiently sparse, then there exists an efficient method guarantees solving $\boldsymbol{Y}=\boldsymbol{A} * \boldsymbol{X}_{0}$ up to scaled-shift ambiguity when both $\boldsymbol{A}_{0}$ and $\boldsymbol{X}_{0}$ are unknown.
- We provide a geometric analysis of the objective landscape when solving short-and-sparse deconvolution via optimization method. This understanding of geometry helps us designing a more efficient algorithm in practice.


### 1.6 Outline

The rest of the thesis is organized as follows. In Chapter 2 we will introduce the scanning line probe procedure in mathematical terms, and provide some primitive theoretical result on how sparse deconvolution problem interact with line scans. In Chapter 3 we present the short-and-sparse deconvolution problem, in which we provide the theory and an efficient algorithm, exactly solving both $\left(\boldsymbol{A}, \boldsymbol{X}_{0}\right)$ when unknown. In the last section Chapter 4. we summarize future works for each of these projects, as well as related open problem of interest.

## Chapter 2

## Compressed Sensing Microscopy with

## Scanning Line Probe

In this chapter we study the compressed sensing microscopy with scanning line probe. In applications of scanning probe microscopy, images are acquired by raster scanning a point probe across a sample. Viewed from the perspective of compressed sensing (CS), this pointwise sampling scheme is inefficient, especially when the target image is structured. While replacing point measurements with delocalized, incoherent measurements has the potential to yield order-of-magnitude improvements in scan time, implementing the delocalized measurements of CS theory is challenging. In this paper we study a partially delocalized probe construction, in which the point probe is replaced with a continuous line, creating a sensor which essentially acquires line integrals of the target image. We show through simulations, rudimentary theoretical analysis, and experiments, that these line measurements can image sparse samples far more efficiently than traditional point measurements, provided the local features in the sample are enough separated. Despite this promise, practical reconstruction from line measurements poses additional difficulties: the measurements are partially coherent, and real measurements exhibit nonidealities. We show how to overcome these limitations using natural strategies (reweighting to cope with coherence, blind calibration for nonidealities), culminating in an end-to-end demonstration.


Figure 2.1: Scanning electrochemical microscope with continuous line probe $\mathbf{O K L}^{+}$18]. Left: the lab made SECM device with line probe, mounted on an automated probe arms with a rotating sample stage. Right: closeup side view of the line probe near the sample surface.

### 2.1 Introduction

Scanning probe microscopy (SPM) is a fundamental technique for imaging interactions between a probe and the sample of interest. Unlike traditional optical microscopy, the resolution achievable by SPM is not constrained by the diffraction limit, making SPM especially advantageous for nanoscale, or atomic level imaging, which has widespread applications in chemistry, biology and materials science [WR94]. Conventional implementations of SPM typically adopt a raster scanning strategy, which utilizes a probe with small and sharp tip, to form a pixelated heatmap image via point-by-point measurements from interactions between the probe tip and the surface. Despite their capability of nanoscale imaging, SPM with pointwise measurement is inherently slow, especially when scanning a large area or producing high-resolution images.

When the target signal is highly structured, compressed sensing (CS) [D ${ }^{+} 06$, CW08, FR17] suggests it is possible to design a data acquisition scheme in which the number of measurements is largely dependent on the signal complexity, instead of the signal size, from which the signal can be efficiently reconstructed algorithmically. In nanoscale microscopy, images are often spatially sparse and structured. CS theory suggests for such signals, localized measurements such as pointwise samples are inefficient. In contrast, delocalized, spatially spread measurements are better suited for reconstructing a sparse image.

Unfortunately, the dense (delocalized) sensing schemes suggested by CS theory (and used in other applications, e.g., [LDSP08, SBC ${ }^{+}$12,VRR11]) are challenging to implement in the settings of micro/nanoscale imaging. Motivated by these concerns, $\left[\mathrm{OKL}^{+} 18\right]$ introduced a new type of semilocalized probe, known as a line probe, which integrates the signal intensity along a straight line, and studied it in the context of a particular microscopy modality known as scanning electrochemical microscopy (SECM) [BFLZ80, BFP ${ }^{+}$91]. In SECM with line probe, the working end of the probe constitutes a straight line, produces a single measurement by


Figure 2.2: Scanning procedure of SECM with continuous line electrode probe. The user begins with mounting the sample on a rotational stage of microscope and chooses $m$ scanning angles. The microscope then carries on sweeping the line probe across the sample, and measures the accumulated current generated between the interreaction of probe and the sample at equispaced intervals of moving distance. After a sweep ends, the sample is rotated to another scanning angle and the scanning sweep procedure repeats, until all $m$ line scans are finished. Collecting all scan lines, and providing the information of the scanning angles, the microscope system parameters (such as the point spread function) and the sparse representing basis of image, the final sample image is produced via computation with sparse reconstruction algorithm.
collecting accumulated redox reaction current induced by the probe and sample. These line measurements are semilocalized, sample a spatially sparse image more efficiently than measurements from point probes, and "has an edge" on high resolution imaging since a thin and sharp line probe can be manufactured with ease. Moreover, experiments in $\left[\mathrm{OKL}^{+} 18\right]$ suggest that a combination of line probes and compressed sensing reconstruction could potentially yield order-of-magnitude reductions in imaging time for sparse samples.

Realizing the promise of line probes (both in SECM and in microscopy in general) demands a more careful study of the mathematical and algorithmic problems of image reconstruction from line scans. Because these measurements are structured, they deviate significantly from conventional CS theory, and basic questions such as the number of line scans required for accurate reconstruction are currently unanswered. Moreover, practical reconstruction from line scans requires modifications to accommodate nonidealities in the sensing system. In this paper, we will address both of these questions through rudimentary analysis and experiments, showing that if the local features are either small or separated, then stable image reconstruction from line scans is attainable.

In the following, we will first describe the scanning procedure and introduce the line scan model under mathematical context in Section 2.2 Later, Section 2.4 presents several important properties of line scans as a measurement model, including a rudimentary study of compressed sensing with line scans over spatially sparse image. Lastly Section 2.5, Section 2.6 are the algorithm and experiment sections, in which we elaborate the algorithm for image reconstruction from line scans, and the result on both the simulated and the actual SECM examples.

### 2.1.1 Contribution

- We expose the lowpass property of line scans, and with rudimentary analysis showing that the exact reconstruction of a sparse image is possible with only three line scans provided these features are well-separated.
- We show the difficulty of image reconstruction from line scan in practice, due to the high coherence of measured signal and inaccurate estimate of point-spread-function. Our reconstruction algorithm addresses above issues.
- Finally, we display the complete algorithm for image reconstruction of SECM with line scans, which includes an efficient algorithm for computation of line scans, showing our improvement of reconstruction result compares to [OKL $\left.{ }^{+} 18\right]$.


### 2.1.2 Related work

### 2.1.2.1 Compressed sensing tomography

Line measurements also arise in computational tomography (CT) imaging, a classical imaging modality which has long history and strong precedent in literature [Hou73, Kak79. Her09], with great variety of applications ranging from medical imaging to material science [WV ${ }^{+}$87, .Fra92, DLB $^{+}$05]. Classical CT reconstruction recovers an image from densely sampled line scans, by approximately solving an inverse problem [NDMD ${ }^{+98}$, SL74]. These methods do not incorporate the prior knowledge of the structure of the target image, and degrade sharply when only a few CT scans are available. Compressed sensing offers an attractive means of reducing the number of measurements needed for accurate CT image reconstruction, and has been employed in applications ranging from medical imaging to (cryogenic) electron transmission microscopy [CTL08, Mal13, $\left.\mathrm{GBVdB}^{+} 12, \mathrm{SHL}^{+} 11, \mathrm{LSMH13}, \mathrm{DNT}+17, \mathrm{BDD}^{+} 12, ~ \mathrm{NdLPL}^{+} 13\right]$. The dominant approach assumes that the target image is sparse in a Fourier or wavelet basis, and reconstructs it via $\ell^{1}$ minimization or related techniques. Images in SECM and related modalities typically exhibit much stronger structure: they often consist some number of small particles [ $\mathrm{DZC}^{+} 10$, BLK15], or other repeated motifs [CSL $\left.{ }^{+} 18\right]$. In this situation, CS is especially promising. On the other hand, as we will see below, understanding the interaction between line scans and spatially localized features demands that we move beyond conventional CS theory.

### 2.1.2.2 Mathematical theory of line scans: Radon transform and image super-resolution

The question of recoverability from line measurements is related to the theory of the Radon transform, which corresponds to a limiting situation in which line scans at every angle are available [Rad05, Cor63, Nat01]. The Radon transform is invertible, meaning perfect reconstruction is possible (albeit not stable) in this limiting situation. Due to the projection slice theorem [Hel10], the line projections are inherently lowpass, and so the line scan reconstruction problem is related to superresolution imaging [FREM04]. When the image of interest consists of sparse point sources, the image can be stably recovered from its low-frequency components, provided the point sources are sufficiently separated [CFG14b]. Similarly, we can hope to achieve stable recovery of localized features from line scans as long as the features are sufficiently separated.

### 2.2 Line scans measurement model

To implement line scans for SECM, a line probe Figure 2.1 is mounted on an automated arm which positions the probe onto the sample surface. The line scan signal is generated by placing this line probe in different places, and measuring the integrated current induced by the interaction between the line probe and the electroactive part of the sample. In a pragmatic scanning procedure Figure 2.2, the user will choose distinct scanning angles $\theta_{1}, \ldots, \theta_{m}$. For each angle $\theta$, the line probe is oriented in direction $\boldsymbol{u}_{\theta}=(\cos \theta, \sin \theta)$ and swept along the normal direction $\boldsymbol{u}_{\theta}^{\perp}=(\sin \theta,-\cos \theta)$. Each sweep of probe generates the projection of the target image along the probe direction $\boldsymbol{u}_{\theta}$; collecting these projections for each $\theta_{i}$, we obtain our complete set of measurements.

### 2.2.1 Line projection

To describe the scanning procedure more precisely, we begin with a mathematical idealization, in which the probe measures a line integral of the image. In this model, when the probe body is oriented in direction $\boldsymbol{u}_{\theta}$ at position $t$, we observe the integral of the image over $\ell_{\theta, t}:=\left\{\boldsymbol{w} \in \mathbb{R}^{2} \mid\left\langle\boldsymbol{u}_{\theta}^{\perp}, \boldsymbol{w}\right\rangle=t\right\}$ :

$$
\begin{align*}
\mathcal{L}_{\theta}[\boldsymbol{Y}](t) & :=\int_{\ell_{\theta, t}} \boldsymbol{Y}(\boldsymbol{w}) d \boldsymbol{w} \\
& =\int_{s} \boldsymbol{Y}\left(s \cdot \boldsymbol{u}_{\theta}+t \cdot \boldsymbol{u}_{\theta}^{\perp}\right) d s \tag{2.1}
\end{align*}
$$



Figure 2.3: Mathematical expression of a single measurement from the line probe. When the stage rotate by $\theta$ clockwise, the relative rotation of probe to sample is counterclockwise by $\theta$. The grey line in the figure represents the rotated line probe, orienting in direction $\boldsymbol{u}_{\theta}=(\cos \theta, \sin \theta)$, and is sweeping in direction $\boldsymbol{u}_{\theta}^{\perp}=(\sin \theta,-\cos \theta)$. When it comes across the point $\boldsymbol{w}_{i}$ where $t=\left\langle\boldsymbol{u}_{\theta}^{\perp}, \boldsymbol{w}_{i}\right\rangle$, it integrates over the contact region $\ell_{\theta, t}$ between the probe and substate and produces a measurement $\boldsymbol{R}_{\theta}(t)$.

Collecting these measurements for all $t$, we obtain a function $\mathcal{L}_{\theta}[\boldsymbol{Y}]$ which is the projection of the image along the direction $\boldsymbol{u}_{\theta}$. We refer to the operation $\mathcal{L}_{\theta}: L^{2}\left(\mathbb{R}^{2}\right) \rightarrow L^{2}(\mathbb{R})$ as a line projection. Combining projections in $m$ directions $\Theta=\left\{\theta_{i}\right\}_{i=1}^{m}$, we obtain an operator $\mathcal{L}_{\Theta}: L^{2}\left(\mathbb{R}^{2}\right) \rightarrow L^{2}(\mathbb{R} \times[m])$ :

$$
\begin{equation*}
\mathcal{L}_{\Theta}[\boldsymbol{Y}]:=\frac{1}{\sqrt{m}}\left[\mathcal{L}_{\theta_{1}}[\boldsymbol{Y}], \ldots, \mathcal{L}_{\theta_{m}}[\boldsymbol{Y}]\right] \tag{2.2}
\end{equation*}
$$

### 2.2.2 Line scans

In reality, it is not possible to fabricate an infinitely sharp line probe, and hence our measurements do not correspond to ideal line projections. The line probe has response in its normal direction, causing the blurring effect that can be modeled as convolution with point spread function $\psi$ along the sweeping direction. In SECM, $\psi$ is typically skewed with a long tail in the sweeping direction. Accounting for this effect is important, if we wish to obtain accurate reconstructions in practice, thereby in this more realistic model, our measurements $\widetilde{\boldsymbol{R}} \in L^{2}(\mathbb{R} \times[m])$ become

$$
\begin{align*}
\widetilde{\boldsymbol{R}} & =\frac{1}{\sqrt{m}}\left[\boldsymbol{\psi} * \mathcal{L}_{\theta_{1}}[\boldsymbol{Y}], \ldots, \boldsymbol{\psi} * \mathcal{L}_{\theta_{m}}[\boldsymbol{Y}]\right] \\
& =: \boldsymbol{\psi} * \mathcal{L}_{\Theta}[\boldsymbol{Y}] \tag{2.3}
\end{align*}
$$

This measurement consists of $m$ functions $\boldsymbol{\psi} * \mathcal{L}_{\theta_{i}}[\boldsymbol{Y}](t)$ of a single (real) variable $t$, which corresponds to the translation of the probe in the $\boldsymbol{u}_{\theta_{i}}^{\perp}$ direction. In practice, we do not measure this function at every $t$, but


Figure 2.4: Side view schematic of SECM line probe OKL $^{+}$18]. Cross-sectional side-view of a CLP in contact with a sample surface. The angle of the CLP with respect to the sample surface $\theta_{\mathrm{CLP}}$ and thickness of the bottom insulating layer of the probe $t_{I}$ determine the mean separation distance between the active sensing element and the sample surface $\left(d_{m}\right)$.
rather collect $n$ equispaced samples. Write the sampling operator as $\mathcal{S}: L^{2}[\mathbb{R}] \rightarrow \mathbb{R}^{n}$, then our discretized measurements $\boldsymbol{R}_{i}$ with scanning angle $\theta_{i}$ is defined as $\boldsymbol{R}_{i}=\mathcal{S}\left\{\widetilde{\boldsymbol{R}}_{i}\right\}$. Collect all $m$ discrete line scans, the final measurement $\boldsymbol{R} \in \mathbb{R}^{n \times m}$ is written as

$$
\begin{equation*}
\boldsymbol{R}=\left[\mathcal{S}\left\{\widetilde{\boldsymbol{R}}_{1}\right\}, \ldots, \mathcal{S}\left\{\widetilde{\boldsymbol{R}}_{m}\right\}\right]=: \mathcal{S}\{\widetilde{\boldsymbol{R}}\} \tag{2.4}
\end{equation*}
$$

Our task is to understand when and how we can reconstruct the target image $\boldsymbol{Y}$ from these samples.

### 2.3 Setup of SECM microscope

### 2.3.1 Fabrication of Redox line probe

A CLP is composed of three layers: an insulating substrate, an electroactive layer, and a thin insulating layer. As shown in Figure 2.4, the electroactive sensing element is sandwiched between the two insulating layers. The thicker of the two insulating layers serves as the probe substrate, while the thinner insulating layer serves as a spacer between the electroactive layer and the sample during imaging that sets the average probe-substrate separation distance $\left(d_{m}\right)$. The thickness of the electroactive layer $\left(t_{E}\right)$, sets the imaging resolution. The CLP also simultaneously senses features along the width of the probe. In contrast, an ultramicroelectrode (UME), or "point-probe", typically consists of a metal wire that has been sealed in glass and polished at the end to obtain an exposed disk-shaped sensing element that is surrounded by a glass ring.

The required to attain a quasi-steady state response is expected to decrease for nanoscale CLPs
CLPs were fabricated using a procedure similar to that described by Wehmeyer et al. for nanoband electrodes [WDW85]. First, $50 \mu \mathrm{~m}$ thick platinum foil was laminated to an insulating polycarbonate substrate using a two-part 5 min Araldite epoxy. In order to ensure a tight seal with minimal gaps between the platinum and the polycarbonate substrate, a vice was used to apply pressure uniformly for several hours while the epoxy cured. The top surface of the Pt foil was electrically insulated using Kapton tape (thickness $\approx 70 \mu m$ ). The end of the CLP was polished using $1 \mu \mathrm{~m}$ alumina lapping paper (McMaster-Carr), followed by 0.3 and $0.05 \mu \mathrm{~m}$ alumina slurries. Electrodes were characterized with optical microscopy and cyclic voltammetry employing the oxidation of $1 \mathrm{mM} \mathrm{K}_{4}\left[\mathrm{Fe}(\mathrm{CN})_{6}\right]$ as a redox probe.

### 2.3.2 Execution of a single CLP line scan

Before a line scan is carried out, the sample stage must be positioned such that (i) the center of the sample area to be imaged is aligned with the midpoint of the CLP, (ii) the distance from the CLP midpoint to the center of the imaging area is set to half of the desired scan length, and (iii) the sample has the proper rotational orientation with respect to the $X$-scan direction such that the CLP scan will occur at the user-specified scan angle $\theta_{s}$. After lowering the CLP using the $Z$-positioner, the line scan measurement begins by initiating potentiostatic control of the substrate and CLP potential, during which the CLP and substrate currents are measured as a function of time. Current measured during these chronoamperometry (CA) measurements is recorded after an initial hold time, typically $240 s$, which allows for dampening of transient signals from the CLP and/or substrate before imaging starts. Next, the CA data for the CLP are recorded and saved to the PC while the $X$-positioner is used to move the sample stage at the user specified step size and dwell time over the user specified scan distance.

### 2.3.3 Sample repositioning between successive scans

Once a line scan finishes, the $Z$-positioner lifts the CLP off of the substrate and the sample stage must be repositioned for the next scan to be measured at a new scan angle $\theta_{s}$. Figures Figure 2.5 illustrate the procedure used for repositioning the sample stage between scans. After the stage is rotated by the userspecified angle $\theta_{s}$ with respect to the rotational center $x_{r}, y_{r}$, the stage then translates in the $X-Y$ plane with the probe position remaining fixed in the $X-Y$ coordinate system. For every substrate position $(x, y)$, its newly translated position $T(x, y)$ is calculated by assigning the location of the rotational center $\left(x_{r}, y_{r}\right)$ of


Figure 2.5: Schematic top views for rotation and translation scheme for CLP sample stage [DKS ${ }^{+}$19]. We illustrate example contains 3 electroactive disks. The scanning area of the previous scan $A_{N}$ is shaded in green, while the axis of rotation is located at $\left(x_{r}, y_{r}\right)$ and marked with a dot (left). After the stage is rotated with respect to $\left(x_{r}, y_{r}\right)$ by an angle $\theta_{s}$ (mid), then translate of the sample stage using the $X$ - and $Y$-motors to a new position corresponding to the start location for the next scan (right). The new scan area, $A_{N+1}$, for the next scan is shaded gray, while the black circle marks the common area of analysis for both scan angles, which contains all of the electroactive objects of interest.
the stage and its rotational angle $\theta_{s}$ at the current scan, using the following equation:

$$
T(x, y)=\left(\begin{array}{cc}
\cos \theta_{s}-1 & \sin \theta_{s}  \tag{2.5}\\
-\sin \theta_{s} & \cos \theta_{s}-1
\end{array}\right)\binom{x_{r}}{y_{r}}+\binom{x_{r}}{y_{r}} .
$$

The translation automatically relocates the substrate to the scanning area $A_{N+1}$, which contains the identical inscribed circle as that of initial area $A_{1}$. The translation scheme allows us to perform a sequence of CLP scans without the need to position the stage rotational center right at the center of the substrate. This is important since the center of the area to be imaged can be located far from the axis of rotation for the rotational stage. Accurate translation of the stage between scans can be ensured as long as the (i) location of the rotation center of the stage $\left(x_{r}, y_{r}\right)$ relative to the bottom end of the probe is known and (ii) all the reactive species reside within the inscribed circle of scanning area $A_{1}$. A more detailed description of this positioning algorithm is presented in Algorithm 1.

### 2.4 Promises and problems of line scans

The line projections $\mathcal{L}_{\theta}$ measurements enjoy two major advantages as an imaging modality model: (i) comparing to the pointwise measurements, the line projections are more delocalized, hence can be more efficient while measuring a spatially sparse signal; and (ii) it is easier to build a sharp edge for the line probe

```
Algorithm 1 CLP-SECM Automatic Scanning Procedure
Require: Probe length \(L\), reactive part of sample lies within inscribed circle of scan area, scan angles \(\theta_{1}, \ldots, \theta_{k}\)
    where \(\theta_{1}=0^{\circ}\), and first center of rotation \(\left(x_{r}, y_{r}\right)^{(1)}\) as relative position to left end of probe plus \(\left(-\frac{L}{2}, \frac{L}{2}\right)\),
    for \(i=1, \ldots, k\) do
        Scan the sample from \(\left[\left(-\frac{L}{2}, \frac{L}{2}\right),\left(\frac{L}{2}, \frac{L}{2}\right)\right]\) to \(\left[\left(-\frac{L}{2},-\frac{L}{2}\right)\left(\frac{L}{2},-\frac{L}{2}\right)\right]\);
        if \(i=k\) then
            break;
        else
            1. Move the stage to where the scan starts, such that probe position is at \(\left[\left(-\frac{L}{2}, \frac{L}{2}\right),\left(\frac{L}{2}, \frac{L}{2}\right)\right]\);
            2. Rotate the stage by angle \(\theta_{\Delta}=-\theta_{i+1}+\theta_{i}\);
            3. Move the stage by \(\left(x_{\Delta}, y_{\Delta}\right)\) where
\[
\binom{x_{\Delta}}{y_{\Delta}}=\left(\begin{array}{cc}
\cos \theta_{\Delta}-1 & \sin \theta_{\Delta} \\
-\sin \theta_{\Delta} & \cos \theta_{\Delta}-1
\end{array}\right)\binom{x_{r}^{(i)}}{y_{r}^{(i)}} ;
\]
4. Get the new rotational center as \(\left(x_{r}, y_{r}\right)^{(i+1)} \leftarrow\left(x_{r}, y_{r}\right)^{(i)}+\left(x_{\Delta}, y_{\Delta}\right)\);
        end if
    end for
```

(even sharper then the diameter for tip of point probe), which is amenable to detect the ultra-high frequency components in the probe sweeping direction. It makes possible fast and high resolution imaging for scanning microscopes.

Nevertheless, the line projection comes with a few apparent disadvantages as a means for imaging. Supposedly, if we uses infinitely many line projection at every angles in $[0,2 \pi)$ as measurements, then projection slice theorem suggests such measurement procedure is invertible, from which the image can always be perfectly reconstructed. However, this imaging modality is not stable, since when viewing in Fourier domain, the line projections are lowpass. This means even if a single line projection can be highly sensitive to the directional high frequency components, the cumulative line projections is not. Thus, in order to stably reconstruct an image consist of multiple localized features, it is required for the localized features in image to be either sufficiently small or separated.

The other disadvantage of line projections can be viewed from the CS perspective, that the line measurements are not incoherent to the sparse signal representing bases-even if the local features are well separated. This means in practice, when using only a few line scans for reconstruction, the number of lines required for exact reconstruction with line scans can not be reasoned via conventional CS theory. More importantly, the coherence of line projections could become a cause for concern during algorithmic reconstruction; it leads the reconstructed image to have incorrect magnitudes via conventional method.

Lastly, as with most of the imaging systems, deblurring from the effect of PSF $\boldsymbol{\psi}$ has always been an
important and fundamental task for imaging algorithm. We will demonstrate some example for showing difficulties on modeling PSF in SECM system with line probe. Later on in the next section, we provide an algorithmic solution addressing both issues from the coherence of line projections and incomplete information of PSF.

### 2.4.1 Compressed Sensing of line projections for highly localized image

Compressed sensing, in its simplest form, asserts that if the target signal has sparse representation, then only a few measurements that are incoherent to the representing basis would suffice for exact reconstruction. Since in many cases of microscopic imaging the underlying signal is structured and spatially localized, CS theory suggests the delocalized measurements, such as line projections, is more preferable than point measurements for more efficient scanning speed.

Assimilating ideas from $C S$, we study the conditioning of the line projection $\mathcal{L}_{\Theta}$ when it is restricted to an image with sparsely populated motifs $D \in L^{2}\left(\mathbb{R}^{2}\right)$. First, we demonstrate the following rudimentary analysis, providing sufficient conditions for to be recovery of such images from line measurements via total variation minimization [KKS17]:

Proposition 2.4.1. [Certificate of TV-norm minimization] Let $\boldsymbol{X}_{0}=\sum_{\boldsymbol{w} \in \mathcal{W}} \alpha_{\boldsymbol{w}} \boldsymbol{\delta}_{\boldsymbol{w}}{ }^{1}$ with $|\mathcal{W}|<\infty$. Given continuous compactly supported circular symmetric $\boldsymbol{D} \in L^{2}\left(\mathbb{R}^{2}\right)$, scanning angles $\Theta=\left\{\theta_{1}, \ldots, \theta_{m}\right\}$ and measurement $\widetilde{\boldsymbol{R}}=\mathcal{L}_{\Theta}\left[\boldsymbol{D} * \boldsymbol{X}_{0}\right]$. Suppose there exists $\widetilde{\boldsymbol{Q}}$ as finite sum of weighed Diracs such that

$$
\begin{cases}\boldsymbol{D} * \mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{Q}}](\boldsymbol{w})=\operatorname{sign}\left(\alpha_{\boldsymbol{w}}\right), & \boldsymbol{w} \in \mathcal{W}  \tag{2.6}\\ \left|\boldsymbol{D} * \mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{Q}}](\boldsymbol{w})\right|<1, & \boldsymbol{w} \notin \mathcal{W}\end{cases}
$$

If the Gram matrix $\boldsymbol{G} \in \mathbb{R}^{|\mathcal{W}| \times|\mathcal{W}|}$, defined as

$$
\begin{equation*}
\boldsymbol{G}_{i j}=\left\langle\mathcal{L}_{\Theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{i}}\right], \mathcal{L}_{\Theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\right\rangle, \boldsymbol{w}_{i}, \boldsymbol{w}_{j} \in \mathcal{W} \tag{2.7}
\end{equation*}
$$

is positive definite, then $\boldsymbol{X}_{0}$ is the unique optimal solution to

$$
\begin{equation*}
\min _{\boldsymbol{X} \in \operatorname{BV}\left(\mathbb{R}^{2}\right)} \int_{\boldsymbol{w}}|\boldsymbol{X}|(d \boldsymbol{w}) \quad \text { s.t. } \quad \widetilde{\boldsymbol{R}}=\mathcal{L}_{\Theta}[\boldsymbol{D} * \boldsymbol{X}] \tag{2.8}
\end{equation*}
$$

Proof. See Appendix A.1.1.

[^1]

Figure 2.6: Proof sketch for sufficiency of image recovery from three line projections. Given a sample with separated tiny discs $\boldsymbol{w}_{1}, \ldots \boldsymbol{w}_{k}$ (black dots), randomly choosing three lines projection forms lines $\widetilde{\boldsymbol{R}}_{1}, \widetilde{\boldsymbol{R}}_{2}, \widetilde{\boldsymbol{R}}_{3}$, in which all the discs after line projection (red dots) are well-separated. From each of these lines, we construct the dual $\widetilde{\boldsymbol{Q}}$ as center of red dots, and a back projection image form the dual (dash lines), forming the set $\cup_{j=1}^{k} \ell_{\theta_{i}, t_{j}}$. Intersection of three such line sets is exactly the set of ground truth disc centers.

Specifically, when the signal image is highly spatially sparse and its components are well separated, the line projections can be a very efficient measurement model. A concrete example is demonstrated in Theorem 2.4.2, where we assume the sparse component of the image signal are small and separated discs; if the radius of the discs are sufficiently small, then, perhaps surprisingly, only three line projections is required to exactly reconstruct the image via efficient algorithm.

Lemma 2.4.2. [Reconstruction from three line projection] Consider an image consists of $k \geq 2$ discs radius $r$. If the centers $\boldsymbol{w}_{1}, \ldots \boldsymbol{w}_{k}$ are at least separated by $\frac{2}{C} k^{2} r$, then three continuous line projections with probe direction chosen independent uniformly at random suffice to recover the image with probability at least $1-C$ via solving (2.8).

Proof. See Appendix A.1.4
The proof idea can be depicted pictorially in Figure 2.6 in which we show the construction of dual certificate $\widetilde{\boldsymbol{Q}}$, and the back projection operation $\mathcal{L}_{\ominus}^{*}$ on $\widetilde{\boldsymbol{Q}}$ which we used in the proof to certify the optimality. In fact, as we will show later, the operation $\mathcal{L}_{\Theta}^{*}$ is the cornerstone for most of the reconstruction algorithms in computed tomography, as well as in our sparse reconstruction algorithm.

### 2.4.2 Reconstructability from line projections of localized image in practice

While the microscopic images are often sparse in spatial domain, they rarely satisfy the conditions of Theorem 2.4.2, in which the local features are uncharacteristically small and far apart. In the following, we
will show in practical application of line scans, when the image consists of multiple localized motifs, the performance of line measurements degrades once the ratio between the size of motifs and its separating distance increases.

### 2.4.2.1 Coherence of line projection of two localized motif

We start from a simple case considering an image with two motifs located at different locations. Define a $2 \times 2$ Gram matrix $G$ with its $i j$-th entries being coherence [DET06] between line projected signal of two motifs $\boldsymbol{D}$ with center at $\boldsymbol{w}_{i}$ and $\boldsymbol{w}_{j}$ respectively,

$$
\begin{equation*}
\boldsymbol{G}_{i j}=\left\langle\mathcal{L}_{\Theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{i}}\right], \mathcal{L}_{\Theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\right\rangle \tag{2.9}
\end{equation*}
$$

If the off-diagonal entry $\boldsymbol{G}_{i j}$ is small in magnitude compared to the diagonal entries $\boldsymbol{G}_{i i}, \boldsymbol{G}_{j j}$, then it suffices to reconstruct the image exactly with efficient algorithm. Conversely, if $\boldsymbol{G}$ is ill-conditioned or even rankdeficient, then exact recovery will be impossible.

Lemma 2.4.3. [Coherence of line projection Gaussians] Let $\boldsymbol{D}$ be the two-dimensional Gaussian functions with covariance $r \boldsymbol{I}^{2}$ and normalized in a sense that $\left\|\mathcal{L}_{0}[\boldsymbol{D}]\right\|_{L^{2}}=1$. If $\theta$ is uniformly random, then the expectation of inner product between two line projected $\boldsymbol{D}$ at different locations $\boldsymbol{w}_{i}, \boldsymbol{w}_{j}$ is bounded by

$$
\begin{equation*}
\left(1-\frac{d^{2}}{8 r^{2}}\right) \mathbf{1}_{\{d \leq 2 r\}}+\frac{r}{2 d} \mathbf{1}_{\{d>2 r\}} \leq \mathbb{E}_{\theta}\left\langle\mathcal{L}_{\theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{i}}\right], \mathcal{L}_{\theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\right\rangle \leq \frac{1}{\sqrt{1+d^{2} / 4 r^{2}}} \tag{2.10}
\end{equation*}
$$

where $d=\left\|\boldsymbol{w}_{i}-\boldsymbol{w}_{j}\right\|_{2}$.

Proof. See Appendix A.1.3.
Theorem 2.4.3 shows the coherence between line projections of two bell-shaped motif with radius $\approx r$ and center distance $d$ is dominated by the distance-to-diameter ratio $d / 2 r$. Because of the projection slice theorem, the matrix $\mathbb{E}_{\theta} \boldsymbol{G}$ is always positive definitive. However, its condition number greatly increases when the image consists of highly overlapping local features. When the ratio is small, say $d / 2 r<1$, in which the two projected motifs are overlapping, then $\mathbb{E}_{\theta} \boldsymbol{G}_{i j}$ will be close to one as with the diagonals, implies $\mathbb{E}_{\theta} \boldsymbol{G}$ become severely ill-conditioned even in the two-sparse case. Generally speaking, line scans are not CS-theoretical optimal sampling method for sparse recovery for image of superposing discs.


Figure 2.7: Least eigenvalue of $\widetilde{\boldsymbol{G}}$ with Gaussian motifs on hexagonal lattice. We show an example image of local features which are placed on the lattice locations (left), and calculate the least eigenvalue with varying number of motifs and distance-to-diameter ratio (right). When the motifs are highly overlapping $d / 2 r=0.5$, then $\widetilde{\boldsymbol{G}}$ is almost rank-deficient; when $d / 2 r \geq 1$, then $\widetilde{\boldsymbol{G}}$ is stably full rank regardless of number of motifs. The result remains almost identical when the lattice is of other form such as rectangular grid, we therefore consider the distance-to-diameter ratio is the determining factor for injectivity of line projections even in more general settings.

### 2.4.2.2 Injectivity of line projection of multiple motifs with minimum separation

To extend the study of the coherence of matrix $\boldsymbol{G}$ to samples that contain $k>2$ motifs $\boldsymbol{D}$. We first investigate a model configuration whose motif centers are allocated on a hexagonal lattice with edges of length $d$. It turns out that the smallest eigenvalue of an approximation $\boldsymbol{G}$ with respect to the locations $\left\{\boldsymbol{w}_{1}, \ldots, \boldsymbol{w}_{k}\right\}$ is largely determined by the distance-to-diameter ratio $d / 2 r$, and depends only weakly on the total number of motifs.

In Figure 2.7. we calculate an approximation of $\mathbb{E}_{\theta} \boldsymbol{G}$ with $\widetilde{\boldsymbol{G}}$, where

$$
\begin{equation*}
\widetilde{\boldsymbol{G}}_{i j}=\left(1+\left\|\boldsymbol{w}_{i}-\boldsymbol{w}_{j}\right\|_{2}^{2} / 4 r^{2}\right)^{-1 / 2} \tag{2.11}
\end{equation*}
$$

is obtained from the upper bound in Theorem 2.4.3 with motifs being the Gaussian function of deviation $r$ placed on hexagonal lattice. We show that when these motifs are highly overlapping with distance-to-diameter ratio $d / 2 r=0.5$, the least eigenvalue of $\widetilde{G}$ is very close to zero and the matrix is nearly rank-deficient; when the motifs are separated, say $d / 2 r \geq 1$, the least eigenvalue of $\widetilde{\boldsymbol{G}}$ is steadily larger then zero and approaches one as the ratio $d / 2 r$ increases. Interestingly, in our experiments the least eigenvalue does not depend strongly on the number of motifs, suggesting that the distance-to-diameter ratio is the dominant factor for injectivity
of line projections on motifs with hexagonal placement ${ }^{2}$. Since the hexagonal configuration is the densest circle packing on a plane [Tót14], we suspect that $\lambda_{\min }\left(\mathbb{E}_{\theta} \boldsymbol{G}\right)$ is also determined by the ratio $d / 2 r$ for every configurations satisfying the minimum separation property.

This conjecture gains more ground when viewing this problem from the point source super-resolution perspective [CFG14b]. It is known that an image consisting of point measures $\boldsymbol{x}=\sum_{i} \alpha_{i} \boldsymbol{\delta}_{\boldsymbol{w}_{i}}$ can be stably recovered from its low frequency information (with frequency cutoff $f_{c}$ ) whenever the point sources have minimum separation $d>C / f_{c}$ for some constant $C$, regardless of the number of such point measures in $\boldsymbol{x}$. In our scenario, we will show that the expected line projection $\mathbb{E}_{\theta} \mathcal{L}_{\theta}^{*} \mathcal{L}_{\theta}$ is also a low-pass filter; and since the local features $\boldsymbol{D}$ is also often consists of low frequency components, our line projections $\mathcal{L}_{\Theta}[\boldsymbol{D} * \boldsymbol{X}]$ can be modeled as the low-pass measurements from sparse map $\boldsymbol{X}$, implying stable and efficient sparse reconstruction is possible as long as $\boldsymbol{X}$ is enough separated under infinitely many line measurements of all angles.

Lemma 2.4.4. [Lowpass property of line projections] Suppose $\boldsymbol{D}$ is two-dimensional Gaussian of covariance $r^{2} \boldsymbol{I}$ with $\left\|\mathcal{L}_{0}[\boldsymbol{D}]\right\|_{L^{2}}=1$ and $\boldsymbol{X}$ is finite summation of Dirac measure. If $\theta$ is uniformly random, then $\mathbb{E}_{\theta} \boldsymbol{D} * \mathcal{L}_{\theta}^{*} \mathcal{L}_{\theta}[\boldsymbol{D} * \boldsymbol{X}]$ is a low-pass filter $\mathcal{K}$ on $\boldsymbol{X}$ with cut-off frequency $f_{c}$ satisfies

$$
\begin{equation*}
f_{c}=\frac{1}{r} \cdot \min \left\{2 r^{2} \varepsilon^{-1}, \sqrt{\left|\log \left(8 r^{2} \varepsilon^{-1}\right)\right|}+0.2\right\} \tag{2.12}
\end{equation*}
$$

in a sense that $\max _{\|\boldsymbol{\xi}\|_{2} \geq f_{c}}\left|\mathcal{F}_{2}\{\mathcal{K}\}(\boldsymbol{\xi})\right| \leq \varepsilon$.

Proof. See Appendix A.1.4

Remark 2.4.5. When radius of motif is sufficiently large, then the cut-off frequency $f_{c}$ is dominated by the cut-off frequency of motif, roughly $C / r$, and is sufficient to recover its locations as long as the separation $d$ satisfies $d>C^{\prime} r$ (reflects the observation of Figure 2.7). In cases with small (pointy) $\boldsymbol{D}$, the cut-off frequency is mainly determined by the low-pass property of line projection, which requires minimum separation $d>C \varepsilon / r$ for exact reconstruction.

Finally, base on [CFG14b], when the separation condition is ensured, the image of separated discs can be recovered from infinitely many line projections via total variation minimization (or $\ell^{1}$ when $\boldsymbol{X}_{0}$ on discrete grid), regardless of number of discs.

[^2]

Figure 2.8: The point spread function of line probe. The PSF of line probe is skewed in the probe sweeping direction. We show an estimated PSF with close form used for reconstruction (left); and the software (LabVIEW) simulated PSF whose shape and intensity changes as the contacting angle varies (right).

### 2.4.3 Obstacles of image reconstruction from line scans

Besides the apparent nonideality of coherence of line scan measurements which is not CS theoretical optimal, this specific sampling method and its corresponding hardware limitations causes other practical nuisances during image reconstruction.

High coherence of line scans To show the coherence is a cause for concern, we rewrite the linear operator $\mathcal{L}_{\Theta}[\boldsymbol{D} * \cdot]$ as $\boldsymbol{A}$, and consider the nonnegative Lasso

$$
\begin{equation*}
\min _{\boldsymbol{X} \geq 0} \lambda\|\boldsymbol{X}\|_{1}+\frac{1}{2}\|\boldsymbol{A}[\boldsymbol{X}]-\boldsymbol{R}\|_{2}^{2} \tag{2.13}
\end{equation*}
$$

using the observed signal $\boldsymbol{R}=\boldsymbol{A}\left[\boldsymbol{X}_{0}\right]$ and linear, column normalized and coherent sampling method $\boldsymbol{A}$. Denote $\Omega$ as the support set of solution of (2.13), write $\boldsymbol{A}_{\Omega}$ as the submatrix of $\boldsymbol{A}$ restricted on columns of support $\Omega$, the unique solution $\boldsymbol{X}$ of program (2.13) (provided if $\boldsymbol{A}_{\Omega}$ is injective) can be written as

$$
\begin{cases}\boldsymbol{X}_{i j}=\left[\boldsymbol{X}_{0 i j}-\lambda\left(\boldsymbol{A}_{\Omega}^{*} \boldsymbol{A}_{\Omega}\right)^{-1} \mathbf{1}\right]_{+} & \boldsymbol{w}_{i j} \in \Omega  \tag{2.14}\\ \boldsymbol{X}_{i j}=0 & \boldsymbol{w}_{i j} \notin \Omega\end{cases}
$$

When $\boldsymbol{A}$ is coherent, columns of $\boldsymbol{A}$ have large inner product, implies many entries of the matrix $\boldsymbol{A}_{\Omega}^{*} \boldsymbol{A}_{\Omega}$ have large, positive off-diagonal entries close to its diagonals. When the sparse penalty $\lambda$ is large in (2.13), its solution will have incorrect relative magnitudes since $\boldsymbol{A}_{\Omega}^{*} \boldsymbol{A}_{\Omega}$ is not close to identity matrix as conventional CS measurements [CT05]. When $\lambda$ is small, the solution of program will be highly sensitive to noise, occasionally


Figure 2.9: Signal model of superposing electroactive species at different location. Left: An optical microscope view of a disc Right: the heatmap image of the substate $\boldsymbol{Y}$ is convolution between electroactive species $\boldsymbol{D}$ and its activation map $\boldsymbol{X}_{0}$.
lead to incorrect results.

Incomplete information of PSF of line scans Another layer of complexity for line probe scans is the difficulty to correctly identify its PSF due to hardware limitations, especially when operating line scans in nanoscale. For instance in Figure 2.8, we show if the contacting angle between the probe and the sample varies, the corresponding PSF changes drastically in both the peak magnitude and the shape. It turns out that even with seemingly small changes of probe condition, the corresponding PSF can be inevitably variated.

### 2.5 Reconstruction from line scans

In this section, we will introduced the algorithm for SECM image reconstruction with line scans. In all following experiments, we consider a representative class of images $\boldsymbol{Y}$ characterized by superposing reactive species $\boldsymbol{D}$ at locations $\mathcal{W}=\left\{\boldsymbol{w}_{1}, \ldots, \boldsymbol{w}_{|\mathcal{W}|}\right\} \subset \mathbb{R}^{2}$ with intensities $\left\{\alpha_{1}, \ldots, \alpha_{|\mathcal{W}|}\right\} \subset \mathbb{R}_{+}$. Define the activation map $\boldsymbol{X}_{0}$ as sum of Dirac measure at $\mathcal{W}$, then $\boldsymbol{Y}$ can simply be written as convolution between $\boldsymbol{D}$ and $\boldsymbol{X}_{0}$ :

$$
\begin{equation*}
\boldsymbol{Y}=\boldsymbol{D} * \boldsymbol{X}_{0}=\sum_{j=1}^{|\mathcal{W}|} \alpha_{j} \boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}} . \tag{2.15}
\end{equation*}
$$

The imaging reconstruction problem then can be cast as finding the best fitting sparse map $\widehat{\boldsymbol{X}}$ from line scans $\boldsymbol{R}=\mathcal{S}\left\{\boldsymbol{\Psi} * \mathcal{L}_{\Theta}[\boldsymbol{Y}]\right\}$, and the reconstructed image is simply $\boldsymbol{D} * \widehat{\boldsymbol{X}}$. Since all associated operations on $\boldsymbol{X}_{0}$ (convolution with $\boldsymbol{D}, \boldsymbol{\psi}$ and line projection $\mathcal{L}_{\Theta}$ ) are all linear, this becomes a sparse estimation problem, which can be solved via the Lasso. In practice, due the resolution limit of probe and the sampling operation $\mathcal{S}$, we do not aiming to find exact $\boldsymbol{X}$ in a continuous space. Instead, we will solve the discretized version of
this sparse recovery problem, which assume $\boldsymbol{X}$ resides on a grid. As such, the associated Lasso problem can be written as:

$$
\begin{equation*}
\min _{\boldsymbol{X} \geq 0} \lambda \sum_{i j} \boldsymbol{X}_{i j}+\frac{1}{2}\left\|\boldsymbol{R}-\mathcal{S}\left\{\boldsymbol{\Psi} * \mathcal{L}_{\Theta}[\boldsymbol{D} * \boldsymbol{X}]\right\}\right\|_{2}^{2} \tag{2.16}
\end{equation*}
$$

### 2.5.1 Sparse recovery with Lasso from line projections

In light of Section 2.4.2 the measurement performance using infinitely many line scans is almost dependent only on the distance-to-diameter ratio of the local features. Since in practice, only finite number line scan is available, we want to study how many line scans will be sufficient for efficient and exact sparse image reconstruction. We do this by studying the performance of algorithm while assuming the line scan are idealized where the PSF is ideally all-pass in the sense that $\psi=\boldsymbol{\delta}$.

Figure 2.10 shows the reconstruction performance from line scans with varying number of lines used and number of discs in the target image $\boldsymbol{Y}$. Each image $\boldsymbol{Y}$ is generated by randomly populating the discs of size $r$ while satisfying $d / 2 r \geq 1$ via rejection sampling, and the scan angles are also uniformly random chosen. Here, two experiment settings are presented. The first is assumed that the imaging area of line scan is fixed (so the density increases linearly with more discs) and the second is considering the cases where the density is a constant (so the imaging area is proportional to the disc amount). In the phase transition (PT) image (Figure 2.10. left), each pixel represents the average of 50 experiments; and in each experiment, given random image $\boldsymbol{Y}$ and its line scans of randomly chosen angles, if solving (2.16) correctly identify the support map of $\boldsymbol{Y}$, then the algorithm succeeds, and vice versa. It shows clear transition lines in both PT images, and the comparison of scanning time between line/point probes shows clear improvement of scanning efficiency.

Interestingly if we compare the result with CS theory, which asserts the number measurement of samples required is close to linear proportional to signal sparsity; here, though the line scans are not CS-optimal, both PT images exhibits similar phenomenon. When the image size is fixed (up), total number of samples $m$ is proportional to the line count $N$, with PT transition line showing linear proportionality between number of line scans and discs $N \propto k$, gives $m \propto k$; on the other hand, when the image density is fixed, the number of samples $m$ is proportional to $N \times \sqrt{k^{3}}$ while the transition line in PT is showing $N \propto \sqrt{k}$, again suggests linear proportionality between the number of measurements and sparsity would be $m \propto N \sqrt{k} \propto k$. To wrap up, these experimental results hinted that if minimum separation of discs are ensured, then to ensure exact

[^3]Fixed image size


Fixed image density



Figure 2.10: Phase transition $\left[\mathrm{OKL}^{+}\right.$18] of fixed image size (top) and fixed density (bot) on support recovery with Lasso. In each experiments, $d / 2 r \geq 1$ is ensured. In either cases, the phase transitions (left) show the number of samples required is almost linearly proportional to the number of discs for exact reconstruction. And the the advancement of scanning efficiency (right) is presented in comparison with the point probe scans. For the fixed size case, we let (image area) $/($ disc area $) \approx 1200$; for the fixed density case, we let density $\approx(1 / 6) \cdot(\max$ density).
signal reconstruction with efficient algorithm, the number of samples required is approximately linearly proportional to the sparsity of image.

Finally, to formally elucidate the sample time reduction from point probe to line scans, we compare the consumed scanning time using different probes in both settings under specific scenarios. Figure 2.10. right). In the fixed area experiment we let the image area be $3 \times 3 \mathrm{~mm}^{2}$ and the disc radius and the image resolution are both $50 \mu \mathrm{~m}$ (image area/disc area ratio around 1200); for the fixed density we let all experiments have equal density 20 discs $/ \mathrm{mm}^{2}$ (nearly $1 / 6$ of maximum density in separating case) with same resolution. Both of the results show clear improvement of scanning efficiency, with reduction of scanning time by 3 to 10 times under these signal settings.

In either case, line measurements are substantially more efficient than measurements with a point probe.

Realizing this gain in practice requires us to modify the Lasso to cope with the following nonidealities: (i) line scans are coherent, (ii) the PSF $\psi$ is typically only partially known, and (iii) naive approaches to computing with line scans are inefficient when the target resolution is large. Below, we show how to address these issues, and give a complete reconstruction algorithm.

### 2.5.2 Computation of line projection

### 2.5.2.1 Fast computation of discrete line projection

The line projection of an image $\boldsymbol{Y}$ in direction of angle $\theta$ is equivalent to the line projection at $0^{\circ}$ of clockwise rotated $\boldsymbol{Y}$ by angle $\theta$. This enables an efficient line projection computationally via fast image rotation with shear transform in Fourier domain [LOK97].

The clockwise rotation of image $\boldsymbol{Y}$ by angle $\theta$ is

$$
\operatorname{Rot}_{\theta}[\boldsymbol{Y}](x, y)=\boldsymbol{Y}\left(\left[\begin{array}{cc}
\cos \theta & -\sin \theta  \tag{2.17}\\
\sin \theta & \cos \theta
\end{array}\right]\left[\begin{array}{l}
x \\
y
\end{array}\right]\right)
$$

where the rotational matrix can be decomposed into three shear transforms

$$
\left[\begin{array}{cc}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{array}\right]=\left[\begin{array}{cc}
1 & 0 \\
\tan \frac{\theta}{2} & 1
\end{array}\right]\left[\begin{array}{cc}
1 & -\sin \theta \\
0 & 1
\end{array}\right]\left[\begin{array}{cc}
1 & 0 \\
\tan \frac{\theta}{2} & 1
\end{array}\right] ;
$$

write both $x, y$-shear transforms as

$$
\begin{aligned}
\operatorname{Shr}_{-\mathrm{x}_{s}}[\boldsymbol{Y}](x, y) & =\boldsymbol{Y}(x+s y, y), \\
\operatorname{Shr}-\mathrm{y}_{t}[\boldsymbol{Y}](x, y) & =\boldsymbol{Y}(x, y+t x),
\end{aligned}
$$

then

Each of the shear transform can be efficiently computed in Fourier domain. Define

$$
\begin{equation*}
\widehat{\mathcal{S}}_{x, t}(u, y)=e^{j 2 \pi t y u}, \quad \widehat{\mathcal{S}}_{y, t}(x, v)=e^{j 2 \pi t x v} \tag{2.19}
\end{equation*}
$$

and $\mathcal{F}_{x}, \mathcal{F}_{y}$ as $n$-DFT in $x, y$-domain, where

$$
\begin{equation*}
\mathcal{F}_{x}\{\boldsymbol{Y}\}(u, y)=\sum_{x} \boldsymbol{Y}(x, y) e^{-j 2 \pi x u} \tag{2.20}
\end{equation*}
$$

$$
\begin{equation*}
\mathcal{F}_{y}\{\boldsymbol{Y}\}(x, v)=\sum_{y} \boldsymbol{Y}(x, y) e^{-j 2 \pi y v} \tag{2.21}
\end{equation*}
$$

From $\sqrt{2.19}-(2.21)$, the $y$-shearing transform can be written as

$$
\begin{align*}
\boldsymbol{Y}(x, y+t x) & =\sum_{y^{\prime}} \boldsymbol{Y}\left(x, y^{\prime}\right) \boldsymbol{\delta}\left(y^{\prime}-t x-y\right) \\
& =\frac{1}{n} \sum_{y^{\prime}} \boldsymbol{Y}\left(x, y^{\prime}\right) \sum_{v} e^{-j 2 \pi v\left(y^{\prime}-t x-y\right)} \\
& =\frac{1}{n} \sum_{v}\left(\sum_{y^{\prime}} \boldsymbol{Y}\left(x, y^{\prime}\right) e^{-j 2 \pi v\left(y^{\prime}-t x\right)}\right) e^{j 2 \pi v y} \\
& =\mathcal{F}_{y}^{-1}\left[\mathcal{F}_{y}[\boldsymbol{Y}] \circ \widehat{\mathcal{S}}_{y, t}\right] \tag{2.22}
\end{align*}
$$

and $x$-shear transform likewise,

$$
\begin{equation*}
\boldsymbol{Y}(x+t y, y)=\mathcal{F}_{x}^{-1}\left[\mathcal{F}_{x}[\boldsymbol{Y}] \circ \widehat{\mathcal{S}}_{x, t}\right] \tag{2.23}
\end{equation*}
$$

Combine 2.17-2.23, we obtain a computational efficient algorithm for line projections Algorithm 2,

```
Algorithm 2 Fast computational discrete line projections
Require: Discrete image \(\boldsymbol{Y} \in \mathbb{R}^{n \times n}\), line scan angles \(\left\{\theta_{1}, \ldots, \theta_{m}\right\}\).
    for \(i=1, \ldots, m\) do
        \(y\)-shearing: \(\boldsymbol{Y} \leftarrow \mathcal{F}_{y}^{-1}\left[\mathcal{F}_{y}[\boldsymbol{Y}] \circ \widehat{\mathcal{S}}_{y, \tan \left(\theta_{i} / 2\right)}\right] ;\)
        \(x\)-shearing: \(\boldsymbol{Y} \leftarrow \mathcal{F}_{x}^{-1}\left[\mathcal{F}_{x}[\boldsymbol{Y}] \circ \widehat{\mathcal{S}}_{x,-\sin \theta_{i}}\right]\);
        \(y\)-shearing: \(\boldsymbol{Y} \leftarrow \mathcal{F}_{y}^{-1}\left[\mathcal{F}_{y}[\boldsymbol{Y}] \circ \widehat{\mathcal{S}}_{y, \tan \left(\theta_{i} / 2\right)}\right]\);
        for \(t=1, \ldots, n\) do
            \(\boldsymbol{R}_{i}(t) \leftarrow \frac{1}{\sqrt{m}} \sum_{y} \boldsymbol{Y}(t, y) ;\)
        end for
    end for
Ensure: Discrete lines \(\mathcal{L}_{\Theta}[\boldsymbol{Y}]=\left\{\boldsymbol{R}_{1}, \ldots, \boldsymbol{R}_{m}\right\} \in \mathbb{R}^{n \times m}\)
```

Since the image $\boldsymbol{Y}$ is discrete, its rotation will naturally incurs interpolation error. To mitigate its effect, it is advised to limit the rotation operation to angle $\theta \in\left[-45^{\circ}, 45^{\circ}\right]$ in Algorithm 2 , then flip the image vertically or horizontally to form the image rotated by $\left[-180^{\circ}, 180^{\circ}\right)$.

Although the Fourier rotation method demands $O\left(n^{2} \log n\right)$ for computational time, which is slightly larger then the direct rotation $O\left(n^{2}\right)$, in practice we found Fourier rotation more appealing: its actual computational time is usually slightly better then other methods, since it gets around the problematic pixelated interpolation from direct rotation; and more importantly, its adjoint is easier to be calculated in a similarly explicit manner as well.


Figure 2.11: Back projection image from the scan lines. We demonstrate a simple example (left) where four discs are line projected with angles $\left\{0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}\right\}$ then undergo convolution with the simulated PSF (mid). Here, the arrows indicates the probe sweeping direction. The back projection image (right) is the superposition of back projection image of each line; and the back projection of a single line $\boldsymbol{R}_{\theta}$ assigns value $\boldsymbol{R}_{\theta}(t)$ along the sweeping directions (arrows) onto the support $\ell_{\theta, t}$ for every $t$.

### 2.5.2.2 Adjoint of line projection

The adjoint operator ${ }^{4}$ of line projections $\mathcal{L}_{\Theta}^{*}: L^{2}([m] \times \mathbb{R}) \rightarrow L^{2}\left(\mathbb{R}^{2}\right)$ is deeply connected with the wellknown tomography image reconstruction technique back projection. The adjoint of a single line projection $\mathcal{L}_{\theta_{i}}^{*}: L^{2}(\mathbb{R}) \rightarrow L^{2}\left(\mathbb{R}^{2}\right)$ of scanning angle $\theta_{i}$ is exactly the back projection of a continuous line $\widetilde{\boldsymbol{R}}_{i}$ which generates an image $\mathcal{L}_{\theta_{i}}^{*}\left[\widetilde{\boldsymbol{R}}_{i}\right]$ whose value over $\ell_{\theta_{i}, t}$ defined in in A.8 is equivalent to $\widetilde{\boldsymbol{R}}_{i}(t)$ :

$$
\begin{equation*}
\mathcal{L}_{\theta_{i}}^{*}\left[\widetilde{\boldsymbol{R}}_{i}\right](\boldsymbol{w})=\widetilde{\boldsymbol{R}}_{i}(t), \quad \forall \boldsymbol{w} \in \ell_{\theta_{i}, t} \tag{2.24}
\end{equation*}
$$

then incorporate with definition of $\ell_{\theta_{i}, t}$, we obtain a simpler form for $\mathcal{L}_{\theta_{i}}^{*}$ as

$$
\begin{equation*}
\mathcal{L}_{\theta_{i}}^{*}\left[\widetilde{\boldsymbol{R}}_{i}\right](\boldsymbol{w})=\widetilde{\boldsymbol{R}}_{i}\left(\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp} \boldsymbol{w}\right\rangle\right) . \tag{2.25}
\end{equation*}
$$

Extending the derivation of 2.25 to $m$-lines $\widetilde{\boldsymbol{R}}$, the back projection of $m$ angles $\mathcal{L}_{\Theta}^{*}$ on $\widetilde{\boldsymbol{R}}$ is the superposition of images from all $m$ back projected lines $\mathcal{L}_{\theta_{i}}^{*}\left[\widetilde{\boldsymbol{R}}_{i}\right]$ of different scanning angles:

$$
\begin{align*}
\mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{R}}](\boldsymbol{w}) & =\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \mathcal{L}_{\theta_{i}}^{*}\left[\widetilde{\boldsymbol{R}}_{i}\right](\boldsymbol{w}) \\
& =\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \widetilde{\boldsymbol{R}}_{i}\left(\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}\right\rangle\right) . \tag{2.26}
\end{align*}
$$

In the following proposition, we show that the line projections defined in 2.26 is indeed the adjoint operator of line projections.

Proposition 2.5.1. The back projection $\mathcal{L}_{\Theta}^{*}$ in 2.26 is the adjoint of line projection $\mathcal{L}_{\Theta}$ in $\left[2.2\right.$, where $\left\langle\widetilde{\boldsymbol{R}}, \mathcal{L}_{\Theta}[\boldsymbol{Y}]\right\rangle=$

[^4]$\left\langle\mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{R}}], \boldsymbol{Y}\right\rangle$.

### 2.5.2.3 Fast computation of discrete back projection

Similar to the line projection, the discrete back projection of a single line $\boldsymbol{R}_{i} \in \mathbb{R}^{n}$ at angle $\theta$ is the image $\boldsymbol{Y}_{i}=\left[\boldsymbol{R}_{i}, \boldsymbol{R}_{i}, \cdots, \boldsymbol{R}_{i}\right] \in \mathbb{R}^{n \times n}$ counterclockwise rotated by $\theta$, and the back projection of multiple lines is the sum of all such images, as shown in Figure 2.11 The discrete back projection thereby can be also calculated efficiently in Fourier domain, as presented in Algorithm 3 .

```
Algorithm 3 Fast computational discrete back projections
Require: Discrete lines \(\left\{\boldsymbol{R}_{1}, \ldots, \boldsymbol{R}_{m}\right\} \in \mathbb{R}^{n \times m}\), line scan angles \(\left\{\theta_{1}, \ldots, \theta_{m}\right\}\).
    Initialize \(\boldsymbol{Y} \leftarrow \mathbf{0} \in \mathbb{R}^{n \times n}\);
    for \(i=1, \ldots, m\) do
        for \(x=1, \ldots, n\) do
            \(\boldsymbol{Y}_{i}(x,:) \leftarrow \frac{1}{\sqrt{m}} \boldsymbol{R} ;\)
        end for
        \(y\)-shearing: \(\boldsymbol{Y}_{i} \leftarrow \mathcal{F}_{y}^{-1}\left[\mathcal{F}_{y}\left[\boldsymbol{Y}_{i}\right] \circ \widehat{\mathcal{S}}_{y,-\tan \left(\theta_{i} / 2\right)}\right] ;\)
        \(x\)-shearing: \(\boldsymbol{Y}_{i} \leftarrow \mathcal{F}_{x}^{-1}\left[\mathcal{F}_{x}\left[\boldsymbol{Y}_{i}\right] \circ \widehat{\mathcal{S}}_{x, \sin \theta_{i}}\right]\);
        \(y\)-shearing: \(\boldsymbol{Y}_{i} \leftarrow \mathcal{F}_{y}^{-1}\left[\mathcal{F}_{y}\left[\boldsymbol{Y}_{i}\right] \circ \widehat{\mathcal{S}}_{y,-\tan \left(\theta_{i} / 2\right)}\right] ;\)
        \(\boldsymbol{Y} \leftarrow \boldsymbol{Y}+\boldsymbol{Y}_{i}\)
    end for
Ensure: Discrete image \(\boldsymbol{Y} \in \mathbb{R}^{n \times n}\)
```

Remark 2.5.2. The discrete back projection from Algorithm 3 is the adjoint operator of discrete line projection from Algorithm 2. which satisfies $\left\langle\widetilde{\boldsymbol{R}}, \mathcal{L}_{\Theta}[\boldsymbol{Y}]\right\rangle=\left\langle\mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{R}}], \boldsymbol{Y}\right\rangle$.

### 2.5.3 Coping with nonidealities of practical line scans with reconstruction algorithm

As aforementioned in Section 2.4.3. in practice solving vanilla Lasso formulation in 2.16) does not provide convincing solution for practical problems in SECM with line probe, due to high coherence of line projections with respect to sparse signals and the nonidealities of PSF. These issues can be remedied by implementing well known techniques such as reweighting and blind calibration method.

### 2.5.3.1 Reweighting Lasso for coherent measurements

To cope with the coherence phenomenon, we adopt the reweighting scheme [CWB08] by solving Lasso formulation 2.16 multiple times while updating penalty variable $\boldsymbol{\lambda}$ in each iterate. At $k$-th iterate, the


Figure 2.12: SECM image reconstruction with pure Lasso and reweighted Lasso. We apply three algorithm to reconstruct the image (left) with 6 line scans with simulated PSF in Figure 2.8. The reconstruction from Lasso with large $\lambda$ (mid left) has unbalanced magnitude due to the coherence of line scans, and from Lasso with small $\lambda$ (mid right) gives blurry image by weakened sparsity regularizer. Reweighing Lasso can adjust the sparse regularizer in each iteration and consistently gives good result.
algorithm chooses the regularizer $\boldsymbol{\lambda}$ in $\left(2.16\right.$ base on the previous outcome of lasso solution $\boldsymbol{X}^{(k)}$, where

$$
\begin{equation*}
\boldsymbol{\lambda}_{i j}^{(k)} \leftarrow C\left(\boldsymbol{X}_{i j}^{(k-1)}+\varepsilon\right)^{-1} \tag{2.27}
\end{equation*}
$$

Reweighting [CWB08] is a technique in sparse recovery which is typically utilized for enhancing the sparsity regularizer, by solving Lasso formulation (2.16) multiple times while updating penalty variable $\boldsymbol{\lambda}$ in each iterate. At $k$-th iterate, the algorithm chooses the regularizer $\boldsymbol{\lambda}_{i j}^{(k)}$ base on the previous outcome of lasso solution $\boldsymbol{X}^{(k)}$, where

$$
\begin{equation*}
\boldsymbol{\lambda}_{i j}^{(k)} \leftarrow C\left(\boldsymbol{X}_{i j}^{(k-1)}+\varepsilon\right)^{-1} \tag{2.28}
\end{equation*}
$$

with $\varepsilon$ being the machine precision constant and $C$ being close to the smooth part in 2.16). The effect of reweighting method is two-fold: (i) it is a majorization-minimization algorithm of sparse regression using log-norm as sparsity surrogate [CWB08], hence, discovers sparse solution more effectively compares to the use of $\ell^{1}$-norm in Lasso; and (ii) the sparsity surrogate in final stages of reweighting approaches $\ell^{0}$-norm, by seeing $\frac{\boldsymbol{X}_{i j}^{(k+1)}}{\boldsymbol{X}_{i j}^{(k)}+\varepsilon} \approx 1$ if $\boldsymbol{X}_{i j}^{(k)} \neq 0$ as $k \rightarrow \infty$. As a result, in the final stages, problem (2.16) effectively turns into least squares, restricted to the support of $\boldsymbol{X}$, which produces a sparse solution with correct magnitude. Figure 2.12(left) displays an example of reweighting scheme, showing better reconstruction result than vanilla Lasso.

In Figure 2.12, we display an example of reweighting scheme versus the vanilla Lasso with different penalty variable in a noiseless scenario. When $\lambda$ is large, the reconstructed $\boldsymbol{X}$ does not recover correct relative magnitudes; when $\lambda$ is small, the effect of sparsity surrogate is weakened, resulting imprecise support recovery and offers blurry image. Using reweighting method correctly reconstruct the exact result. To show how the addressed modification in Lasso algorithm improves success rate of image reconstruction from line


Figure 2.13: SECM image reconstruction with reweighed Lasso and reweighed calibrating Lasso. We simulate a line scan with uneven magnitude (left), and reconstruct the image (mid left) with two algorithm. The algorithm with reweighting only (mid right) cannot identify the correct support; where the reweighting plus calibration (right) method well approximates the image.
scans, we present a series of simulated experiments, comparing the reconstruction between the vanilla Lasso with different $\lambda$ settings and reweighting method in Figure 2.14. Each data point consists of average of 30 experiments; in each of the experiment, the ground truth discs are generated at random with minimum separation (rejection sampling), which is then reconstructed from 8 random lines scans if disc number $<16$, or 16 lines scans when disc number $>16$. All discs are assumed to have equal magnitude. The correctness of the image reconstruction is measured by calculating the relative error between the pixel values of image, which measures the difference between normalized ground truth image and reconstructed image. The experiments show the reweighting method steadily outperforms the vanilla Lasso under various settings when measurements are incoherent.

### 2.5.3.2 Blind calibration for incomplete PSF information

Due to natural physical limitation, the incorrect estimation of PSF can be inevitable especially in nanoscale. A remedy therein is to parameterize the PSF to accommodate all of its possible variations which can leads to significant impact on the accuracy of reconstruction result. We assume $\boldsymbol{\psi}\left(\boldsymbol{p}_{i}\right)$ is a single instance of PSF with parameter $\boldsymbol{p}_{i}$, where the vector $\boldsymbol{p}_{i}$ can represent the peak value, the width of peak, and the rise/decay of PSF in Figure 2.8 for the scan of angle $\theta_{i}$. For the reconstruction algorithm, we replace the PSF $\psi$ in (2.16) with the parameterized version $\boldsymbol{\psi}\left(\boldsymbol{p}_{i}\right)$, and optimize both the parameter $\boldsymbol{p}_{i}$ and the sparse map $\boldsymbol{X}$ via alternating minimization.

Figure 2.13 exhibits a simulated example in which the PSF of line scans has unbalanced magnitudes due to the variation of probe scanning angle. In this example, the line scan with largest overall magnitude is four times as much as the smallest, which shows the comparison of image reconstruction results from algorithm of reweighting or of reweighting plus rescaling calibration. The figures show the calibration achieves successful


Figure 2.14: Performance of reweighting method versus Lasso. We use 8 line scans when the disc number is below 16, and 16 line scans when disc number is above for reconstruction. The experiments show reweighting method outperforms vanilla Lasso with various penalty variable $\lambda$ setting w.r.t. normalized (to 1) magnitude difference between the ground truth images and reconstructed images.
reconstruction while the former non-calibration method fell short on this simulated problem which has more than enough line scans are utilized to reconstruct a simplistic four disc example.

### 2.5.4 Image reconstruction algorithm from line scans

Finally we formally state the complete algorithm Algorithm 5 for reconstruction of SECM image from its line scans. The algorithm solves multiple iteration of

$$
\begin{equation*}
\min _{\boldsymbol{X} \geq 0, \boldsymbol{p} \in \mathcal{P}} \sum_{i j} \boldsymbol{\lambda}_{i j}^{(k)} \boldsymbol{X}_{i j}+\sum_{i=1}^{m} \frac{1}{2}\left\|\mathcal{S}\left\{\boldsymbol{\psi}\left(\boldsymbol{p}_{i}\right) * \mathcal{L}_{\theta_{i}}[\boldsymbol{D} * \boldsymbol{X}]\right\}-\boldsymbol{R}_{i}\right\|_{2}^{2} \tag{2.29}
\end{equation*}
$$

while updating the penalty variable $\boldsymbol{\lambda}^{(k)}$ in each iterate base on 2.28. To solve a single iterate of 2.29 , the algorithm utilize an accelerated alternating minimization method specifically for non-smooth, non-convex objective called iPalm algorithm [PS16] stated in Algorithm 4 Since this formulation is nonconvex and the gradients of objective 2.29 could have terribly large gradient Lipchitz constant locally, we adopt backtracking method for choosing step size of each individual gradient steps. The our real data experiments, the analytic form of $\operatorname{PSF} \widehat{\boldsymbol{\psi}}(\boldsymbol{p})$ is relized as a two-side-exponential decaying function. Define a one-side exponential-decay function as

$$
\begin{equation*}
\mathcal{E}_{c, \alpha}(t)=(c \cdot t+1)^{-\alpha} \mathbf{1}_{\{t>0\}}, \quad \check{\mathcal{E}}_{c, \alpha}(t)=\mathcal{E}_{c, \alpha}(-t) \tag{2.30}
\end{equation*}
$$

```
Algorithm \(4 \mathrm{iPalm}\left(\boldsymbol{X}_{\text {init }}, p_{\text {init }}, \boldsymbol{\lambda}, h, \mathcal{P}\right)\) : Accelerated iPalm for calibrating sparse regression
Require: Initialization \(\boldsymbol{X}_{\text {init }} \in \mathbb{R}^{n \times n}\) and \(\boldsymbol{p}_{\text {init }} \in \mathcal{P}\), sparse penalty \(\boldsymbol{\lambda} \in \mathbb{R}^{n \times n}\), smooth function \(h\), and number
    of iterations \(L\).
    Let \(\boldsymbol{X}^{(0)} \leftarrow \boldsymbol{X}_{\text {init }} ; \boldsymbol{p}^{(0)} \leftarrow p_{\text {init }} ; \alpha \leftarrow 0.9 ; t_{X 0}, t_{p 0} \leftarrow 1\)
    for \(\ell=1, \ldots, L\) do
        / / Accelerated Proximal Gradient for map \(\boldsymbol{X}\).
        \(\boldsymbol{Y}^{(\ell)} \leftarrow \boldsymbol{X}^{(\ell)}+\alpha\left(\boldsymbol{X}^{(\ell)}-\boldsymbol{X}^{(\ell-1)}\right) ;\)
        \(t \leftarrow t_{X 0} ;\)
        repeat
            \(t \leftarrow t / 2 ;\)
            \(\boldsymbol{X}^{(\ell+1)} \leftarrow \operatorname{Soft}_{t \boldsymbol{\lambda}}^{+}\left[\boldsymbol{Y}^{(\ell)}-t \partial_{\boldsymbol{X}} h\left(\boldsymbol{Y}^{(\ell)}, \boldsymbol{p}^{(\ell)}\right)\right] ;\)
        until \(h\left(\boldsymbol{X}^{(\ell+1)}, \boldsymbol{p}^{(\ell)}\right) \leq h\left(\boldsymbol{Y}^{(\ell)}, \boldsymbol{p}^{(\ell)}\right)\)
                        \(+\left\langle\bar{\partial}_{\boldsymbol{X}} h\left(\boldsymbol{Y}^{(\ell)}, \boldsymbol{p}^{(\ell)}\right), \boldsymbol{X}^{(\ell+1)}-\boldsymbol{Y}^{(\ell)}\right\rangle\)
                        \(+\frac{1}{2 t}\left\|\boldsymbol{X}^{(\ell+1)}-\boldsymbol{Y}^{(\ell)}\right\|_{2}^{2} ;\)
        \(t_{X 0} \leftarrow 4 t ;\)
        / / Accelerated Proximal Gradient for parameters \(p\).
        \(\boldsymbol{q}^{(\ell)} \leftarrow \boldsymbol{p}^{(\ell)}+\alpha\left(\boldsymbol{p}^{(\ell)}-\boldsymbol{p}^{(\ell-1)}\right) ;\)
        \(t \leftarrow t_{p 0}\);
        repeat
            \(t \leftarrow t / 2 ;\)
            \(\boldsymbol{p}^{(\ell+1)} \leftarrow \operatorname{Proj}_{\mathcal{P}}\left[\boldsymbol{q}^{(\ell)}-t \partial_{p} h\left(\boldsymbol{X}^{(\ell+1)}, \boldsymbol{q}^{(\ell)}\right)\right] ;\)
        until \(h\left(\boldsymbol{X}^{(\ell+1)}, \boldsymbol{p}^{(\ell+1)}\right) \leq h\left(\boldsymbol{X}^{(\ell+1)}, \boldsymbol{q}^{(\ell)}\right)\)
                        \(+\left\langle\partial_{p} h\left(\boldsymbol{X}^{(\ell+1)}, \boldsymbol{q}^{(\ell)}\right), \boldsymbol{p}^{(\ell+1)}-\boldsymbol{q}^{(\ell)}\right\rangle\)
                        \(+\frac{1}{2 t}\left\|\boldsymbol{p}^{(\ell+1)}-\boldsymbol{q}^{(\ell)}\right\|_{2}^{2} ;\)
        \(t_{p 0} \leftarrow 4 t ;\)
    end for
Ensure: \(\left(\boldsymbol{X}^{(L)}, \boldsymbol{p}^{(L)}\right)\) as the approximated minimizers of \(\min _{\boldsymbol{X} \geq 0, \boldsymbol{p} \in \mathcal{P}} \sum_{i j} \boldsymbol{\lambda}_{i j} \boldsymbol{X}_{i j}+h(\boldsymbol{X}, \boldsymbol{p})\)
```

then $\widehat{\boldsymbol{\psi}}(\boldsymbol{p})=\widehat{\boldsymbol{\psi}}\left(c_{\ell}, \alpha_{\ell}, c_{r}, \alpha_{r}, \sigma\right)$ is defined as

$$
\begin{equation*}
\widehat{\boldsymbol{\psi}}(\boldsymbol{p})=\left[\breve{\mathcal{E}}_{c_{\ell}, \alpha_{\ell}}+\mathcal{E}_{c_{r}, \alpha_{r}}\right] * f_{\sigma} \tag{2.31}
\end{equation*}
$$

where $f_{\sigma}$ is zero-mean Gaussian function with deviation $\sigma$.
Here, we will write the smooth part of 2.29 as

$$
h(\boldsymbol{X}, \boldsymbol{p}):=\sum_{i=1}^{m} \frac{1}{2}\left\|\mathcal{S}\left\{\boldsymbol{\psi}\left(\boldsymbol{p}_{i}\right) * \mathcal{L}_{\theta_{i}}[\boldsymbol{D} * \boldsymbol{X}]\right\}-\boldsymbol{R}_{i}\right\|_{2}^{2}
$$

in the algorithms.

```
Algorithm 5 Reconstruct SECM image with line scans via reweighted iPalm.
Require: Line scans \(\left\{\boldsymbol{R}_{i}\right\}_{i=1}^{m}\), scan angles \(\left\{\theta_{i}\right\}_{i=1}^{m}\), profile \(\boldsymbol{D}\), estimated psf \(\widehat{\boldsymbol{\psi}}\), initial guess of system param-
    eters of line scans \(\boldsymbol{p}_{\text {init }}\) within convex set \(\mathcal{P}\), and number of iterations \(K\).
    Let \(\boldsymbol{X}^{(0)} \leftarrow \mathbf{0}, \boldsymbol{p}^{(0)} \leftarrow \boldsymbol{p}_{\text {init }}\),
    Let \(h(\boldsymbol{X}, \boldsymbol{p}) \leftarrow \sum_{i=1}^{m} \frac{1}{2}\left\|\widehat{\boldsymbol{\psi}} * \mathcal{L}_{\theta_{i}}[\boldsymbol{p}][\boldsymbol{D} * \boldsymbol{X}]-\boldsymbol{R}_{i}\right\|_{2}^{2} ;\)
    for \(k=1, \ldots, K\) do
        if \(k=1\) then \(\boldsymbol{\lambda} \leftarrow C \max _{i j}\left\{\mathcal{L}_{\Theta}^{*}[\widehat{\boldsymbol{\psi}} * \boldsymbol{R}]_{i j}\right\} \cdot \mathbf{1}\);
        else
            \(\forall i, j \in[n], \boldsymbol{\lambda}_{i j}^{(k)} \leftarrow C h\left(\boldsymbol{X}^{(k)}, \boldsymbol{p}^{(k)}\right) /\left(\boldsymbol{X}_{i j}^{(k-1)}+\varepsilon\right) ;\)
        end if
        \(\left(\boldsymbol{X}^{(k+1)}, \boldsymbol{p}^{(k+1)}\right) \leftarrow \mathrm{iPalm}\left(\boldsymbol{X}^{(k)}, \boldsymbol{p}^{(k)}, \boldsymbol{\lambda}, h, \mathcal{P}\right) ;\)
    end for
Ensure: Reconstructed image \(\boldsymbol{Y} \leftarrow \boldsymbol{D} * \boldsymbol{X}^{(K)}\)
```


### 2.6 Real data experiments

We present two sets of experiments to demonstrate an end-to-end result of line probe SECM.
Figure 2.15 displays the comparison of the line probe/point probe scan on a simplistic three disc samples ( $75 \mu \mathrm{~m}$ in radius, platinum). In these experiments, the point probe tip diameter and the line probe edge thickness are equivalent $(\approx 20 \mu \mathrm{~m})$, and the probe moving speed ( 100 ms ), the sampling rate ( $10 \mu \mathrm{~m}$ ), and the probe end material (platinum) are identical as well. Four images are shown here, including the optical closeup image for the three discs, the line scans, and the reconstruction image of either point probe or the line probe. In the optical image, the arrow (scan direction) represents the line probe sweeping direction when $\theta_{s}=0^{\circ}$, which generates the $0^{\circ}$ line scan. The three discs sample is then rotated by $\theta_{s}$ ( $45^{\circ}$ in this case) clockwise, proceeds with another sweep of line probe, produces the $45^{\circ}$ line scans. This routine continues until all seven scans are carried out.

In the reconstructed images, the black circles indicated the ground truth size and location of the platinum discs derived from the optimal image. The reconstruction algorithm Algorithm 5 is setup with 6 reweighting iterates, where each iterates runs 50 iterates of Ipalm. We can see the reconstructed result from the point probe exhibits distortion in the image due to the skewness of probe PSF along its proceeding direction during raster scans; while the image of line scan reconstruction presents three circular features with its size and locations are agreeing with the ground truth, since the skewness of PSF has been successfully corrected by the reconstruction algorithm.

In Figure 2.16, we reconstructed the image of samples consist of platinum discs arranged in a more complicating configuration. Two sets of the experiment are presented here, which are the samples consist of 8


Figure 2.15: Real signal experiments on three platinum discs [DKS ${ }^{+}$19]. We show the reconstruction result of a three disc sample (up-left), which is scanned with line probe in seven different directions (up-right). The arrow in optical image represents the line probe sweeping direction, while as $\theta_{s}$ stands for clockwise rotation of the sample. The black circle indicates the correct disc location in each images. Compare to the point probe, in which the shifts of disc location are resulted from the skew of PSF (down-left), our line scan reconstruction accurately recovers the exact location (down-right). For both of the reconstructed images, the resolution is $10 \mu \mathrm{~m}$ per pixel.
or 10 discs, while the disc diameter/image resolution/probe dimension/sampling rate are all identical to the three discs case in Figure 2.15 The reconstruction algorithm are also setup similarly, with reweighting(ipalm) procedures with 6(50) iterates, generating the images of interest of much larger dimension. Notice that here we use $7(9)$ line scans on $8(10)$ disc sample respectively, and demonstrate both of the resulting reconstructed image and the location map, in which the location map is a binary image defined by $\mathbf{1}_{\left\{\boldsymbol{X}_{i j} \geq 0.5\|\boldsymbol{X}\|_{\infty}\right\}}$ at ( $i, j$ )-th entry.

We can see for these more complicating images, our algorithm are still able to reconstruct the image of platinum discs with correct location and shape. The corresponding location maps are approximately recovered, with most of the discs locations are represented by a single one-sparse vector, and some other locations are represented by a two-sparse vector due to the inevitable discretization error.

Our code for the reconstruction of SECM image from scans from line probe can be found via the following link:
https://github.com/clpsecm/clpsecm_imaging


Figure 2.16: Real signal experiments of 8, $\mathbf{1 0}$ platinum discs. Showing the optimal image of the 8 discs (up) and 10 discs (down) sample, and their corresponding line scans, reconstructed image and reconstructed disc location map. In optical image, the arrows represent the line probe sweeping direction, while as $\theta_{s}$ stands for clockwise rotation of the sample. In both examples, our algorithm is able to successfully obtain these images of the discs, with most of the disc locations can be approximately represented by an one-sparse vector. Here, the image resolution is $20 \mu \mathrm{~m}$ per pixel.

### 2.7 Summary \& Discussion

This section presents the development of a novel scanning probe microscope technique involving the use of line probe. The microscope operates line integrals in each measurement, such measurements are non-local, hence more efficient then conventional raster scans for microscopic image with localized sparse structure. This paper shows the increment in efficiency of line probe via rudimentary analysis and experiments; and proposes a simple modification in conventional CS algorithm for image reconstruction, with its effect on both the simulated and the actual datasets. Due to the strong relation between computational tomography and line scans, we also view our work can potentially being applied to ares of CT or other similar imaging modalities involving the use of projection measurements.

We envision the possibilities of future work are in multiple dimension. First, the current studied microscopic images are circumscribed in sparse convolutional model; while it has an immediate access to
applications such as lattice structure imaging in material science, we aim to expand the potential application of line scans to more general purpose imaging problems. Furthermore, unlike many other imaging modalities, in SPM the design of probe topography (i.e. the sampling pattern) is not limited to a straight line, therefore it is possible adopt various different probe design to achieve CS-like sample reduction. Lastly, in this paper we have shown via simple reasoning and experiments to exhibit the relationship between the complexity of image and the required number of line scan measurements to achieve exact reconstruction. We consider rigorously demonstrating the relationship can also be an interesting direction in CS , especially since the line scans are not the CS optimal measurement model.

## Acknowledgment

For the SECM work, I want to thank Prof. Daniel Esposito for contributing the principle idea, as well as my colleagues Dr. Anna Elisabeth Dorfi and Prof. Glen O'Neil for the construction of the microscope, execution of data collections and many helpful advices for reconstruction algorithm designs.

## Chapter 3

## Short-and-Sparse Deconvolution

In this chapter 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.

### 3.1 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 $\boldsymbol{y}=\boldsymbol{a}_{0} * \boldsymbol{x}_{0} \in \mathbb{R}^{n} 1$ This SaS model exhibits a basic scaled shift symmetry: for any nonzero scalar $\alpha$ and cyclic shift $s_{\ell}[\cdot]$,

$$
\begin{equation*}
\left(\alpha s_{\ell}\left[\boldsymbol{a}_{0}\right]\right) *\left(\frac{1}{\alpha} s_{-\ell}\left[\boldsymbol{x}_{0}\right]\right)=\boldsymbol{y} . \tag{3.1}
\end{equation*}
$$

Because of this symmetry, we only expect to recover $a_{0}$ and $x_{0}$ up to a signed shift (see Figure 3.1). Our problem of interest can be stated more formally as:

Problem 3.1.1 (Short-and-Sparse Deconvolution). Given the cyclic convolution ${ }^{2} \boldsymbol{y}=\boldsymbol{a}_{0} * \boldsymbol{x}_{0} \in \mathbb{R}^{n}$ 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 ${ }^{3}$, and nonconvex formulations exhibit a complicated optimization landscape, with many 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 ${ }^{4}$ of a shift $s_{\ell}\left[\boldsymbol{a}_{0}\right]$, 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 $\left(\boldsymbol{a}_{0}, \boldsymbol{x}_{0}\right)$. Our algorithm is based on a formulation first introduced in [ZLK $\left.{ }^{+} 17\right]$, 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 $\left(\boldsymbol{a}_{0}, \boldsymbol{x}_{0}\right)$ 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 3.2 introduces our optimization approach and modeling assumptions. Section 3.3 introduces our main results - both

[^5]

Figure 3.1: 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ and $s_{-\ell}\left[\boldsymbol{x}_{0}\right]$, 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.
geometric and algorithmic - and compares them to the literature. Section 3.4 3.5 describes the main ideas of our analysis. Finally, Section 3.8 discusses two main limitations of our analysis and describes directions for future work.

### 3.2 Formulation and Assumptions

### 3.2.1 Nonconvex SaS over the Sphere

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

$$
\begin{equation*}
\min _{\boldsymbol{a}, \boldsymbol{x}} \frac{1}{2} \underset{\text { Data Fidelity }}{\|\boldsymbol{a} * \boldsymbol{x}-\boldsymbol{y}\|_{2}^{2}+\underset{\text { Sparsity }}{\lambda\|\boldsymbol{x}\|_{1}} \text { s.t. } \quad\|\boldsymbol{a}\|_{2}=1 .} \tag{3.2}
\end{equation*}
$$

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

$$
\begin{equation*}
\varphi_{\text {lasso }}(\boldsymbol{a}) \equiv \min _{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}-\boldsymbol{y}\|_{2}^{2}+\lambda\|\boldsymbol{x}\|_{1}\right\} \tag{3.3}
\end{equation*}
$$

denote the optimal Lasso cost, we see that (3.2) simply optimizes $\varphi_{\text {lasso }}$ with respect to $\boldsymbol{a}$ :

$$
\begin{equation*}
\min _{\boldsymbol{a}} \varphi_{\text {lasso }}(\boldsymbol{a}) \text { s.t. }\|\boldsymbol{a}\|_{2}=1 . \tag{3.4}
\end{equation*}
$$

In (3.2)-(3.4), 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 3.4 is defined in terms of the optimal Lasso cost. This function is challenging to analyze, especially far away from $a_{0}$. $[\mathrm{ZLK}+17]$ analyzes the local minima of a simplification of (3.4), obtained by approximating ${ }^{5}$ the data fidelity term as

$$
\begin{align*}
\frac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}-\boldsymbol{y}\|_{2}^{2} & =\frac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}\|_{2}^{2}-\langle\boldsymbol{a} * \boldsymbol{x}, \boldsymbol{y}\rangle+\frac{1}{2}\|\boldsymbol{y}\|_{2}^{2} \\
& \approx \frac{1}{2}\|\boldsymbol{x}\|_{2}^{2}-\langle\boldsymbol{a} * \boldsymbol{x}, \boldsymbol{y}\rangle+\frac{1}{2}\|\boldsymbol{y}\|_{2}^{2} \tag{3.5}
\end{align*}
$$

This yields a simpler objective function

$$
\begin{equation*}
\varphi_{\ell^{1}}(\boldsymbol{a})=\min _{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{x}\|_{2}^{2}-\langle\boldsymbol{a} * \boldsymbol{x}, \boldsymbol{y}\rangle+\frac{1}{2}\|\boldsymbol{y}\|_{2}^{2}+\lambda\|\boldsymbol{x}\|_{1}\right\} \tag{3.6}
\end{equation*}
$$

We make one further simplification to this problem, replacing the nondifferentiable penalty $\|\cdot\|_{1}$ with a smooth approximation $\rho(\boldsymbol{x}){ }^{6}$ Our analysis allows for a variety of smooth sparsity surrogates $\rho(\boldsymbol{x})$; for concreteness, we state our main results for the particular penalty $y^{7}$

$$
\begin{equation*}
\rho(\boldsymbol{x})=\sum_{i}\left(\boldsymbol{x}_{i}^{2}+\delta^{2}\right)^{1 / 2} \tag{3.7}
\end{equation*}
$$

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,

$$
\begin{equation*}
\varphi_{\rho}(\boldsymbol{a})=\min _{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{x}\|_{2}^{2}-\langle\boldsymbol{a} * \boldsymbol{x}, \boldsymbol{y}\rangle+\frac{1}{2}\|\boldsymbol{y}\|_{2}^{2}+\lambda \rho(\boldsymbol{x})\right\} \tag{3.8}
\end{equation*}
$$

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

$$
\begin{equation*}
\min _{\boldsymbol{a}} \varphi_{\rho}(\boldsymbol{a}) \quad \text { s.t. } \quad \boldsymbol{a} \in \mathbb{S}^{p-1} \tag{3.9}
\end{equation*}
$$

Here, we set $p=3 p_{0}-2$. 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 $W=\left\{-p_{0}+1, \ldots, 2 p_{0}-1\right\}$, then for any shifts $\ell \in\left\{-p_{0}+1, \ldots, p_{0}-1\right\}$, the support of $\ell$-shifted short signal $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ is entirely contained in interval $W$. We will give a provable method which recovers a scaled version of one of these canonical shifts.

[^6]

Figure 3.2: 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 3.3.1 which, loosely speaking, asserts that when $\theta \lesssim 1 /\left(p_{0} \sqrt{\mu}+\sqrt{p_{0}}\right)$, our method succeeds.

### 3.2.2 Analysis Setting and Assumptions

For convenience, we assume that $\boldsymbol{a}_{0}$ has unit $\ell^{2}$ norm, i.e., $\boldsymbol{a}_{0} \in \mathbb{S}^{p_{0}-1} 8$ Our analysis makes two main assumptions, on the short motif $a_{0}$ and the sparse map $x_{0}$, 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 $\mu\left(\boldsymbol{a}_{0}\right)$ to be the largest inner product between distinct shifts:

$$
\begin{equation*}
\mu\left(\boldsymbol{a}_{0}\right)=\max _{\ell \neq 0}\left|\left\langle\boldsymbol{a}_{0}, s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\rangle\right| \tag{3.10}
\end{equation*}
$$

The quantity $\mu\left(\boldsymbol{a}_{0}\right)$ is bounded between 0 and 1 . Our theory allows any $\mu$ smaller than some numerical constant. Figure 3.2 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 $\mu\left(\boldsymbol{a}_{0}\right) \approx 0$ Here, the observed signal $y$ consists of a superposition of sharp pulses. This is arguably the easiest instance of SaS deconvolution.

[^7]- Generic. If $\boldsymbol{a}_{0}$ is chosen uniformly at random from the sphere $\mathbb{S}^{p_{0}-1}$, its coherence is bounded as $\mu\left(\boldsymbol{a}_{0}\right) \lesssim \sqrt{1 / p_{0}}$ with high probability.
- Tapered Generic Lowpass. Here, $a_{0}$ 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 $\mu\left(\boldsymbol{a}_{0}\right) \lesssim \beta$. In this model, $\mu$ does not have to diminish as $p_{0}$ grows - it can be a fixed constant ${ }^{10}$

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

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 $\theta$ :

$$
\begin{equation*}
\boldsymbol{x}_{0 i}=\boldsymbol{\omega}_{i} \boldsymbol{g}_{i}, \tag{3.11}
\end{equation*}
$$

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

$$
\begin{equation*}
\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \mathrm{BG}(\theta) \tag{3.12}
\end{equation*}
$$

Here, $\theta$ is the probability that a given entry $\boldsymbol{x}_{0 i}$ 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 $\theta \gtrsim 1 / p_{0}$.

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 sparsitycoherence relationship is captured in the form

$$
\begin{equation*}
\theta \lesssim 1 /\left(p_{0} \sqrt{\mu}+\sqrt{p_{0}}\right) . \tag{3.13}
\end{equation*}
$$

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 3.2. we illustrate these tradeoffs for the three models described above.

[^8]
### 3.3 Main Results: Geometry and Algorithms

In this section, we introduce our main results - on the geometry of $\varphi_{\rho}$ Section 3.3.1) and its algorithmic implications Section 3.3.2. Finally, in Section 3.3.3. we compare these results with the literature on deconvolution.

### 3.3.1 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 $\pm s_{\ell}\left[\boldsymbol{a}_{0}\right]$. The shifts $\pm s_{\ell}\left[\boldsymbol{a}_{0}\right]$ 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 $\pm s_{\ell}\left[\boldsymbol{a}_{0}\right]$.

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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ of the ground truth $\boldsymbol{a}_{0}$. In Figure 3.3. we plot the function value of $\varphi_{\rho}$ over

$$
\mathcal{B}_{\ell^{2}, r}\left(s_{\ell}\left[\boldsymbol{a}_{0}\right]\right) \cap \mathbb{S}^{p-1}
$$

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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$.
- There is a local minimizer very close to $s_{\ell}\left[\boldsymbol{a}_{0}\right]$.


Figure 3.3: 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$.

Geometry near the span of two shifts. We next visualize the objective function $\varphi_{\rho}$ near the linear span of two different shifts $s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right]$ and $s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]$. More precisely, we plot $\varphi_{\rho}$ near the intersection Figure 3.4 left) of the sphere $\mathbb{S}^{p-1}$ and the linear subspace

$$
\mathcal{S}_{\left\{\ell_{1}, \ell_{2}\right\}}=\left\{\boldsymbol{\alpha}_{1} s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right]+\boldsymbol{\alpha}_{2} s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right] \mid \boldsymbol{\alpha}_{1}, \boldsymbol{\alpha}_{2} \in \mathbb{R}\right\}
$$



Figure 3.4: Geometry of $\varphi_{\rho}$ near the span $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}\right\}}$ of two shifts of $\boldsymbol{a}_{0}$. Left: each pair of shifts $s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right], s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]$ defines a linear subspace $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}\right\}}$ of $\mathbb{R}^{p}$. Center/right: every local minimum of $\varphi_{\rho}$ near $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}\right\}}$ (red line) is close to either $s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right]$ or $s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]$; there is a negative curvature in the middle of $s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right], s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]$, 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$.
- These are the only local minimizers in the vicinity of $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}\right\}}$. In particular, the objective function $\varphi$ exhibits negative curvature along $S_{\left\{\ell_{1}, \ell_{2}\right\}}$ at any superposition $\boldsymbol{\alpha}_{1} s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right]+\boldsymbol{\alpha}_{2} s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]$ whose weights $\boldsymbol{\alpha}_{1}$ and $\boldsymbol{\alpha}_{2}$ are balanced, i.e., $\left|\boldsymbol{\alpha}_{1}\right| \approx\left|\boldsymbol{\alpha}_{2}\right|$.
- 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 3.5, left) of the sphere $\mathbb{S}^{p-1}$ with the linear span of three shifts $s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right], s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right], s_{\ell_{3}}\left[\boldsymbol{a}_{0}\right]$ of the true kernel $\boldsymbol{a}_{0}$ :

$$
\mathcal{S}_{\left\{\ell_{1}, \ell_{2}, \ell_{3}\right\}}=\left\{\boldsymbol{\alpha}_{1} s_{\ell_{1}}\left[\boldsymbol{a}_{0}\right]+\boldsymbol{\alpha}_{2} s_{\ell_{2}}\left[\boldsymbol{a}_{0}\right]+\boldsymbol{\alpha}_{3} s_{\ell_{3}}\left[\boldsymbol{a}_{0}\right] \mid \boldsymbol{\alpha}_{1}, \boldsymbol{\alpha}_{2}, \boldsymbol{\alpha}_{3} \in \mathbb{R}\right\}
$$

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 3.5: Geometry of $\varphi_{\rho}$ over the span $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}, \ell_{3}\right\}}$ of three shifts of $\boldsymbol{a}_{0}$. The subspace $\mathcal{S}_{\left\{\ell_{1}, \ell_{2}, \ell_{3}\right\}}$ 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 $s_{\ell_{i}}\left[\boldsymbol{a}_{0}\right]$, and are the only minimizers near $\mathcal{S}_{\ell_{1}, \ell_{2}, \ell_{3}}$.

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

$$
\begin{equation*}
\boldsymbol{\tau} \subseteq\left\{-p_{0}+1, \ldots, p_{0}-1\right\} \tag{3.14}
\end{equation*}
$$

define a linear subspace

$$
\begin{equation*}
\mathcal{S}_{\boldsymbol{\tau}}=\left\{\sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_{\ell} s_{\ell}\left[\boldsymbol{a}_{0}\right] \mid \boldsymbol{\alpha}_{-p_{0}+1}, \ldots, \boldsymbol{\alpha}_{p_{0}-1} \in \mathbb{R}\right\} . \tag{3.15}
\end{equation*}
$$

The subspace $\mathcal{S}_{\tau}$ is the linear span of the shifts $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ indexed by $\ell$ in the set $\tau$. Our geometric theory will show that with high probability the function $\varphi_{\rho}$ has no spurious local minimizers near any $\mathcal{S}_{\tau}$ for which $\tau$ is not too large - say, $|\boldsymbol{\tau}| \leq 4 \theta p_{0}$. Combining all of these subspaces into a single geometric object, define the union of subspaces

$$
\begin{equation*}
\Sigma_{4 \theta p_{0}}=\bigcup_{|\tau| \leq 4 \theta p_{0}} \mathcal{S}_{\boldsymbol{\tau}} . \tag{3.16}
\end{equation*}
$$

Figure 3.6(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 $\Sigma_{4 \theta p_{0}}$.


Figure 3.6: 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 $\mathcal{S}_{\tau}$. Right: $\varphi_{\rho}$ has good geometry near this union of subspaces.

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 3.3.1 (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 $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }}$ $\operatorname{BG}(\theta) \in \mathbb{R}^{n}$ with sparsity rate

$$
\begin{equation*}
\theta \in\left[\frac{c_{1}}{p_{0}}, \frac{c_{2}}{p_{0} \sqrt{\mu}+\sqrt{p_{0}}}\right] \cdot \frac{1}{\log ^{2} p_{0}} . \tag{3.17}
\end{equation*}
$$

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 $n \geq \operatorname{poly}\left(p_{0}\right)$, with high probability, every local minimizer $\overline{\boldsymbol{a}}$ of $\varphi_{\rho}$ over $\Sigma_{4 \theta p_{0}}$ satisfies $\left\|\overline{\boldsymbol{a}}-\sigma s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\|_{2} \leq$ $c \max \left\{\mu, p_{0}^{-1}\right\}$ for some signed shift $\sigma s_{\ell}\left[\boldsymbol{a}_{0}\right]$ of the true kernel. Above, $c_{1}, c_{2}>0$ are positive numerical constants.

Proof. This follows from Theorem 3.4.1.

The upper bound on $\theta$ in (3.17) yields the tradeoff between coherence and sparsity described in Figure 3.2 Simply put, when $a_{0}$ is better conditioned (as a kernel), its coherence $\mu$ is smaller and $x_{0}$ can be denser.

At a technical level, our proof of Theorem 3.3.1 shows that (i) $\varphi_{\rho}(\boldsymbol{a})$ is strongly convex in the vicinity of each signed shift, and that at every other point $a$ 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, $\left\|\overline{\boldsymbol{a}}-s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\|_{2} \lesssim \mu$ ) for us to exactly recover $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ : we will give a refinement algorithm that produces $\left( \pm s_{\ell}\left[\boldsymbol{a}_{0}\right], \pm s_{-\ell}\left[\boldsymbol{x}_{0}\right]\right)$.

### 3.3.2 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 $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(k)}, \ldots$ 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 $\overline{\boldsymbol{a}}$, we refine it to produce a signed shift of $\left(\boldsymbol{a}_{0}, \boldsymbol{x}_{0}\right)$ 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 6) and state our main algorithmic result Theorem 3.3.2, which asserts that under appropriate conditions this method produces a signed shift of $\left(\boldsymbol{a}_{0}, \boldsymbol{x}_{0}\right)$.

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:

$$
\begin{equation*}
\boldsymbol{y}=\boldsymbol{a}_{0} * \boldsymbol{x}_{0}=\sum_{\ell \in \operatorname{supp}\left(\boldsymbol{x}_{0}\right)} \boldsymbol{x}_{0 \ell} s_{\ell}\left[\boldsymbol{a}_{0}\right] \tag{3.18}
\end{equation*}
$$

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

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

$$
\begin{equation*}
\left[\mathbf{0}^{p_{0}-1} ; \boldsymbol{y}_{0} ; \cdots ; \boldsymbol{y}_{p_{0}-1} ; \mathbf{0}^{p_{0}-1}\right] \tag{3.19}
\end{equation*}
$$

Zero padding provides enough space to accommodate any shift $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ with $\ell \in \boldsymbol{\tau}$. We then perform one step


Figure 3.7: 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ 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).
of the generalized power method ${ }^{11}$, writing

$$
\begin{equation*}
\boldsymbol{a}^{(0)}=-\boldsymbol{P}_{\mathbb{S}^{p}-1} \nabla \varphi_{\ell^{1}}\left(\boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\mathbf{0}^{p_{0}-1} ; \boldsymbol{y}_{0} ; \cdots ; \boldsymbol{y}_{p_{0}-1} ; \mathbf{0}^{p_{0}-1}\right]\right), \tag{3.20}
\end{equation*}
$$

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 3.5 after interpreting the gradient $\nabla \varphi_{\rho}$ in terms of its action on the shifts $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ in Section 3.4. For now, we note that this operation has the effect of (approximately) filling in the missing pieces of the truncated shifts $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ - see Figure 3.7 for an example. We will prove that with high probability $\boldsymbol{a}^{(0)}$ is indeed close to $\Sigma_{4 \theta p_{0}}$.

[^9]The next key observation is that the function $\varphi_{\rho}$ grows as we move away from the subspace $\mathcal{S}_{\tau}-$ see Figure 3.8 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

$$
\begin{equation*}
\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\boldsymbol{a}^{(k)}-t \boldsymbol{g}^{(k)}-t^{2} \boldsymbol{v}^{(k)}\right] \tag{3.21}
\end{equation*}
$$



Figure 3.8: Growth of $\varphi_{\rho}$ away from $\mathcal{S}_{\tau}$. Because $\varphi_{\rho}$ grows away from $\mathcal{S}_{\tau}$, small-stepping descent methods stay near $\mathcal{S}_{\tau}$.
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 $\overline{\boldsymbol{a}}$ that is near some signed shift $\pm s_{\ell}\left[\boldsymbol{a}_{0}\right]$ of the ground truth.

Refinement: Rounding a near-solution with homotopy alternating minimization. The second step of our algorithm rounds the local minimizer $\overline{\boldsymbol{a}} \approx \sigma s_{\ell}\left[\boldsymbol{a}_{0}\right]$ to produce an exact solution $\widehat{\boldsymbol{a}}=\sigma s_{\ell}\left[\boldsymbol{a}_{0}\right]$. As a byproduct, it also exactly recovers the corresponding signed shift of the true sparse signal, $\widehat{\boldsymbol{x}}=\sigma s_{-\ell}\left[\boldsymbol{x}_{0}\right]$.

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 $I^{(k)}$ :

$$
\begin{equation*}
\frac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}-\boldsymbol{y}\|_{2}^{2}+\lambda \sum_{i \notin I^{(k)}}\left|\boldsymbol{x}_{i}\right| . \tag{3.22}
\end{equation*}
$$

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

$$
\begin{equation*}
\widehat{a} * \widehat{x} \approx y \tag{3.23}
\end{equation*}
$$

This can be viewed as a homotopy or continuation method $\mathrm{OPTOO}^{\mathrm{EHJ}}{ }^{+} 04$. For concreteness, at $k$-th iteration


Initial $\boldsymbol{a}^{(0)}$

$\boldsymbol{a}^{(100)}$


Converged $\overline{\boldsymbol{a}}$


Est. $\widehat{\boldsymbol{a}}$ and true $\boldsymbol{a}_{0}$

Figure 3.9: 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 $\boldsymbol{a}_{0}$.
the algorithm reads:

$$
\begin{align*}
\text { Update } \boldsymbol{x}: & \boldsymbol{x}^{(k+1)} \leftarrow \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{a}^{(k)} * \boldsymbol{x}-\boldsymbol{y}\right\|_{2}^{2}+\lambda^{(k)} \sum_{i \notin I^{(k)}}\left|\boldsymbol{x}_{i}\right|,  \tag{3.24}\\
\text { Update } \boldsymbol{a}: & \boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\underset{\boldsymbol{a}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)}-\boldsymbol{y}\right\|_{2}^{2}\right],  \tag{3.25}\\
\text { Update } \lambda \text { and } I: & \lambda^{(k+1)} \leftarrow \frac{1}{2} \lambda^{(k)}, \quad I^{(k+1)} \leftarrow \operatorname{supp}\left(\boldsymbol{x}^{(k+1)}\right) . \tag{3.26}
\end{align*}
$$

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 $\overline{\boldsymbol{a}}$ is sufficiently nearby.

Algorithm and Main Algorithmic Result. Our overall algorithm is summarized as Algorithm 6. Figure 3.9 illustrates the main steps of this algorithm. Our main algorithmic result states that under closely related hypotheses as above, Algorithm 6 produces a signed shift of the ground truth $\left(\boldsymbol{a}_{0}, \boldsymbol{x}_{0}\right)$ :

Theorem 3.3.2 (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 $\max _{i \neq j}\left|\left\langle\boldsymbol{\iota}_{p_{0}}^{*} s_{i}\left[\boldsymbol{a}_{0}\right], \boldsymbol{\iota}_{p_{0}}^{*} s_{j}\left[\boldsymbol{a}_{0}\right]\right\rangle\right| \leq \mu$ and $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \mathrm{BG}(\theta) \in \mathbb{R}^{n}$ with $\theta, \mu$ satisfying

$$
\begin{equation*}
\theta \in\left[\frac{c_{1}}{p_{0}}, \frac{c_{2}}{\left(p_{0} \sqrt{\mu}+\sqrt{p_{0}}\right) \log ^{2} p_{0}}\right], \quad \mu \leq \frac{c_{3}}{\log ^{2} n} \tag{3.32}
\end{equation*}
$$

for some constant $c_{1}, c_{2}, c_{3}>0$. If the signal lengths $n, p_{0}$ satisfy $n>\operatorname{poly}\left(p_{0}\right)$ and $p_{0}>\operatorname{polylog}(n)$, then there exist $\delta, \eta_{v}>0$ such that with high probability, Algorithm 6 produces $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}})$ that are equal to the ground truth up to signed shift symmetry:

$$
\begin{equation*}
\left\|(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}})-\sigma\left(s_{\ell}\left[\boldsymbol{a}_{0}\right], s_{-\ell}\left[\boldsymbol{x}_{0}\right]\right)\right\|_{2} \leq \varepsilon \tag{3.33}
\end{equation*}
$$

for some $\sigma \in\{-1,1\}$ and $\ell \in\left\{-p_{0}+1, \ldots, p_{0}-1\right\}$ if $K_{1}>\operatorname{poly}\left(n, p_{0}\right)$ and $K_{2}>\operatorname{polylog}\left(n, p_{0}, \varepsilon^{-1}\right)$.

Proof. See Theorem 3.5.1 and Theorem 3.5.2

```
Algorithm 6 Short and Sparse Deconvolution
Input: Observation \(\boldsymbol{y}\), motif length \(p_{0}\), sparsity \(\theta\), shift-coherence \(\mu\), and curvature threshold \(-\eta_{v}\).
    Minimization:
    Set \(\boldsymbol{a}^{(0)} \leftarrow-\boldsymbol{P}_{\mathbb{S}^{p-1}} \nabla \varphi_{\rho}\left(\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\mathbf{0}^{p_{0}-1} ; \boldsymbol{y}_{0} ; \cdots ; \boldsymbol{y}_{p_{0}-1} ; \mathbf{0}^{p_{0}-1}\right]\right)\).
    Set \(\lambda=0.1 / \sqrt{p_{0} \theta}{ }^{12}\) and \(\delta>0\) in \(\varphi_{\rho}\). For \(k=1,2, \ldots, K_{1}\), let
\[
\begin{equation*}
\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\boldsymbol{a}^{(k)}-t \boldsymbol{g}^{(k)}-t^{2} \boldsymbol{v}^{(k)}\right] \tag{3.27}
\end{equation*}
\]
```

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 $\left\langle\boldsymbol{v}^{(k)}, \boldsymbol{g}^{(k)}\right\rangle \geq 0$, otherwise let $\boldsymbol{v}^{(k)}=\mathbf{0}$; and $t \in(0,0.1 / n \theta]$ satisfies

$$
\begin{equation*}
\varphi_{\rho}\left(\boldsymbol{a}^{(k+1)}\right)<\varphi_{\rho}\left(\boldsymbol{a}^{(k)}\right)-\frac{1}{2} t\left\|\boldsymbol{g}^{(k)}\right\|_{2}^{2}-\frac{1}{4} t^{4} \eta_{v}\left\|\boldsymbol{v}^{(k)}\right\|_{2}^{2} \tag{3.28}
\end{equation*}
$$

to obtain a near local minimizer $\overline{\boldsymbol{a}} \leftarrow \boldsymbol{a}^{\left(K_{1}\right)}$.
Refinement:
$\overline{\text { Set } \boldsymbol{a}^{(0)} \leftarrow \overline{\boldsymbol{a}}}, \lambda^{(0)} \leftarrow 10(p \theta+\log n)(\mu+1 / p)$ and $I^{(0)} \leftarrow \mathcal{S}_{\lambda(0)}[\operatorname{supp}(\breve{\boldsymbol{y}} * \overline{\boldsymbol{a}}])$. For $k=1,2, \ldots, K_{2}$, let
$\boldsymbol{x}^{(k+1)} \leftarrow \operatorname{argmin}_{\boldsymbol{x}} \frac{1}{2}\left\|\boldsymbol{a}^{(k)} * \boldsymbol{x}-\boldsymbol{y}\right\|_{2}^{2}+\lambda^{(k)} \sum_{i \notin I^{(k)}}\left|\boldsymbol{x}_{i}\right|$,
$\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\operatorname{argmin}_{\boldsymbol{a}} \frac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)}-\boldsymbol{y}\right\|_{2}^{2}\right]$,
$\lambda^{(k+1)} \leftarrow \lambda^{(k)} / 2, \quad I^{(k+1)} \leftarrow \operatorname{supp}\left(\boldsymbol{x}^{(k+1)}\right)$,
to obtain $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}}) \leftarrow\left(\boldsymbol{a}^{\left(K_{2}\right)}, \boldsymbol{x}^{\left(K_{2}\right)}\right)$.
Output: Return $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{x}})$.

When solving SaS deconvolution via minimizing bilinear Lasso objective (3.3) in practice, the algorithm is analogous to the provable method introduced in Algorithm 6. where the curvilinear descent and the refinement step can be realized as alternating gradient descent of both variables $\boldsymbol{a}, \boldsymbol{x}$ in 3.3. Unlike Algorithm 6 , this alternating gradient method has yet come with theoretical guarantees, but shown to be an effective and efficient method for SaS deconvolution problems both in simulation and in reality $\left[\mathrm{LQK}^{+} 19 \mid\right.$.

### 3.3.3 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 [ $\overline{\left.\text { PSG }^{+} 16\right]}$, where the observed signal exhibits repetitive short neuron excitation patterns occurring sparsely across time and/or space. Similarly,

[^10]electron microscopy images $\left[\mathrm{CLC}^{+} 17\right]$ arising in study of nanomaterials often exhibit repeated motifs.
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 [ $\mathrm{FSH}^{+} 06$, 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. $\left[\mathrm{ZLK}^{+} 17\right]$ 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 $a_{0}$. 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ to the window $\left\{0, \ldots, p_{0}-1\right\}$. 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$, under complicated conditions
on the kernel $\boldsymbol{a}_{0}$. Our core optimization problem is very similar to [ZLK $\left.{ }^{+} 17\right]$. 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 $x_{0}$ 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 $i=1, \ldots, K$ 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 $y$ into $K$ pieces; this apparent reduction fails due to boundary effects.

### 3.3.4 Notations

Vectors and indices. All vectors/matrices are written in bold font $\boldsymbol{a} / \boldsymbol{A}$; indexed values are written as $\boldsymbol{a}_{i}, \boldsymbol{A}_{i j}$. 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 $\{0,1, \ldots, n-1\}$. We write $[n]=\{0, \ldots, n-1\}$. We often use captial italic symbols $I, J$ for subsets of $[n]$. We abuse notation slightly and write $[-p]=\{n-p+1, \ldots, n-1,0\}$ and $[ \pm p]=$ $\{n-p+1, \ldots, n-1,0,1, \ldots, p-1\}$. 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 $\boldsymbol{\iota}_{I}: \mathbb{R}^{|I|} \rightarrow \mathbb{R}^{n}$ injects the input vector to higher dimensional Euclidean space, via $\left(\boldsymbol{\iota}_{I} \boldsymbol{x}\right)_{i}=\boldsymbol{x}_{I^{-1}(i)}$ for $i \in I$ and 0 otherwise. Its adjoint operator $\iota_{I}^{*}$ can be understood as subset selection operator which picks up entries of coordinates $I$. A common zero-filling operator through out this paper $\iota$ is abbreviation of $\iota_{[p]}$, which is often being addressed as zero-padding operator and its adjoint $\iota^{*}$ 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: $I * J=\operatorname{supp}\left(\mathbf{1}_{I} * \mathbf{1}_{J}\right)$. Similarly, the shift operator $s_{\ell}[\cdot]: \mathbb{R}^{p} \rightarrow \mathbb{R}^{n}$ is circular with modulo-n without specification: $\left(s_{\ell}[\boldsymbol{a}]\right)_{j}=\left(\boldsymbol{\iota}_{[p]} \boldsymbol{a}\right)_{j-\ell}$. Notice that here $\boldsymbol{a}$ can be shorter $p \leq n$. 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 $b \in \mathbb{R}^{n}$,

$$
\begin{equation*}
C_{a} b=a * b \tag{3.34}
\end{equation*}
$$

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 $\breve{\boldsymbol{C}}_{\boldsymbol{a}}$ as $\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{e}_{j}=s_{j}[\widetilde{\boldsymbol{a}}]$ and $\breve{\boldsymbol{C}}_{\boldsymbol{a}} \boldsymbol{e}_{j}=s_{-j}[\boldsymbol{a}]$. The two operators will satisfy

$$
\begin{equation*}
C_{a}^{*} b=\check{a} * b, \quad \check{C}_{a} b=a * \check{b} \tag{3.35}
\end{equation*}
$$

### 3.4 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ 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 $\pm s_{\ell}\left[\boldsymbol{a}_{0}\right]$.

### 3.4.1 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 $\boldsymbol{a}_{0}$ :

$$
\begin{equation*}
\boldsymbol{a}=\sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_{\ell} s_{\ell}\left[\boldsymbol{a}_{0}\right] \tag{3.36}
\end{equation*}
$$

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

$$
\begin{equation*}
\boldsymbol{a}=\sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_{\ell} s_{\ell}\left[\boldsymbol{a}_{0}\right]+\sum_{\ell \notin \boldsymbol{\tau}} \boldsymbol{\alpha}_{\ell} s_{\ell}\left[\boldsymbol{a}_{0}\right] \tag{3.37}
\end{equation*}
$$

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 $\beta \in \mathbb{R}^{n}$, whose entries are the inner products between $\boldsymbol{a}$ and the shifts of $\boldsymbol{a}_{0}: \boldsymbol{\beta}_{\ell}=\left\langle\boldsymbol{a}, s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\rangle$. Since the columns of $\boldsymbol{C}_{\boldsymbol{a}_{0}}$ are the shifts of $\boldsymbol{a}_{0}$, we can write

$$
\begin{align*}
\beta & =C_{a_{0}}^{*} \iota a  \tag{3.38}\\
& =C_{a_{0}}^{*} \iota \iota^{*} C_{a_{0}} \alpha=: M \alpha \tag{3.39}
\end{align*}
$$

The matrix $\boldsymbol{M}$ is the Gram matrix of the truncated shifts $\iota^{*} s_{\ell}\left[\boldsymbol{a}_{0}\right]: \boldsymbol{M}_{i j}=\left\langle\boldsymbol{\iota}^{*} s_{i}\left[\boldsymbol{a}_{0}\right], \boldsymbol{\iota}^{*} s_{j}\left[\boldsymbol{a}_{0}\right]\right\rangle$. 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.2.

### 3.4.2 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:

$$
\begin{equation*}
\varphi_{\ell^{1}}(\boldsymbol{a})=-\frac{1}{2}\left\|\mathcal{S}_{\lambda}[\check{\boldsymbol{y}} * \boldsymbol{a}]\right\|_{2}^{2} . \tag{3.40}
\end{equation*}
$$

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 $\mathcal{S}_{\lambda}[x]$ shrinks the elements of $x$ towards zero. Small elements become identically zero, resulting in a sparse vector.

## Gradient: Sparsifying the Correlations $\beta$

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

$$
\begin{align*}
\nabla \varphi_{\ell^{1}}(\boldsymbol{a}) & =-\iota^{*} \boldsymbol{C}_{a_{0}} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}\left[\breve{C}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \iota a\right] \\
& =-\iota^{*} \boldsymbol{C}_{a_{0}} \breve{C}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{\beta}\right] \\
& =-\iota^{*} \boldsymbol{C}_{a_{0}} \boldsymbol{\chi}[\boldsymbol{\beta}], \tag{3.41}
\end{align*}
$$

where we have simplified the notation by introducing an operator $\chi: \mathbb{R}^{n} \rightarrow \mathbb{R}^{n}$ 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 $a_{0}$ with coefficients given by the entries of $\boldsymbol{\chi}[\boldsymbol{\beta}]$ :

$$
\begin{equation*}
-\nabla \varphi_{\ell^{1}}(\boldsymbol{a})=\sum_{\ell} \chi[\boldsymbol{\beta}]_{\ell} s_{\ell}\left[\boldsymbol{a}_{0}\right] . \tag{3.42}
\end{equation*}
$$

The operator $\chi$ 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,

$$
\frac{1}{n \theta} \cdot \boldsymbol{\chi}[\boldsymbol{\beta}]_{\ell} \approx \begin{cases}\boldsymbol{\beta}_{\ell}-\lambda, & \boldsymbol{\beta}_{\ell}>\lambda  \tag{3.43}\\ \boldsymbol{\beta}_{\ell}+\lambda, & \boldsymbol{\beta}_{\ell}<-\lambda \\ 0, & \text { otherwise }\end{cases}
$$

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 3.10 (middle left) and compare to Figure 3.10(left). Because $\boldsymbol{\chi}[\boldsymbol{\beta}]$ suppresses small entries of $\boldsymbol{\beta}$, the strongest contributions to $-\nabla \varphi_{\ell^{1}}$ in (3.42) will come from


Figure 3.10: 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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$, 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 $\boldsymbol{\beta}\left(\boldsymbol{a}^{+}\right)$after, one projected gradient step $\boldsymbol{a}^{+}=\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a}-t \cdot \operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})\right]$. Notice that the small entries of $\boldsymbol{\beta}$ are shrunk towards zero. Right: the $\operatorname{gradient} \operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})$ is large whenever it is easy to sparsify $\boldsymbol{\beta}$; in particular, when the largest entry $\boldsymbol{\beta}_{(0)} \gg \boldsymbol{\beta}_{(1)} \gg 0$.
shifts $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ with large $\boldsymbol{\beta}_{\ell}$. In particular, the Euclidean gradient is large whenever there is a single preferred shift $s_{\ell}\left[\boldsymbol{a}_{0}\right]$, 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

$$
\begin{align*}
\operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a}) & =\boldsymbol{P}_{\boldsymbol{a}^{\perp}} \nabla \varphi_{\ell^{1}}(\boldsymbol{a}) \\
& =-\boldsymbol{P}_{\boldsymbol{a}^{\perp}} \sum_{\ell} \chi_{\ell}[\boldsymbol{\beta}] s_{\ell}\left[\boldsymbol{a}_{0}\right] . \tag{3.44}
\end{align*}
$$

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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ 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 $\underbrace{13}$ Since the tangent space consists of those vectors orthogonal to $\boldsymbol{a}$, this is possible whenever $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ is not too aligned with $\boldsymbol{a}$, i.e., $\boldsymbol{a}$ is not too close to $s_{\ell}\left[\boldsymbol{a}_{0}\right]$.

Our technical lemma quantifies this situation in terms of the ordered entries of $\boldsymbol{\beta}$. Write $\left|\boldsymbol{\beta}_{(0)}\right| \geq\left|\boldsymbol{\beta}_{(1)}\right| \geq \ldots$, with corresponding shifts $s_{(0)}\left[\boldsymbol{a}_{0}\right], s_{(1)}\left[\boldsymbol{a}_{0}\right], \ldots$ There is a strong gradient whenever $\left|\boldsymbol{\beta}_{(0)}\right|$ is significantly larger than $\left|\boldsymbol{\beta}_{(1)}\right|$ and $\left|\boldsymbol{\beta}_{(1)}\right|$ is not too small compared to $\lambda$ : in particular, when $\frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|>\left|\boldsymbol{\beta}_{(1)}\right|>\frac{\lambda}{4 \log ^{2} \theta^{-1}}$.

[^11]In this situation, gradient descent drives $\boldsymbol{a}$ toward $s_{(0)}\left[\boldsymbol{a}_{0}\right]$, reducing $\left|\boldsymbol{\beta}_{(1)}\right|, \ldots$, 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 B.3.

## Hessian: Negative Curvature Breaks Symmetry

When there is no single preferred shift, i.e., when $\left|\boldsymbol{\beta}_{(1)}\right|$ is close to $\left|\boldsymbol{\beta}_{(0)}\right|$, the gradient can be small. Similarly, when $\boldsymbol{a}$ is very close to $\pm s_{(0)}\left[\boldsymbol{a}_{0}\right]$, 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

$$
\begin{equation*}
\widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a})=-\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{P}_{I} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota} \tag{3.45}
\end{equation*}
$$

with $I=\operatorname{supp}\left(\mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \iota \boldsymbol{a}\right]\right)$. 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

$$
\begin{equation*}
\widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a}) \approx-\sum_{\ell} s_{\ell}\left[\boldsymbol{a}_{0}\right] s_{\ell}\left[\boldsymbol{a}_{0}\right]^{*}\left(\frac{\partial}{\partial \boldsymbol{\beta}_{\ell}} \boldsymbol{\chi}_{\ell}[\boldsymbol{\beta}]\right) \tag{3.46}
\end{equation*}
$$

can be obtained from (3.42) noting that $\frac{\partial}{\partial \boldsymbol{a}} \boldsymbol{\chi}_{\ell}[\boldsymbol{\beta}]=\sum_{j} s_{j}\left[\boldsymbol{a}_{0}\right] \frac{\partial}{\partial \boldsymbol{\beta}_{j}} \boldsymbol{\chi}_{\ell}[\boldsymbol{\beta}]$, that $\frac{\partial}{\partial \boldsymbol{\beta}_{j}} \boldsymbol{\chi}_{\ell}[\boldsymbol{\beta}] \approx 0$ for $j \neq \ell$, and that

$$
\frac{1}{n \theta} \cdot \frac{\partial \boldsymbol{\chi}_{\ell}[\boldsymbol{\beta}]}{\partial \boldsymbol{\beta}_{\ell}} \approx \begin{cases}0 & \left|\boldsymbol{\beta}_{\ell}\right| \ll \lambda  \tag{3.47}\\ 1 & \left|\boldsymbol{\beta}_{\ell}\right| \gg \lambda\end{cases}
$$

Again, we corroborate this approximation pictorially - see Figure 3.11
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 $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ corresponding to some $\left|\boldsymbol{\beta}_{\ell}\right| \geq \lambda$, or whenever $\boldsymbol{v}$ is a linear combination of such shifts. In particular, if for some $j,\left|\boldsymbol{\beta}_{(0)}\right|,\left|\boldsymbol{\beta}_{(1)}\right|, \ldots,\left|\boldsymbol{\beta}_{(j)}\right| \gg \lambda$, then $\varphi_{\ell^{1}}$ will exhibit negative curvature in any



Negative curvature: $\boldsymbol{\beta}_{(1)}>\frac{4}{5} \boldsymbol{\beta}_{(0)}$
Strong convexity: $\boldsymbol{\beta}_{(1)}<\nu \boldsymbol{\lambda}$

Figure 3.11: Hessian Breaks Symmetry. Left: contribution of $-s_{i}\left[\boldsymbol{a}_{0}\right] s_{i}\left[\boldsymbol{a}_{0}\right]^{*}$ to the Euclidean hessian. If $\left|\boldsymbol{\beta}_{i}\right| \gg \lambda$ the Euclidean hessian exhibits a strong negative component in the $s_{i}\left[\boldsymbol{a}_{0}\right]$ direction. The Riemmanian hessian exhibits negative curvature in directions spanned by $s_{i}\left[\boldsymbol{a}_{0}\right]$ with corresponding $\left|\boldsymbol{\beta}_{\boldsymbol{i}}\right| \gg \lambda$ and positive curvature in directions spanned by $s_{i}\left[\boldsymbol{a}_{0}\right]$ with $\left|\boldsymbol{\beta}_{i}\right| \ll \lambda$. Middle: this creates negative curvature along the subspace $\mathcal{S}_{\tau}$ 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.
direction $\boldsymbol{v} \in \operatorname{span}\left(s_{(0)}\left[\boldsymbol{a}_{0}\right], s_{(1)}\left[\boldsymbol{a}_{0}\right], \ldots, s_{(j)}\left[\boldsymbol{a}_{0}\right]\right)$.

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

$$
\widetilde{\operatorname{Hess}}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})=\boldsymbol{P}_{\boldsymbol{a}^{\perp}}\left(\begin{array}{c}
\widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a})  \tag{3.48}\\
\text { Curvature of } \varphi_{\ell^{1}}
\end{array}+\begin{array}{c}
\left\langle-\nabla \varphi_{\ell^{1}}(\boldsymbol{a}), \boldsymbol{a}\right\rangle \cdot \boldsymbol{I} \\
\text { Curvature of the sphere }
\end{array}\right) \boldsymbol{P}_{\boldsymbol{a}^{\perp}} .
$$

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 $\left\langle-\nabla \varphi_{\ell^{1}}(\boldsymbol{a}), \boldsymbol{a}\right\rangle$ 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 $\left|\boldsymbol{\beta}_{(0)}\right|,\left|\boldsymbol{\beta}_{(1)}\right|, \ldots,\left|\boldsymbol{\beta}_{(j)}\right| \gg \lambda, \varphi_{\ell^{1}}$ will exhibit negative curvature in any tangent direction $\boldsymbol{v} \perp \boldsymbol{a}$ which is in the linear span

$$
\operatorname{span}\left(s_{(0)}\left[\boldsymbol{a}_{0}\right], s_{(1)}\left[\boldsymbol{a}_{0}\right], \ldots, s_{(j)}\left[\boldsymbol{a}_{0}\right]\right)
$$

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 $\mathcal{S}_{\tau}$. The Riemannian Hessian is
positive in these directions, creating positive curvature orthogonal to the subspace $\mathcal{S}_{\tau}$.
(iii) Strong convexity around minimizers: Around a minimizer $s_{\ell}\left[\boldsymbol{a}_{0}\right]$, only a single entry $\boldsymbol{\beta}_{\ell}$ is large. Any tangent direction $\boldsymbol{v} \perp \boldsymbol{a}$ is nearly orthogonal to the subspace $\operatorname{span}\left(s_{\ell}\left[\boldsymbol{a}_{0}\right]\right)$, and hence is a direction of positive (Riemmanian) curvature. The objective function $\varphi_{\rho}$ is strongly convex around the target solutions $\pm s_{\ell}\left[\boldsymbol{a}_{0}\right]$.

Figure 3.11visualizes 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 B.4.

### 3.4.3 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 $\mathcal{S}_{\tau}$ can be covered by regions of negative curvature, large gradient, and regions of strong convexity containing target solutions $\pm s_{\ell}\left[\boldsymbol{a}_{0}\right]$. 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 $\mathcal{S}_{\tau}$. For technical reasons, it turns out to be convenient to do this in terms of the coefficients $\alpha$ in the representation

$$
\begin{equation*}
\boldsymbol{a}=\sum_{\ell \in \boldsymbol{\tau}} \boldsymbol{\alpha}_{\ell} s_{\ell}\left[\boldsymbol{a}_{0}\right]+\sum_{\ell^{\prime} \in \boldsymbol{\tau}^{c}} \boldsymbol{\alpha}_{\ell^{\prime}} s_{\ell^{\prime}}\left[\boldsymbol{a}_{0}\right] \tag{3.49}
\end{equation*}
$$

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 $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}$ as a measure of the distance from $\boldsymbol{a}$ to $\mathcal{S}_{\boldsymbol{\tau}}$. A technical wrinkle arises, because the representation 3.49 is not unique. We resolve this issue by choosing the $\boldsymbol{\alpha}$ that minimizes $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}$, writing:

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)=\inf \left\{\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}: \sum_{\ell} \boldsymbol{\alpha}_{\ell} s_{\ell}\left[\boldsymbol{a}_{0}\right]=\boldsymbol{a}\right\} \tag{3.50}
\end{equation*}
$$

The distance $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)$ 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 $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)$ is not too large. Similar to 3.16 , we define an object which contains all points that are close to some $\mathcal{S}_{\boldsymbol{\tau}}$, in the above sense:

$$
\begin{equation*}
\Sigma_{4 \theta p_{0}}^{\gamma}:=\bigcup_{|\boldsymbol{\tau}| \leq 4 \theta p_{0}}\left\{\boldsymbol{a}: d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \gamma\right\} \tag{3.51}
\end{equation*}
$$

The aforementioned geometric properties hold over this set:

Theorem 3.4.1 (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

$$
\begin{equation*}
\theta \in\left[\frac{c^{\prime}}{p_{0}}, \frac{c}{p_{0} \sqrt{\mu}+\sqrt{p_{0}}}\right] \cdot \frac{1}{\log ^{2} p_{0}} \tag{3.52}
\end{equation*}
$$

for some constants $c^{\prime}, c>0$. Set $\lambda=0.1 / \sqrt{p_{0} \theta}$ in $\varphi_{\rho}$ where $\rho(x)=\sqrt{x^{2}+\delta^{2}}$. There exist numerical constants $C, c^{\prime \prime}, c^{\prime \prime \prime}, c_{1}-c_{4}>0$ such that if $\delta \leq \frac{c^{\prime \prime} \lambda \theta^{8}}{p^{2} \log ^{2} n}$ and $n>C p_{0}^{5} \theta^{-2} \log p_{0}$, then with probability at least $1-c^{\prime \prime \prime} / n$, for every $\boldsymbol{a} \in \Sigma_{4 \theta p_{0}}^{\gamma}$, we have:

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

$$
\begin{equation*}
\lambda_{\min }\left(\operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a})\right) \leq-c_{1} n \theta \lambda \tag{3.53}
\end{equation*}
$$

- (Large gradient): If $\nu_{1}\left|\boldsymbol{\beta}_{(0)}\right| \geq\left|\boldsymbol{\beta}_{(1)}\right| \geq \nu_{2}(\theta) \lambda$, then

$$
\begin{equation*}
\left\|\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})\right\|_{2} \geq c_{2} n \theta \frac{\lambda^{2}}{\log ^{2} \theta^{-1}} \tag{3.54}
\end{equation*}
$$

- (Convex near shifts): If $\nu_{2}(\theta) \lambda \geq\left|\boldsymbol{\beta}_{(1)}\right|$, then

$$
\begin{equation*}
\operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a}) \succ c_{3} n \theta \boldsymbol{P}_{\boldsymbol{a}^{\perp}} \tag{3.55}
\end{equation*}
$$

- (Retraction to subspace): If $\frac{\gamma}{2} \leq d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \gamma$, then for every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{a}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}$, there exists $\boldsymbol{\zeta}$ satisfying $\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})=\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\zeta}$, such that

$$
\begin{equation*}
\left\langle\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\rangle \geq c_{4}\left\|\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}\right\|_{2}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \tag{3.56}
\end{equation*}
$$

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

$$
\begin{equation*}
\min _{\substack{\ell \in[ \pm p] \\ \sigma \in\{ \pm 1\}}}\left\|\boldsymbol{a}-\sigma s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\|_{2} \leq \frac{1}{2} \max \left\{\mu, p_{0}^{-1}\right\} \tag{3.57}
\end{equation*}
$$

where $\nu_{1}=\frac{4}{5}, \nu_{2}(\theta)=\frac{1}{4 \log ^{2} \theta^{-1}}$ and $\gamma=\frac{c \cdot \operatorname{poly}(\sqrt{1 / \theta}, \sqrt{1 / \mu})}{\log ^{2} \theta^{-1}} \cdot \frac{1}{\sqrt{p_{0}}}$.
Proof. See Appendix B.6.5.
The retraction property elaborated in (3.56) implies that the negative gradient at $a$ points in a direction that decreases $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)$. 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 $\operatorname{grad}\left[\varphi_{\rho}\right](\underline{\boldsymbol{a}})$. Then $\underline{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}=\mathbf{0}, \underline{\boldsymbol{\zeta}}_{\boldsymbol{\tau}^{c}} \approx \mathbf{0}$ and

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

Our main geometric claim in Theorem 3.3.1 is a direct consequence of Theorem 3.4.1. Moreover, it suggests that as long as we can minimize $\varphi_{\rho}$ within the region $\Sigma_{4 \theta p_{0}}^{\gamma}$, we will solve the SaS deconvolution problem.

### 3.5 Provable Algorithm

In light of Theorem 3.4.1, in this section we introduce a two-part algorithm Algorithm 6, which first applies the curvilinear descent method to find a local minimum of $\varphi_{\rho}$ within $\Sigma_{4 \theta p_{0}}^{\gamma}$, followed by refinement algorithm that uses alternating minimization to exactly recover the ground truth. This algorithm exactly solves SaS deconvolution problem.

### 3.5.1 Minimization

There are three major issues in finding a local minimizer within $\Sigma_{4 \theta p_{0}}^{\gamma}$. We want $\ldots$
(i) Initialization. the initializer $\boldsymbol{a}^{(0)}$ to reside within $\Sigma_{4 \theta p_{0}}^{\gamma}$,
(ii) Negative curvature. the method to avoid stagnating near the saddle points of $\varphi_{\rho}$,
(iii) No exit. the descent method to remain inside $\Sigma_{4 \theta p_{0}}^{\gamma}$.

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

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

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

is the normalized gradient vector from a chunk of data $\boldsymbol{a}^{(-1)}:=\boldsymbol{P}_{\left[p_{0}\right]}\left(\boldsymbol{a}_{0} * \widetilde{\boldsymbol{x}}_{0}\right)$ with $\widetilde{\boldsymbol{x}}_{0}$ a normalized BernoulliGaussian random vector of length $2 p_{0}-1$. Since $\nabla \varphi_{\rho} \approx \nabla \varphi_{\ell^{1}}$, expand the gradient $\nabla \varphi_{\ell^{1}}$ and rewrite the
gradient $\nabla_{\ell^{1}}\left(\boldsymbol{a}^{(-1)}\right)$ in shift space, we get

$$
\begin{aligned}
-\nabla \varphi_{\rho^{1}}\left(\boldsymbol{a}^{(-1)}\right) & \approx \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{P}_{\left[p_{0}\right]}\left(\boldsymbol{a}_{0} * \widetilde{\boldsymbol{x}}_{0}\right)\right] \\
& =\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\chi}\left[\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{P}_{\left[p_{0}\right]} \boldsymbol{C}_{\boldsymbol{a}_{0}} \widetilde{\boldsymbol{x}}_{0}\right] \\
& \approx \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\chi}\left[\widetilde{\boldsymbol{x}}_{0}\right] \\
& \approx n \theta \cdot \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \mathcal{S}_{\lambda}\left[\widetilde{\boldsymbol{x}}_{0}\right]
\end{aligned}
$$

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

$$
\begin{equation*}
\max _{i \neq j}\left|\left\langle\boldsymbol{\iota}_{p_{0}}^{*} s_{i}\left[\boldsymbol{a}_{0}\right], \boldsymbol{\iota}_{p_{0}}^{*} s_{j}\left[\boldsymbol{a}_{0}\right]\right\rangle\right| \leq \mu \ll 1 \tag{3.58}
\end{equation*}
$$

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

$$
\begin{equation*}
\boldsymbol{a}^{(0)} \approx \boldsymbol{P}_{\mathbb{S}^{p-1}} \iota^{*} \boldsymbol{C}_{a_{0}} \mathcal{S}_{\lambda}\left[\widetilde{\boldsymbol{x}}_{0}\right] \tag{3.59}
\end{equation*}
$$

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

$$
\begin{equation*}
\boldsymbol{a}^{(0)} \in \Sigma_{4 \theta p_{0}}^{\gamma} \tag{3.60}
\end{equation*}
$$

Furthermore, since $\widetilde{\boldsymbol{x}}_{0}$ is normalized, the largest magnitude for entries of $\left|\widetilde{\boldsymbol{x}}_{0}\right|$ is likely to be around $1 / \sqrt{2 p_{0} \theta}$. 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 $\left|\widetilde{\boldsymbol{x}}_{0}\right|$. We suggest setting $\lambda$ in $\varphi_{\rho}$ as

$$
\begin{equation*}
\lambda=\frac{c}{\sqrt{p_{0} \theta}} . \tag{3.61}
\end{equation*}
$$

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

Minimize $\varphi_{\rho}$ within $\Sigma_{4 \theta p_{0}}^{\gamma}$. 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 $\Sigma_{4 \theta p_{0}}^{\gamma}$ - in particular, negative curvature in symmetrybreaking 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 $\boldsymbol{v}^{*} \boldsymbol{g} \geq 0$. In particular, we set

$$
\begin{equation*}
\boldsymbol{a}^{+} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\boldsymbol{a}-t \boldsymbol{g}-t^{2} \boldsymbol{v}\right] \tag{3.62}
\end{equation*}
$$

For small $t$,

$$
\begin{equation*}
\varphi\left(\boldsymbol{a}^{+}\right) \approx \varphi(\boldsymbol{a})+\langle\boldsymbol{g}, \boldsymbol{\xi}\rangle+\frac{1}{2} \boldsymbol{\xi}^{*} \boldsymbol{H} \boldsymbol{\xi} \tag{3.63}
\end{equation*}
$$

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

$$
\begin{equation*}
\left|d_{\alpha}\left(\boldsymbol{a}^{+}, \mathcal{S}_{\tau}\right)-d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\tau}\right)\right| \leq \frac{\gamma}{2} \tag{3.64}
\end{equation*}
$$

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

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{a}^{+}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \tag{3.65}
\end{equation*}
$$

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

$$
\begin{equation*}
\forall k=1,2, \ldots, \quad a^{(k)} \in \Sigma_{4 \theta p_{0}}^{\gamma} \tag{3.66}
\end{equation*}
$$

We conclude this section with the following theorem:

Theorem 3.5.1 (Convergence of retractive curvilinear search). Suppose signals $\boldsymbol{a}_{0}, \boldsymbol{x}_{0}$ satisfy the conditions of Theorem 3.4.1. $\theta>10^{3} c / p_{0}(c>1)$, and $\boldsymbol{a}_{0}$ is $\mu$-truncated shift coherent $\max _{i \neq j}\left|\left\langle\boldsymbol{\iota}_{p_{0}}^{*} s_{i}\left[\boldsymbol{a}_{0}\right], \iota_{p_{0}}^{*} s_{j}\left[\boldsymbol{a}_{0}\right]\right\rangle\right| \leq \mu$. Write $\boldsymbol{g}=\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})$ and $\boldsymbol{H}=\operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a})$. 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 $\boldsymbol{v}^{*} \boldsymbol{g} \geq 0$; otherwise let $\boldsymbol{v}=\mathbf{0}$. Define a sequence $\left\{\boldsymbol{a}^{(k)}\right\}_{k \in \mathbb{N}}$ where $\boldsymbol{a}^{(0)}$ equals 3.20 and for $k=1,2, \ldots, K_{1}$ :

$$
\begin{equation*}
\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\boldsymbol{a}^{(k)}-t \boldsymbol{g}^{(k)}-t^{2} \boldsymbol{v}^{(k)}\right] \tag{3.67}
\end{equation*}
$$

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

$$
\begin{equation*}
\varphi_{\rho}\left(\boldsymbol{a}^{(k+1)}\right)<\varphi_{\rho}\left(\boldsymbol{a}^{(k)}\right)-\frac{1}{2}\left(t\left\|\boldsymbol{g}^{(k)}\right\|_{2}^{2}+\frac{1}{2} t^{4} \eta_{v}\left\|\boldsymbol{v}^{(k)}\right\|_{2}^{2}\right) \tag{3.68}
\end{equation*}
$$

then with probability at least $1-1 / c$, there exists some signed shift $\overline{\boldsymbol{a}}= \pm s_{i}\left[\boldsymbol{a}_{0}\right]$ where $i \in\left[ \pm p_{0}\right]$ such that $\left\|\boldsymbol{a}^{(k)}-\overline{\boldsymbol{a}}\right\|_{2} \leq \mu+1 / p$ for all $k \geq K_{1}=\operatorname{poly}(n, p)$. Here, $\eta_{v}=c^{\prime} n \theta \lambda$ for some $c^{\prime}<c_{1}$ in Theorem 3.4.1

Proof. See Appendix B.7.2.

### 3.5.2 Local Refinement

In this section, we describe and analyze an algorithm which refines an estimate $\overline{\boldsymbol{a}} \approx \boldsymbol{a}_{0}$ of the kernel to exactly recover $\left(\boldsymbol{a}_{0}, \boldsymbol{x}_{0}\right)$. Set

$$
\begin{equation*}
\boldsymbol{a}^{(0)} \leftarrow \overline{\boldsymbol{a}}, \quad \lambda^{(0)} \leftarrow C(p \theta+\log n)(\mu+1 / p), \quad I^{(0)} \leftarrow \operatorname{supp}\left(\mathcal{S}_{\lambda}\left[\boldsymbol{C}_{\overline{\boldsymbol{a}}}^{*} \boldsymbol{y}\right]\right) \tag{3.69}
\end{equation*}
$$

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

$$
\begin{align*}
& \boldsymbol{x}^{(k+1)} \leftarrow \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{a}^{(k)} * \boldsymbol{x}-\boldsymbol{y}\right\|_{2}^{2}+\lambda^{(k)} \sum_{i \notin I^{(k)}}\left|\boldsymbol{x}_{i}\right|,  \tag{3.70}\\
& \boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\underset{\boldsymbol{a}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)}-\boldsymbol{y}\right\|_{2}^{2}\right]  \tag{3.71}\\
& \lambda^{(k+1)} \leftarrow \frac{1}{2} \lambda^{(k)}, \quad I^{(k+1)} \leftarrow \operatorname{supp}\left(\boldsymbol{x}^{(k+1)}\right) . \tag{3.72}
\end{align*}
$$

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

$$
\begin{equation*}
\min _{\sigma \in \pm 1, \ell \in\left[ \pm p_{0}\right]}\left\|\boldsymbol{a}^{(k)}-\sigma \cdot s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\|_{2} \leq C^{\prime} 2^{-k} \tag{3.73}
\end{equation*}
$$

Modified coherence and support density assumptions It should be clear that exact recovery is unlikely if $x_{0}$ 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:

$$
\begin{equation*}
\kappa_{I}:=6 \max \{\theta p, \log n\} \tag{3.74}
\end{equation*}
$$

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

$$
\begin{equation*}
I=\operatorname{supp}\left(\boldsymbol{x}_{0}\right), \quad I^{(k)}=\operatorname{supp}\left(\boldsymbol{x}^{(k)}\right) \tag{3.75}
\end{equation*}
$$

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

$$
\begin{equation*}
\max _{\ell}|I \cap([p]+\ell)| \leq \kappa_{I} \tag{3.76}
\end{equation*}
$$

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

$$
\begin{equation*}
(\mu+1 / p) \cdot \kappa_{I}^{2}<c \tag{3.77}
\end{equation*}
$$

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

Alternating minimization produces $\boldsymbol{a}$ that contracts toward $\boldsymbol{a}_{0}$. Recall that B.25) in Theorem 3.4.1 provides that

$$
\begin{equation*}
\left\|\overline{\boldsymbol{a}}-\boldsymbol{a}_{0}\right\|_{2} \leq(\mu+1 / p), \tag{3.78}
\end{equation*}
$$

which is sufficiently close to $\boldsymbol{a}_{0}$ as long as 3.76) holds true. Here, we will elaborate this by showing a single iteration of alternating minimization algorithm 3.70-3.72 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 (3.70) can be written as

$$
\begin{equation*}
\boldsymbol{x}^{(k+1)}=\boldsymbol{\iota}_{J}\left(\iota_{J}^{*} \boldsymbol{C}_{\boldsymbol{a}^{(k)}}^{*} \boldsymbol{C}_{\boldsymbol{a}^{(k)}} \boldsymbol{\iota}_{J}\right)^{-1} \boldsymbol{\iota}_{J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a}^{(k)}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}_{0}-\lambda^{(k)} \boldsymbol{P}_{J \backslash T} \boldsymbol{\sigma}^{(k+1)}\right), \tag{3.79}
\end{equation*}
$$

and the solution to least squares problem (3.71) will be

$$
\begin{equation*}
\boldsymbol{a}^{(k+1)}=\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^{*} \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}} \boldsymbol{\iota}\right)^{-1}\left(\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota \boldsymbol{a}_{0}\right) \tag{3.80}
\end{equation*}
$$

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 3.79 gives

$$
\begin{align*}
\boldsymbol{x}^{(k+1)} & \approx \boldsymbol{x}_{0}  \tag{3.81}\\
\left(\boldsymbol{x}^{(k+1)}-\boldsymbol{x}_{0}\right) & \approx \boldsymbol{P}_{I}\left(\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}_{0}-\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}(k)} \boldsymbol{x}_{0}\right) \\
& \approx \boldsymbol{P}_{I}\left[\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}_{0}-\boldsymbol{a}^{(k)}\right)\right], \tag{3.82}
\end{align*}
$$

which implies, while assuming $\boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \approx n \theta \boldsymbol{I}$, that from 3.80):

$$
\begin{align*}
\left(\boldsymbol{a}^{(k+1)}-\boldsymbol{a}_{0}\right) & \approx(n \theta)^{-1} \iota^{*} \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota} \boldsymbol{a}_{0}-\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}}^{*} \boldsymbol{C}_{\boldsymbol{x}^{(k+1)}} \iota \boldsymbol{a}_{0} \\
& \approx(n \theta)^{-1} \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{x}_{0}-\boldsymbol{x}^{(k+1)}\right) \\
& \approx(n \theta)^{-1} \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}^{(k)}-\boldsymbol{a}_{0}\right) \tag{3.83}
\end{align*}
$$

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 $(n \theta)^{-1} \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}$ approximates a contraction mapping with fixed point $\boldsymbol{a}_{0}$, as follows:

$$
\begin{align*}
(n \theta)^{-1} \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota} & \approx \boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{e}_{0} \boldsymbol{e}_{0}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \iota \\
& \approx \boldsymbol{a}_{0} \boldsymbol{a}_{0}^{*} \tag{3.84}
\end{align*}
$$

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 3.5.2 (Linear rate convergence of alternating minimization). Suppose $\boldsymbol{y}=\boldsymbol{a}_{0} * \boldsymbol{x}_{0}$ where $\boldsymbol{a}_{0}$ is $\mu$-shift coherent and $\boldsymbol{x}_{0} \sim \operatorname{BG}(\theta)$, 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

$$
\begin{equation*}
\left\|\boldsymbol{a}^{(0)}-\boldsymbol{a}_{0}\right\|_{2} \leq \mu+1 / p, \quad \lambda^{(0)}=5 \kappa_{I}(\mu+1 / p), \quad I^{(0)}=\operatorname{supp}\left(\boldsymbol{C}_{\boldsymbol{a}^{(0)}}^{*} \boldsymbol{y}\right) \tag{3.85}
\end{equation*}
$$

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

$$
\begin{align*}
& \boldsymbol{x}^{(k+1)} \leftarrow \operatorname{argmin}_{\boldsymbol{x}} \frac{1}{2}\left\|\boldsymbol{a}^{(k)} * \boldsymbol{x}-\boldsymbol{y}\right\|_{2}^{2}+\lambda^{(k)} \sum_{i \notin I^{(k)}}\left|\boldsymbol{x}_{i}\right|  \tag{3.86}\\
& \boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S} p-1}\left[\operatorname{argmin}_{\boldsymbol{a}} \frac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)}-\boldsymbol{y}\right\|_{2}^{2}\right]  \tag{3.87}\\
& \lambda^{(k+1)} \leftarrow \frac{1}{2} \lambda^{(k)}, \quad I^{(k+1)} \leftarrow \operatorname{supp}\left(\boldsymbol{x}^{(k+1)}\right) \tag{3.88}
\end{align*}
$$

then

$$
\begin{equation*}
\left\|\boldsymbol{a}^{(k+1)}-\boldsymbol{a}_{0}\right\|_{2} \leq(\mu+1 / p) 2^{-k} \tag{3.89}
\end{equation*}
$$

for every $k=0,1,2, \ldots$
Proof. See Appendix B.8.3.
Remark 3.5.3. The estimates $\boldsymbol{x}^{(k)}$ also converges to the ground truth $\boldsymbol{x}_{0}$ at a linear rate.

### 3.6 From Analysis to Practical Algorithm

The algorithmic idea for minimizing $\varphi_{\rho}$ base on our understanding of its geometry can be similarly applied to the practical algorithm—solving SaSD with Bilinear Lasso formulation 3.90 :

$$
\begin{equation*}
\min _{\boldsymbol{a} \in \mathbb{S}^{p-1}, \boldsymbol{x} \in \mathbb{R}^{n}} \lambda\|\boldsymbol{x}\|_{1}+\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{a} * \boldsymbol{x}\|_{2}^{2} \tag{3.90}
\end{equation*}
$$

Optimization over lifted space the sphere for $\boldsymbol{a}$ For both the Bilinear Lasso and $\varphi_{\rho}$, a unit-norm constraint on $\boldsymbol{a}$ is enforced to break the scaling symmetry between $\boldsymbol{a}_{0}$ and $\boldsymbol{x}_{0}$. In contrast to enforcing $\ell_{1}$-norm constraints where it leads to spurious minimizers for deconvolution problems [LWDF11, BVG13, ZLK ${ }^{+}$17]; choosing the $\ell_{2}$-norm, has surprisingly strong implications for optimization. The $\varphi_{\rho}$ objective, for example, is piecewise concave whenever $\boldsymbol{a}$ is sufficiently far away from any shift of $\boldsymbol{a}_{0}$, but the sphere induces positive curvature near individual shifts to create strong convexity. These two properties combine to ensure recoverability of $\boldsymbol{a}_{0}$.

Likewise, the solutions of the short kernel $\boldsymbol{a} \in \mathbb{R}^{p_{0}}$ is the collection of shifted $\boldsymbol{a}_{0}$ as well, which, is the set of $\left\{s_{-p_{0}+1}\left[\boldsymbol{a}_{0}\right], \ldots, s_{p_{0}-1}\left[\boldsymbol{a}_{0}\right]\right\}$. To ensure all the local minimizers on the subspace $\mathcal{S}_{\boldsymbol{\tau}}$ are the exact shifts, we naturally optimize $\boldsymbol{a}$ in higher dimension space $\mathbb{R}^{3 p_{0}-2=p}$, to effectivly deal with the boundary issues caused by shift symmetry.

Initializing near a few shifts. The landscape of $\varphi_{\rho}$ makes single shifts of $\boldsymbol{a}_{0}$ easy to locate if $\boldsymbol{a}$ is initialized near a span of a few shifts. Fortunately, this is a relatively simple matter in SaSD , since, as we mentioned, that $\boldsymbol{y}$ is itself a sparse superposition of shifts. Therefore, one initialization strategy is to randomly choose a length- $p_{0}$ window $\widetilde{\boldsymbol{y}}=\left[\boldsymbol{y}_{i}, \boldsymbol{y}_{i+1} \ldots \boldsymbol{y}_{i+p_{0}-1}\right]$ from the observation and set

$$
\begin{equation*}
\boldsymbol{a}^{(0)} \doteq \boldsymbol{P}_{\mathbb{S}^{p}-1}\left(\left[\mathbf{0}^{p_{0}-1} ; \widetilde{\boldsymbol{y}}_{i} ; \mathbf{0}^{p_{0}-1}\right]\right) \tag{3.91}
\end{equation*}
$$

This brings $\boldsymbol{a}^{(0)}$ suitably close to the sum of a few shifts of $\boldsymbol{a}_{0}$; any truncation effects are absorbed by padding the ends of $\widetilde{\boldsymbol{y}}_{i}$, which also sets the length for $\boldsymbol{a}$ to be $p=3 p_{0}-2$.

Implications for practical computation. The regionally benign optimization landscape of $\varphi_{\rho}$ guarantees that efficient recovery is possible for SaSD when $\boldsymbol{a}_{0}$ is incoherent. Applications of sparse deconvolution, however, are often motivated by sharpening or resolution tasks [CFG14b, CE16] where the motif $\boldsymbol{a}_{0}$ is smooth and coherent (i.e. $\mu\left(\boldsymbol{a}_{0}\right)$ is large). The $\varphi_{\rho}$ objective is a poor approximation of the Bilinear Lasso in such cases and fails to yield practical algorithms, so we should optimize the Bilinear Lasso directly.

In Section 3.5.1. we introduce the algorithm for finding minimizer of $\varphi_{\rho}$ with marginal minimization over sparse $\boldsymbol{x}$ and apply second order method (Riemannian curvilinear search) for descending $\boldsymbol{a}$ over sphere. Though this construction of algorithm provides exact convergence guarantee, in practice we can simply adopt several of its key ideas for practical algorithms, which works nicely on real data applications using Bilinear lasso objective.

Besides initialize $\boldsymbol{a}^{(0)}$ as 3.91, we initialize $\boldsymbol{x}^{(0)}$ by marginal minimize over $\boldsymbol{x}$ in Bilinear Lasso with fixed $\boldsymbol{a}^{(0)}$ :

$$
\begin{equation*}
\boldsymbol{x}^{(0)}=\underset{\boldsymbol{x}}{\operatorname{argmin}}\left(\lambda\|\boldsymbol{x}\|_{1}+\frac{1}{2}\left\|\boldsymbol{a}^{(0)} * \boldsymbol{x}-\boldsymbol{y}\right\|_{F}^{2}\right) \tag{3.92}
\end{equation*}
$$

and as such

### 3.7 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 $\theta \in\left[10^{-3}, 10^{-2}\right]$ and $p_{0} \in\left[10^{3}, 10^{4}\right]$, we draw 50 instances of synthetic data by choosing $\boldsymbol{a}_{0} \sim \operatorname{Unif}\left(\mathbb{S}^{p_{0}-1}\right)$ and $\boldsymbol{x}_{0} \in \mathbb{R}^{n}$ with $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ where $n=5 \times 10^{5}$. Note that choosing $\boldsymbol{a}_{0}$ this way implies $\mu\left(\boldsymbol{a}_{0}\right) \approx \frac{1}{\sqrt{p_{0}}}$.

For each instance, we recover $\boldsymbol{a}_{0}$ and $\boldsymbol{x}_{0}$ from $\boldsymbol{y}=\boldsymbol{a}_{0} * \boldsymbol{x}_{0}$ by minimizing problem 3.6. For ease of computation, we modify Algorithm 6by replacing curvilinear search with accelerated Riemannian gradient
descent method Algorithm 7, 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

$$
\begin{align*}
\operatorname{Exp}_{\boldsymbol{a}}(\boldsymbol{u}) & :=\cos \left(\|\boldsymbol{u}\|_{2}\right) \cdot \boldsymbol{a}+\sin \left(\|\boldsymbol{u}\|_{2}\right) \cdot \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_{2}}  \tag{3.93}\\
\log _{\boldsymbol{a}}(\boldsymbol{b}) & :=\arccos (\langle\boldsymbol{a}, \boldsymbol{b}\rangle) \cdot \frac{\boldsymbol{P}_{\boldsymbol{a} \perp}(\boldsymbol{b}-\boldsymbol{a})}{\left\|\boldsymbol{P}_{\boldsymbol{a} \perp}(\boldsymbol{b}-\boldsymbol{a})\right\|_{2}} \tag{3.94}
\end{align*}
$$

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

For each recovery instance, we say the local minimizer $\boldsymbol{a}_{\text {min }}$ generated from Algorithm 7 is sufficiently close to a solution of SaS deconvolution problem, if

$$
\begin{equation*}
\operatorname{success}\left(\boldsymbol{a}_{\min }, ; \boldsymbol{a}_{0}\right):=\left\{\max _{\ell}\left|\left\langle s_{\ell}\left[\boldsymbol{a}_{0}\right], \boldsymbol{a}_{\min }\right\rangle\right|>0.95\right\} \tag{3.95}
\end{equation*}
$$

The result is shown in Figure 3.12. Our source code can be accessed via the following address:

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

```
Algorithm 7 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 \(\boldsymbol{a}^{(0)} \leftarrow-\boldsymbol{P}_{\mathbb{S}^{p}-1} \nabla \varphi_{\rho}\left(\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\mathbf{0}^{p_{0}-1} ;\left[\boldsymbol{y}_{0}, \cdots, \boldsymbol{y}_{p_{0}-1}\right] ; \mathbf{0}^{p_{0}-1}\right]\right)\),
    for \(k=1,2, \ldots, K\) do
        Get momentum: \(\boldsymbol{w} \leftarrow \operatorname{Exp}_{\boldsymbol{a}^{(k)}}\left(\eta \cdot \log _{\boldsymbol{a}^{(k-1)}}\left(\boldsymbol{a}^{(k)}\right)\right)\).
        Get negative gradient direction: \(\boldsymbol{g} \leftarrow-\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{w})\).
        Armijo step \(\boldsymbol{a}^{(k+1)} \leftarrow \operatorname{Exp}_{\boldsymbol{w}}(t \boldsymbol{g})\), choosing \(t \in(0,1)\) s.t. \(\varphi_{\rho}\left(\boldsymbol{a}^{(k+1)}\right)-\varphi_{\rho}(\boldsymbol{w})<-t\|\boldsymbol{g}\|_{2}^{2}\).
    end for
Output: Return \(\boldsymbol{a}^{(K)}\).
```


### 3.8 Summary \& Discussion

In this section, we close by discussing the most important limitations of our results when $\boldsymbol{a}_{0}$ is coherent, about scenarios when the signal setting breaches our assumption, especially when $\boldsymbol{x}_{0}$ is either highly sparse or non-symmetric, and highlighting corresponding directions for future work.

The main drawback of our proposed method is that it does not succeed when the target motif $a_{0}$ has


Figure 3.12: Success probability of SaS deconvolution under generic $a_{0}, x_{0}$ 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 7 works with high probability when $\theta \lesssim \frac{1}{\sqrt{p_{0}}}$ in generic case.
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 $a_{0}$. We do not suggest minimizing $\varphi_{\rho}$ in this situation. On the other hand, minimizing the bilinear lasso objective $\varphi_{\text {lasso }}$ over the sphere often succeeds even if the true signal pair $\left(\boldsymbol{a}_{0}, \boldsymbol{x}_{0}\right)$ is coherent and dense.

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 bilinear 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.

Our model assume $x_{0}$ to be Bernoulli-Gaussian vector, which are sparse and symmetric iid random variables. When $\boldsymbol{x}_{0}$ is sparse but non-symmetric, (e.g. Bernoulli), one can apply our result with a simple
symmetrization trick, by using the concatenated observation vectors $[\boldsymbol{y},-\boldsymbol{y}]$ as an input to our algorithm.
When $\boldsymbol{x}_{0}$ is highly sparse and if $\boldsymbol{y}$ is noiseless, it is possible to identify a short copy of $\boldsymbol{a}_{0}$ via looking for a shortest consecutive non-zero entries within $\boldsymbol{y}$. When $\theta \ll 1 / p_{0}$, these isolated copies are very common. Once $\theta$ exceeds $1 / p_{0}$, or when support $x_{0}$ is not Bernoulli random while being more clustered, they become very uncommon. In particular, the probability of an isolated copy is small unless $n \gtrsim \exp \left(p_{0} \theta\right)$. Our proposed approach succeeds when $n \geq \operatorname{poly}\left(p_{0}\right)$.

In applications involving noisy data, optimization approaches often outperform direct inspection, even for samples with isolated copies of $\boldsymbol{a}_{0}$. An intuition for this is that optimization methods aggregate information across the sample. One practical avenue for obtaining the best of both worlds is to try to optimize the choice of data segment used for initialization. This can be a potential improvement for our data-driven initialization scheme, both in theory and in practice.

Finally, there are several directions in which our analysis could be improved. Our lower bounds on the length $n$ of the random vector $x_{0}$ required for success are clearly suboptimal. We also suspect our sparsity-coherence tradeoff between $\mu, \theta$ (roughly, $\theta \lesssim 1 /\left(\sqrt{\mu} p_{0}\right)$ ) is suboptimal, even for the $\varphi_{\rho}$ objective. Articulating optimal sparsity-coherence tradeoffs for is another interesting direction in this line of work. Extending our current result for cases when $\boldsymbol{y}$ is affected by noise can also be a natural next step for future work.

## Chapter 4

## Discussion

Due to the broad application of many field in practice, the associate problem with sparse deconvolution is far beyond the subject discussed in the thesis, and many possible research direction can very likely be branched out from it. Here, we will list some potentially interesting topics, that also has accessed to real world applications, and potentially can be extended and build on from the presented works:

- Convolutional Dictionary Learning. In this scenario, people consider the observation $\boldsymbol{Y}$ consist of linear combination of $k$ convoluting components $\left(\boldsymbol{A}_{i}, \boldsymbol{X}_{i}\right)$, namely $\boldsymbol{Y}=\sum_{i}^{k} \boldsymbol{A}_{i} * \boldsymbol{X}_{i}$, while all of the components $\left\{\left(\boldsymbol{A}_{i}, \boldsymbol{X}_{i}\right)\right\}_{k}$ are unknown. Albiet this can be viewed as a simple extension for short-andsparse deconvolution if the number of pairs $k$ is relatively small, the convolutional dictionary problem is generally way more complicated then SaS deconvolution, and the author believe so far even the existence of stably working of a general purpose algorithm has remained unclear. Nevertheless, it is a highly valuable problem since it can be directly being apply in many fields, and even potentially giving access to study the convolutional neural network.
- Parametric Short-and-Sparse Deconvolution. In many cases, especially in scientific/biological signal processing, the short event signal $\boldsymbol{A}$ can be parametrized base on the understanding of the physi$\mathrm{cal} /$ chemical phenomenon for the event, or the statistic evidence from past data. Parametrize $\boldsymbol{A}$ can potentially increase the solvability of short-and-spare deconvolution, for which we believe deserve a systematic study for this subject since the method is operating in practice in many fields while gaining successes. It is possible that with our study of short-and-sparse deconvolution, a more sophisticated understanding for its parameterized version becomes possible.
- Mixed probe scan in microscopic imaging. In our study of chemical microscopy, we compare the pros and cons between the localized measurement with point probe and the non-local measurement line scans, where the point probe can be general purpose but with inefficient scanning time, while the line probe requires the signal to be structured in order to reconstruct the signal with more efficient scans. A possible study, to gain advantage on both local and non-local measurements, is to combine the multiple type of probe or even active sampling across multiple probe. This will rely on our future understanding for line probe, or even the inventions of other type of non-localized probe.


## Bibliography

[AD88] GR Ayers and J Christopher Dainty. Iterative blind deconvolution method and its applications. Optics letters, 13(7):547-549, 1988.
[AG77] Moshe Abeles and Moise H Goldstein. Multispike train analysis. Proceedings of the IEEE, 65(5):762-773, 1977.
[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.
[BB98] Mario Bertero and Patrizia Boccacci. Introduction to inverse problems in imaging. CRC press, 1998.
[BC11] Heinz H. Bauschke and Patrick L. Combettes. Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer Publishing Company, Incorporated, 1st edition, 2011.
[BCOC14] Marisa Brini, Tito Calì, Denis Ottolini, and Ernesto Carafoli. Neuronal calcium signaling: function and dysfunction. Cellular and molecular life sciences, 71(15):2787-2814, 2014.
[ $\mathrm{BDD}^{+} 12$ ] Peter Binev, Wolfgang Dahmen, Ronald DeVore, Philipp Lamby, Daniel Savu, and Robert Sharpley. Compressed sensing and electron microscopy. In Modeling Nanoscale Imaging in Electron Microscopy, pages 73-126. Springer, 2012.
[BDE09] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM review, 51(1):34-81, 2009.
$\left[\mathrm{BDH}^{+} 13\right]$ 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.
[BFLZ80] Allen J Bard, Larry R Faulkner, Johna Leddy, and Cynthia G Zoski. Electrochemical methods: fundamentals and applications, volume 2. Wiley New York, 1980.
[ $\mathrm{BFP}^{+} 91$ ] Allen J Bard, Fu-Ren F Fan, David T Pierce, Patrick R Unwin, David O Wipf, and Feimeng Zhou. Chemical imaging of surfaces with the scanning electrochemical microscope. Science, 254(5028):68-74, 1991.
[BJT ${ }^{+}$07] Agnes Bogner, P-H Jouneau, Gilbert Thollet, D Basset, and Catherine Gauthier. A history of scanning electron microscopy developments: towards wet-stem imaging. Micron, 38(4):390-401, 2007.
[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.
[BLK15] Carlos A Silvera Batista, Ronald G Larson, and Nicholas A Kotov. Nonadditivity of nanoparticle interactions. Science, 350(6257):1242477, 2015.
[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. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE transactions on image processing, 18(11):24192434, 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):19802007, 2001.
[CE16] Patrizio Campisi and Karen Egiazarian. Blind image deconvolution: theory and applications. CRC press, 2016.
[CE17] Patrizio Campisi and Karen Egiazarian. Blind image deconvolution: theory and applications. CRC press, 2017.
[CFG14a] E. Candes and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. Communications on Pure and Applied Mathematics, 67(6):906-956, 2014.
[CFG14b] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of superresolution. Communications on Pure and Applied Mathematics, 67(6):906-956, 2014.
[Chi16] Yuejie Chi. Guaranteed blind sparse spikes deconvolution via lifting and convex optimization. IEEE Journal of Selected Topics in Signal Processing, 10(4):782-794, June 2016.
[CLC ${ }^{+}$17] Sky Cheung, Yenson Lau, Zhengyu Chen, Ju Sun, Yuqian Zhang, John Wright, and Abhay Pasupathy. Beyond the fourier transform: A nonconvex optimization approach to microscopy analysis. Submitted, 2017.
[CM14] Sunav Choudhary and Urbashi Mitra. Identifiability scaling laws in bilinear inverse problems. arXiv preprint arXiv:1402.2637, 2014.
[CM15] Sunav Choudhary and Urbashi Mitra. Fundamental limits of blind deconvolution part ii: Sparsity-ambiguity trade-offs. arXiv preprint arXiv:1503.03184, 2015.
[Cor63] Allan Macleod Cormack. Representation of a function by its line integrals, with some radiological applications. Journal of applied physics, 34(9):2722-2727, 1963.
[CSL ${ }^{+}$18] Sky C Cheung, John Y Shin, Yenson Lau, Zhengyu Chen, Ju Sun, Yuqian Zhang, John N Wright, and Abhay N Pasupathy. Dictionary learning in fourier transform scanning tunneling spectroscopy. arXiv preprint arXiv:1807.10752, 2018.
[CSL $\left.{ }^{+} 20\right]$ Sky C Cheung, John Y Shin, Yenson Lau, Zhengyu Chen, Ju Sun, Yuqian Zhang, Marvin A Müller, Ilya M Eremin, John N Wright, and Abhay N Pasupathy. Dictionary learning in fourier-transform scanning tunneling spectroscopy. Nature Communications, 11(1):1-11, 2020.
[CT05] Emmanuel Candes and Terence Tao. Decoding by linear programming. arXiv preprint math/0502327, 2005.
[CTL08] Guang-Hong Chen, Jie Tang, and Shuai Leng. Prior image constrained compressed sensing (piccs): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets. Medical physics, 35(2):660-663, 2008.
[CVR14] Subhasis Chaudhuri, Rajbabu Velmurugan, and Renu Rameshan. Blind deconvolution methods: A review. In Blind Image Deconvolution, pages 37-60. Springer, 2014.
[CW98] Tony F Chan and Chiu-Kwong Wong. Total variation blind deconvolution. IEEE transactions on Image Processing, 7(3):370-375, 1998.
[CW08] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]. IEEE signal processing magazine, 25(2):21-30, 2008.
[CWB08] Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted 11 minimization. Journal of Fourier analysis and applications, 14(5-6):877-905, 2008.
[ $\mathrm{D}^{+} 06$ ] David L Donoho et al. Compressed sensing. IEEE Transactions on information theory, 52(4):12891306, 2006.
[DET06] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Transactions on information theory, 52(1):6-18, 2006.
[DKS ${ }^{+}$19] Anna E Dorfi, Han-Wen Kuo, Vera Smirnova, John Wright, and Daniel V Esposito. Design and operation of a scanning electrochemical microscope for imaging with continuous line probes. Manuscript submitted under review, 2019.
[DLB $\left.{ }^{+} 05\right]$ Nebojsa Duric, Peter Littrup, Alex Babkin, David Chambers, Stephen Azevedo, Arkady Kalinin, Roman Pevzner, Mikhail Tokarev, Earle Holsapple, Olsi Rama, et al. Development of ultrasound tomography for breast imaging: Technical assessment. Medical Physics, 32(5):1375-1386, 2005.
[DMWD08] CE DeForest, PCH Martens, and MJ Wills-Davey. Solar coronal structure and stray light in trace. The Astrophysical Journal, 690(2):1264, 2008.
[DNT ${ }^{+}$17] Laurène Donati, Masih Nilchian, Sylvain Trépout, Cédric Messaoudi, Sergio Marco, and Michael Unser. Compressed sensing for stem tomography. Ultramicroscopy, 179:47-56, 2017.
[Don81] David Donoho. On minimum entropy deconvolution. In Applied time series analysis II, pages 565-608. Elsevier, 1981.
[DZC $\left.{ }^{+} 10\right]$ Mark E Davis, Jonathan E Zuckerman, Chung Hang J Choi, David Seligson, Anthony Tolcher, Christopher A Alabi, Yun Yen, Jeremy D Heidel, and Antoni Ribas. Evidence of rnai in humans from systemically administered sirna via targeted nanoparticles. Nature, 464(7291):1067, 2010.
[DZSW11] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and superresolution by adaptive sparse domain selection and adaptive regularization. IEEE Transactions on Image Processing, 20(7):1838-1857, 2011.
$\left.[E H]^{+} 04\right]$ Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. The Annals of statistics, 32(2):407-499, 2004.
[EP10] Yonina C Eldar and Volker Pohl. Recovering signals from lowpass data. IEEE transactions on signal processing, 58(5):2636-2646, 2010.
[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.
[FR17] Simon Foucart and Holger Rauhut. A mathematical introduction to compressive sensing. Bull. Am. Math, 54:151-165, 2017.
[Fra92] Joachim Frank. Electron tomography. Springer, 1992.
[FREM04] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Advances and challenges in super-resolution. International Journal of Imaging Systems and Technology, 14(2):47-57, 2004.
[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.
[FZP17] Johannes Friedrich, Pengcheng Zhou, and Liam Paninski. Fast online deconvolution of calcium imaging data. PLoS computational biology, 13(3):e1005423, 2017.
$[G B V d B+12]$ Bart Goris, Sara Bals, Wouter Van den Broek, Enrique Carbó-Argibay, Sergio Gómez-Graña, Luis M Liz-Marzán, and Gustaaf Van Tendeloo. Atomic-scale determination of surface facets in gold nanorods. Nature materials, 11(11):930, 2012.
[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsâĂŤonline 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.
[GSM06] S Gburek, J Sylwester, and P Martens. The trace telescope point spread function for the 171 å filter. Solar Physics, 239(1-2):531-548, 2006.
[Hel10] Sigurdur Helgason. Integral geometry and Radon transforms. Springer Science \& Business Media, 2010.
[Her09] Gabor T Herman. Fundamentals of computerized tomography: image reconstruction from projections. Springer Science \& Business Media, 2009.
[HHMN13] Vincent C Holmberg, Justin R Helps, K Andre Mkhoyan, and David J Norris. Imaging impurities in semiconductor nanostructures. Chemistry of Materials, 25(8):1332-1350, 2013.
[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.
[HNO06] Per Christian Hansen, James G Nagy, and Dianne P O'leary. Deblurring images: matrices, spectra, and filtering, volume 3. Siam, 2006.
[Hou73] Godfrey N Hounsfield. Computerized transverse axial scanning (tomography): Part 1. description of system. The British journal of radiology, 46(552):1016-1022, 1973.
[Jen10] Grant Jensen. Cryo-EM, Part C: analyses, interpretation, and case studies. Academic Press, 2010.
[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.
[Kak79] Avinash C Kak. Computerized tomography with x-ray, emission, and ultrasound sources. Proceedings of the IEEE, 67(9):1245-1272, 1979.
[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.
[KKS17] Felix Krahmer, Christian Kruschel, and Michael Sandbichler. Total variation minimization in compressed sensing. In Compressed Sensing and its Applications, pages 333-358. Springer, 2017.
[KMR14] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. Communications on Pure and Applied Mathematics, 67(11):1877-1904, 2014.
[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.
[LBC99] Steven J Ludtke, Philip R Baldwin, and Wah Chiu. Eman: semiautomated software for highresolution single-particle reconstructions. Journal of structural biology, 128(1):82-97, 1999.
[LDSP08] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. IEEE signal processing magazine, 25(2):72, 2008.
[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.
[LOK97] Kieran G Larkin, Michael A Oldfield, and Hanno Klemm. Fast fourier method for the accurate rotation of sampled images. Optics communications, 139(1-3):99-106, 1997.
[LQK $\left.{ }^{+} 19\right]$ Yenson Lau, Qing Qu, Han-Wen Kuo, Pengcheng Zhou, Yuqian Zhang, and John Wright. Short-and-sparse deconvolution-a geometric approach. arXiv preprint arXiv:1908.10959, 2019.
[LS00] Lei Li and Terence P Speed. Parametric deconvolution of positive spike trains. Annals of Statistics, pages 1279-1301, 2000.
[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.
[LSMH13] Rowan Leary, Zineb Saghi, Paul A Midgley, and Daniel J Holland. Compressed sensing electron tomography. Ultramicroscopy, 131:70-91, 2013.
[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.
[Mal13] Krzysztof Malczewski. Pet image reconstruction using compressed sensing. In 2013 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), pages 176-181. IEEE, 2013.
[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.
[Nat01] Frank Natterer. The Mathematics of Computerized Tomography. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001.
[NdLPL+13] Olivia Nicoletti, Francisco de La Peña, Rowan K Leary, Daniel J Holland, Caterina Ducati, and Paul A Midgley. Three-dimensional imaging of localized surface plasmon resonances of metal nanoparticles. Nature, 502(7469):80, 2013.
[NDMD $\left.{ }^{+} 98\right]$ John Nuyts, Bruno De Man, Patrick Dupont, Michel Defrise, Paul Suetens, and Luc Mortelmans. Iterative reconstruction for helical ct: a simulation study. Physics in Medicine E Biology, 43(4):729, 1998.
[NG10] Patrick A Naylor and Nikolay D Gaubitch. Speech dereverberation. Springer Science \& Business Media, 2010.
[ $\mathrm{ODB}^{+}$15] Marie Engelene J Obien, Kosmas Deligkaris, Torsten Bullmann, Douglas J Bakkum, and Urs Frey. Revealing neuronal function through microelectrode array recordings. Frontiers in neuroscience, 8:423, 2015.
[OKL $\left.{ }^{+} 18\right]$ Glen D ONeil, Han-Wen Kuo, Duncan N Lomax, John Wright, and Daniel V Esposito. Scanning line probe microscopy: Beyond the point probe. Analytical chemistry, 90(19):11531-11537, 2018.
[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.
$\left[\mathrm{PAF}^{+} 04\right]$ Dale Purves, George J Augustine, David Fitzpatrick, William C Hall, Anthony-Samuel LaMantia, James O McNamara, and S Mark Williams. Neuroscience. Sunderland, MA: Sinauer Associates, 773, 2004.
[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.
[PS16] Thomas Pock and Shoham Sabach. Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems. SIAM Journal on Imaging Sciences, 9(4):1756-1787, 2016.
[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.
[PZBF ${ }^{+}$07] Praveen Pankajakshan, Bo Zhang, Laure Blanc-Féraud, Zvi Kam, Jean-Christophe Olivo-Marin, and Josiane Zerubia. Parametric blind deconvolution for confocal laser scanning microscopy. In 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 6531-6534. IEEE, 2007.
[Rad05] Johann Radon. 1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. Classic papers in modern diagnostic radiology, 5:21, 2005.
[Rei13] Ludwig Reimer. Transmission electron microscopy: physics of image formation and microanalysis, volume 36. Springer, 2013.
[Rud06] Walter Rudin. Real and complex analysis. Tata McGraw-Hill Education, 2006.
[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.
$\left[\mathrm{SBC}^{+} 12\right]$ Vincent Studer, Jérome Bobin, Makhlad Chahid, Hamed Shams Mousavi, Emmanuel Candes, and Maxime Dahan. Compressive fluorescence microscopy for biological and hyperspectral imaging. Proceedings of the National Academy of Sciences, 109(26):E1679-E1687, 2012.
[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.
[SHL $\left.{ }^{+} 11\right]$ Zineb Saghi, Daniel J Holland, Rowan Leary, Andrea Falqui, Giovanni Bertoni, Andrew J Sederman, Lynn F Gladden, and Paul A Midgley. Three-dimensional morphology of iron oxide nanoparticles with reactive concave surfaces. a compressed sensing-electron tomography (cs-et) approach. Nano letters, 11(11):4666-4673, 2011.
[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.
[SKS ${ }^{+}$09] Shigeru Shinomoto, Hideaki Kim, Takeaki Shimokawa, Nanae Matsuno, Shintaro Funahashi, Keisetsu Shima, Ichiro Fujita, Hiroshi Tamura, Taijiro Doi, Kenji Kawano, et al. Relating neuronal firing patterns to functional differentiation of cerebral cortex. PLoS computational biology, 5(7), 2009.
[SL74] Lawrence A Shepp and Benjamin F Logan. The fourier reconstruction of a head section. IEEE Transactions on nuclear science, 21(3):21-43, 1974.
[SNRS ${ }^{+}$08] Sjors HW Scheres, Rafael Núñez-Ramírez, Carlos OS Sorzano, José María Carazo, and Roberto Marabini. Image processing for electron microscopy single-particle analysis using xmipp. Nature protocols, 3(6):977, 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.
[Tót14] L Fejes Tóth. Regular figures. Elsevier, 2014.
[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027, 2010.
[VJJW ${ }^{+}$16] Jairo Velasco Jr, Long Ju, Dillon Wong, Salman Kahn, Juwon Lee, Hsin-Zon Tsai, Chad Germany, Sebastian Wickenburg, Jiong Lu, Takashi Taniguchi, et al. Nanoscale control of rewriteable doping patterns in pristine graphene/boron nitride heterostructures. Nano letters, 16(3):16201625, 2016.
[VRR11] Ashok Veeraraghavan, Dikpal Reddy, and Ramesh Raskar. Coded strobing photography: Compressive sensing of high speed periodic videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(4):671-686, 2011.
[WC16] Liming Wang and Yuejie Chi. Blind deconvolution from multiple sparse inputs. IEEE Signal Processing Letters, 23(10):1384-1388, 2016.
[WDW85] Kenneth R Wehmeyer, Mark R Deakin, and R Mark Wightman. Electroanalytical properties of band electrodes of submicrometer width. Analytical Chemistry, 57(9):1913-1916, 1985.
[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.
[WR94] Roland Wiesendanger and Wiesendanger Roland. Scanning probe microscopy and spectroscopy: methods and applications. Cambridge university press, 1994.
[WT14] Ruxin Wang and Dacheng Tao. Recent progress in image deblurring. arXiv preprint arXiv:1409.6838, 2014.
$\left[W V^{+}\right.$87] Scott L Wellington, Harold J Vinegar, et al. X-ray computerized tomography. Journal of Petroleum Technology, 39(08):885-898, 1987.
[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.
[ZN09] Changyin Zhou and Shree Nayar. What are good apertures for defocus deblurring? In 2009 IEEE international conference on computational photography (ICCP), pages 1-8. IEEE, 2009.
[ZZW13] Liuyang Zhang, Xiaowei Zeng, and Xianqiao Wang. Programmable hydrogenation of graphene for novel nanocages. Scientific reports, 3(1):1-8, 2013.

## Appendix A

## Appendix: Compressed Sensing

## Microscopy with Scanning Line Probe

## A. 1 Analytic Derivations

## A.1.1 Proof of Theorem 2.4.1

Proof. First we show the existence result. Note that $\boldsymbol{X}_{0}$ satisfies the equalitiy constraint (2.8) automatically, and since total variation of Dirac measure is exactly one,

$$
\begin{aligned}
\int\left|\boldsymbol{X}_{0}\right|\left(d \boldsymbol{w}^{\prime}\right) & =\sum_{\boldsymbol{w}} \int\left|\alpha_{\boldsymbol{w}}\right| \boldsymbol{\delta}_{\boldsymbol{w}}\left(d \boldsymbol{w}^{\prime}\right)=\sum_{\boldsymbol{w}}\left|\alpha_{\boldsymbol{w}}\right| \\
& =\sum_{\boldsymbol{w}} \alpha_{\boldsymbol{w}} \cdot \boldsymbol{D} * \mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{Q}}](\boldsymbol{w})
\end{aligned}
$$

then since $\boldsymbol{D}$ is circular symmetric, $\boldsymbol{D} * \mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{Q}}](\boldsymbol{w})=\left\langle\boldsymbol{\delta}_{\boldsymbol{w}}, \boldsymbol{D} * \mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{Q}}]\right\rangle=\left\langle\mathcal{L}_{\Theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}}\right], \widetilde{\boldsymbol{Q}}\right\rangle$, we derive

$$
\int\left|\boldsymbol{X}_{0}\right|\left(d \boldsymbol{w}^{\prime}\right)=\left\langle\mathcal{L}_{\Theta}\left[\boldsymbol{D} * \sum_{\boldsymbol{w}} \alpha_{\boldsymbol{w}} \boldsymbol{\delta}_{\boldsymbol{w}}\right], \widetilde{\boldsymbol{Q}}\right\rangle=\langle\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{Q}}\rangle
$$

which certifies that $\boldsymbol{X}_{0}$ is an optimal solution to the problem since the duality gap $\int\left|\boldsymbol{X}_{0}\right|\left(d \boldsymbol{w}^{\prime}\right)-\langle\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{Q}}\rangle=0$. For uniqueness, let $\boldsymbol{X}^{\prime}=\sum_{\boldsymbol{w}^{\prime} \in \mathcal{W}^{\prime}} \alpha_{\boldsymbol{w}^{\prime}}^{\prime} \boldsymbol{\delta}_{\boldsymbol{w}^{\prime}}$ to be another optimal solution with $\mathcal{W}^{\prime} \nsubseteq \mathcal{W}$, since we know $\boldsymbol{X}^{\prime}$ is primal feasible $\widetilde{\boldsymbol{R}}=\mathcal{L}_{\Theta}\left[\boldsymbol{D} * \boldsymbol{X}^{\prime}\right]$, then

$$
\begin{aligned}
\int\left|\boldsymbol{X}_{0}\right|\left(d \boldsymbol{w}^{\prime}\right) & =\langle\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{Q}}\rangle=\left\langle\mathcal{L}_{\Theta}\left[\boldsymbol{D} * \boldsymbol{X}^{\prime}\right], \widetilde{\boldsymbol{Q}}\right\rangle \\
& =\left\langle\boldsymbol{X}^{\prime}, \boldsymbol{D} * \mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{Q}}]\right\rangle
\end{aligned}
$$

$$
=\sum_{\boldsymbol{w}^{\prime} \in \mathcal{W}^{\prime}} \alpha_{\boldsymbol{w}^{\prime}}^{\prime} \boldsymbol{D} * \mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{Q}}]\left(\boldsymbol{w}^{\prime}\right)
$$

and by knowing $\mathcal{W}^{\prime} \nsubseteq \mathcal{W}$ and using the second condition in (2.6):

$$
\int\left|\boldsymbol{X}_{0}\right|\left(d \boldsymbol{w}^{\prime}\right)<\sum_{\boldsymbol{w}^{\prime} \in \mathcal{W}^{\prime}}\left|\alpha_{\boldsymbol{w}^{\prime}}^{\prime}\right|=\int\left|\boldsymbol{X}^{\prime}\right|\left(d \boldsymbol{w}^{\prime}\right)
$$

thus $\boldsymbol{X}^{\prime}$ is an optimal solution only if $\mathcal{W}^{\prime} \subseteq \mathcal{W}$. Finally uniqueness of $\boldsymbol{X}_{0}$ is a result from injectivity of $\mathcal{L}_{\Theta}[\boldsymbol{D} * \cdot]$ over $\mathcal{W}$ from (2.7).

## A.1.2 Proof of Theorem 2.4.2

Proof. We first argue that with high probability, no pair of discs overlaps within any line scan. Let $\theta_{i} \sim_{\text {i.i.d. }}$ Unif $[-\pi, \pi)$ denote the $i$-th scanning angle. Write $d$ as the minimum distance between all pairs of $\left(\boldsymbol{w}_{i}, \boldsymbol{w}_{j}\right)$, the probability that any particular pair of two discs overlap is bounded as

$$
\begin{align*}
& \mathbb{P}\left[\text { Two discs overlap on line scan } \widetilde{\boldsymbol{R}}_{i}\right] \\
& \quad \leq \mathbb{P}\left[\theta_{i} \in\left[-\sin ^{-1}\left(\frac{2 r}{d}\right), \sin ^{-1}\left(\frac{2 r}{d}\right)\right]\right] \\
&  \tag{A.1}\\
& \quad=\frac{2}{\pi} \sin ^{-1} \frac{2 r}{d}
\end{align*}
$$

Using the assumption that $R<\frac{d}{8}$ to bound $\sin ^{-1}\left(\frac{2 r}{d}\right)<\frac{2 \pi r}{3 d}$ and summing the failure probability over all three line scans and $\frac{k(k-1)}{2}$ pairs of discs, we obtain:

$$
\begin{align*}
\mathbb{P} & {\left[\text { Two of the } k \text { discs overlap at either } \widetilde{\boldsymbol{R}}_{1}, \widetilde{\boldsymbol{R}}_{2}, \widetilde{\boldsymbol{R}}_{3}\right] } \\
& \leq \frac{3 k^{2}}{2} \cdot \mathbb{P}\left[\text { Two discs overlap on line scan } \widetilde{\boldsymbol{R}}_{1}\right] \\
& \leq \frac{3 k^{2}}{\pi} \sin ^{-1}\left(\frac{2 r}{d}\right) \leq \frac{2 k^{2} r}{d} \\
& \leq C \tag{A.2}
\end{align*}
$$

Thus, with probability at least $1-C$, no pair of discs overlaps in any line scan.
Since there are no overlapping discs in any line, a single line projection $\widetilde{\boldsymbol{R}}_{i}(t)$ with scan angle $\theta_{i}$ has largest magnitude at points $t$ where the probe body passes the disc center $\boldsymbol{w}_{j}$. These points of largest magnitude $\beta_{j}$ is located at $\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}_{j}\right\rangle$ on $\widetilde{\boldsymbol{R}}_{i}$, or equivalently,

$$
\begin{equation*}
\mathcal{L}_{\theta_{i}}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\left(\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}_{j}\right\rangle\right)=\beta_{j}, \quad i=1,2,3 \tag{A.3}
\end{equation*}
$$

Using these points, we construct the dual certificate $\widetilde{\boldsymbol{Q}}_{i}$ for angle $\theta_{i}$, where

$$
\begin{equation*}
\widetilde{\boldsymbol{Q}}_{i}=\sum_{j=1}^{k} \frac{1}{\sqrt{3} \beta_{j}} \boldsymbol{\delta}_{\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}_{j}\right\rangle} \tag{A.4}
\end{equation*}
$$

and $\widetilde{\boldsymbol{Q}}=\left[\widetilde{\boldsymbol{Q}}_{1}, \widetilde{\boldsymbol{Q}}_{2}, \widetilde{\boldsymbol{Q}}_{3}\right]$. Using this certificate we verify 2.6 holds. For the equality, calculate at every $\boldsymbol{w}_{j} \in\left\{\boldsymbol{w}_{1}, \ldots \boldsymbol{w}_{k}\right\}:$

$$
\begin{align*}
& \boldsymbol{D} * \mathcal{L}_{\left\{\theta_{1}, \theta_{2}, \theta_{3}\right\}}^{*}[\widetilde{\boldsymbol{Q}}]\left(\boldsymbol{w}_{j}\right) \\
= & \left\langle\boldsymbol{D} * \mathcal{L}_{\left\{\theta_{1}, \theta_{2}, \theta_{3}\right\}}^{*}[\widetilde{\boldsymbol{Q}}], \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right\rangle=\frac{1}{\sqrt{3}} \sum_{i=1}^{3}\left\langle\widetilde{\boldsymbol{Q}}_{i}, \mathcal{L}_{\theta_{i}}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\right\rangle \\
= & \frac{1}{\sqrt{3}} \sum_{i=1}^{3}\left\langle\frac{1}{\sqrt{3} \beta_{j}} \boldsymbol{\delta}_{\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}_{j}\right\rangle}, \mathcal{L}_{\theta_{i}}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\right\rangle \\
= & \frac{1}{3 \beta_{j}} \sum_{i=1}^{3} \mathcal{L}_{\theta_{i}}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\left(\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}_{j}\right\rangle\right)=1 \tag{A.5}
\end{align*}
$$

where the third line is by plugging in $\widetilde{\boldsymbol{Q}}$ and derived with no overlap property; the forth line via plugging in A.3. For the inequality, calculate

$$
\begin{align*}
& \left|\boldsymbol{D} * \mathcal{L}_{\left\{\theta_{1}, \theta_{2}, \theta_{3}\right\}}^{*}[\widehat{\boldsymbol{Q}}](\boldsymbol{w})\right| \\
& \quad=\left|\sum_{i=1}^{3} \sum_{j=1}^{k} \frac{1}{3 \beta_{j}} \mathcal{L}_{\theta_{i}}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}}\right]\left(\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}_{j}\right\rangle\right)\right| \tag{A.6}
\end{align*}
$$

which is derived similarly as A.5. Now, by observing $\mathcal{L}_{\theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}}\right]$ has unique local maximum at $\left\langle\boldsymbol{u}_{\theta}^{\perp}, \boldsymbol{w}\right\rangle$, each summand (w.r.t. $i$ ) in A.6 is strictly less than 1 if $\boldsymbol{w}$ does not satisfy

$$
\begin{equation*}
\exists j \in\{1, \ldots, k\}, \quad\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}\right\rangle=\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}_{j}\right\rangle . \tag{A.7}
\end{equation*}
$$

Accordingly, define the back projection line set $\ell_{\theta_{i}, t_{j}}$ as

$$
\begin{equation*}
\ell_{\theta_{i}, t_{j}}:=\left\{\boldsymbol{w} \in \mathbb{R}^{2} \mid\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}\right\rangle=\left\langle\boldsymbol{u}_{\theta_{i}}^{\perp}, \boldsymbol{w}_{j}\right\rangle\right\} \tag{A.8}
\end{equation*}
$$

we want to show that for every $\boldsymbol{w} \notin\left\{\boldsymbol{w}_{1}, \ldots, \boldsymbol{w}_{k}\right\}$,

$$
\begin{equation*}
\boldsymbol{w} \notin \cap_{i=1}^{3}\left(\cup_{j=1}^{k} \ell_{\theta_{i}, t_{j}}\right) \tag{A.9}
\end{equation*}
$$

then A. 6 is strictly less then 1.
W.l.o.g., write $\boldsymbol{w}_{j \ell}=\ell_{\theta_{1}, t_{j}} \cap \ell_{\theta_{2}, t_{\ell}}$. Suppose the point $\boldsymbol{w}_{j \ell}$ is in the third line set $\cup_{j=1}^{k} \ell_{\theta_{3}, t_{j}}$, then there exists some disc center $\boldsymbol{w}_{q}$ such that $\left\langle\boldsymbol{u}_{\theta_{3}}^{\perp}, \boldsymbol{w}_{j \ell}\right\rangle=\left\langle\boldsymbol{u}_{\theta_{3}}^{\perp}, \boldsymbol{w}_{q}\right\rangle$. Since $\theta_{3}$ is generated uniform randomly, we
conclude that for any $j, \ell$ :

$$
\begin{equation*}
\mathbb{P}\left[\exists q \in 1, \ldots, k \quad \text { s.t. } \quad \boldsymbol{w}_{j \ell} \in \ell_{\theta_{3}, t_{q}}\right]=0 \tag{A.10}
\end{equation*}
$$

The direction $\boldsymbol{u}_{\theta_{3}}$ is not aligned with the line formed by $\boldsymbol{w}_{j \ell}, \boldsymbol{w}_{q}$ almost surely. This proves A.9.
Finally, the diagonal entries of Gram matrix $G$ defined in 2.7 is strictly positive, and the off-diagonal entries $\boldsymbol{G}_{i j}$ can be derived as

$$
\begin{equation*}
\boldsymbol{G}_{i j}=\frac{1}{3} \sum_{t=1}^{3}\left\langle\mathcal{L}_{\theta_{t}}[\boldsymbol{D}] * \boldsymbol{\delta}_{\boldsymbol{w}_{i}}, \mathcal{L}_{\theta_{t}}[\boldsymbol{D}] * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right\rangle=0 \tag{A.11}
\end{equation*}
$$

by no overlap property. Hence $G$ is positive definite. This concludes that solving total variation minimization successfully reconstruct the image from three line projections.

## A.1.3 Proof of Theorem 2.4.3

Proof. Write $\boldsymbol{d}(t)=\mathcal{L}_{0}[\boldsymbol{D}](t)$, where $\boldsymbol{d}$ is a one-dimensional standard Gaussian with deviation $r$. Since $\boldsymbol{D}$ is circular symmetric, the line projection of $\boldsymbol{D}$ in any angle is identical, that is, $\mathcal{L}_{\theta}[\boldsymbol{D}]=\mathcal{L}_{0}[\boldsymbol{D}]$ for every $\theta$. Also write $\boldsymbol{w}_{i}-\boldsymbol{w}_{j}=d(\cos \phi, \sin \phi)$, then

$$
\begin{align*}
&\left\langle\mathcal{L}_{\theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{i}}\right], \mathcal{L}_{\theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\right\rangle \\
&=\left\langle\mathcal{L}_{\theta}[\boldsymbol{D}] * \mathcal{L}_{\theta}\left[\boldsymbol{\delta}_{\boldsymbol{w}_{i}}\right], \mathcal{L}_{\theta}[\boldsymbol{D}] * \mathcal{L}_{\theta}\left[\boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\right\rangle \\
&=\left\langle\boldsymbol{d} * \boldsymbol{d}, \boldsymbol{\delta}_{\left|\boldsymbol{u}_{\theta}^{*}\left(\boldsymbol{w}_{i}-\boldsymbol{w}_{j}\right)\right|}\right\rangle \\
&=(\boldsymbol{d} * \boldsymbol{d})(d \cos (\theta-\phi)) \\
&=\exp \left(\frac{-d^{2} \cos ^{2}(\theta-\phi)}{4 r^{2}}\right) \tag{A.12}
\end{align*}
$$

where the first equality is by interchanging iterated integrals; the second equality is by knowing the adjoint of convolution is correlation and $\boldsymbol{d}$ is symmetric; and the final equality is by observing that $\boldsymbol{d} * \boldsymbol{d}$ is a Gaussian function with variance $\sqrt{2} r$ and $(\boldsymbol{d} * \boldsymbol{d})(0)=1$.

We derive the expectation upper bound of A.12 over $\theta$ as

$$
\begin{aligned}
\mathbb{E}_{\theta}\left\langle\mathcal{L}_{\theta}[\boldsymbol{D}\right. & \left.\left.* \boldsymbol{\delta}_{\boldsymbol{w}_{i}}\right], \mathcal{L}_{\theta}\left[\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}_{j}}\right]\right\rangle \\
& \left.=\frac{1}{\pi} \int_{0}^{\pi} \exp \left(-d^{2} \cos ^{2} \theta\right) / 4 r^{2}\right) d \theta \\
& \leq \frac{1}{\pi} \int_{0}^{\pi} \frac{1}{1+\left(d^{2} \cos ^{2} \theta\right) / 4 r^{2}} d \theta
\end{aligned}
$$

$$
\begin{equation*}
=\frac{1}{\sqrt{1+d^{2} / 4 r^{2}}} \tag{A.13}
\end{equation*}
$$

by utilizing $\exp \left(-x^{2}\right)\left(1+x^{2}\right)<1$ in the second inequality.
As for the lower bound of A.12, from its first equality, we calculate when $d>2 r$, then

$$
\begin{align*}
& \frac{1}{\pi} \int_{0}^{\pi} \\
& \quad \exp \left(-\left(d^{2} \cos ^{2} \theta\right) / 4 r^{2}\right) d \theta \\
& \quad \geq \frac{1}{\pi} \int_{0}^{\pi} \exp \left(-\left(d^{2} / 4 r^{2}\right) \cdot\left(2 r^{2} / d^{2}\right)\right) \cdot \mathbf{1}_{\left\{\cos ^{2} \theta \leq 2 r^{2} / d^{2}\right\}} d \theta \\
& \quad \geq \frac{2}{\pi} \cdot \exp \left(-\frac{1}{2}\right) \cdot\left(\frac{\pi}{2}-\cos ^{-1} \sqrt{2 r^{2} / d^{2}}\right) \\
& \quad \geq \frac{2}{\pi} \cdot \exp \left(-\frac{1}{2}\right) \cdot(\sqrt{2} r / d)  \tag{A.14}\\
& \quad \geq r / 2 d
\end{align*}
$$

using $\cos ^{-1} x \leq \frac{\pi}{2}-x$ for $x \in[0,0.5]$. And when $d \leq 2 r$, we simply have

$$
\begin{equation*}
\frac{1}{\pi} \int_{0}^{\pi} \exp \left(-\left(d^{2} \cos ^{2} \theta\right) / 4 r^{2}\right) d \theta \geq 1-d^{2} / 8 r^{2} \tag{A.15}
\end{equation*}
$$

via Taylor expansion at $d / 2 r=0$

## A.1.4 Proof of Theorem 2.4.4

Proof. We start with restating projection slice theorem as $\mathcal{F}_{1} \mathcal{L}_{\theta}[\boldsymbol{Y}]=\mathcal{S}_{\theta}\left[\mathcal{F}_{2} \boldsymbol{Y}\right]$, where $\mathcal{F}_{1}, \mathcal{F}_{2}$ are unitary Fourier transform in one, two-dimensional Euclidean space respectively, $\mathcal{S}_{\theta}$ is the slice operator defined as $\mathcal{S}_{\theta}[\boldsymbol{Y}](r)=\boldsymbol{Y}\left(r \boldsymbol{u}_{\theta}^{\perp}\right)$ Hel10].

Notice that $\boldsymbol{Y}=\boldsymbol{D} * \boldsymbol{X} \in L^{1} \cap L^{2}\left(\mathbb{R}^{2}\right)$ therefore its Fourier transform is well defined, we expand $\mathcal{L}_{\theta}^{*} \mathcal{L}_{\theta}$ in Fourier domain with slice operator $\mathcal{S}_{\theta}$, write $\widehat{\boldsymbol{Y}}=\mathcal{F}_{2} \boldsymbol{Y}$. and derive

$$
\begin{aligned}
\mathbb{E}_{\theta} & \mathcal{L}_{\theta}^{*} \mathcal{L}_{\theta}[\boldsymbol{Y}](\boldsymbol{w}) \\
& =\mathbb{E}_{\theta} \mathcal{F}_{2}^{*} \mathcal{S}_{\theta}^{*} \mathcal{F}_{1}^{-1 *} \mathcal{F}_{1}^{-1} \mathcal{S}_{\theta} \mathcal{F}_{2} \boldsymbol{Y}(\boldsymbol{w}) \\
& =\mathbb{E}_{\theta} \int_{\boldsymbol{\xi} \in \mathbb{R}^{2}} \exp (j 2 \pi\langle\boldsymbol{\xi}, \boldsymbol{w}\rangle) \cdot \mathcal{S}_{\theta}^{*}\left[\mathcal{S}_{\theta}[\widehat{\boldsymbol{Y}}]\right](\boldsymbol{\xi}) d \boldsymbol{\xi} \\
& =\mathbb{E}_{\theta} \int_{t \in \mathbb{R}} \mathcal{S}_{\theta}[\exp (j 2 \pi\langle\cdot, \boldsymbol{w}\rangle)](t) \cdot \mathcal{S}_{\theta}[\widehat{\boldsymbol{Y}}](t) d t \\
& =\frac{1}{2 \pi} \int_{\theta=0}^{2 \pi} \int_{t \in \mathbb{R}} \exp \left(j 2 \pi t\left\langle\boldsymbol{u}_{\theta}^{\perp}, \boldsymbol{w}\right\rangle\right) \cdot \widehat{\boldsymbol{Y}}\left(t \boldsymbol{u}_{\theta}^{\perp}\right) d t d \theta \\
& =\frac{2}{2 \pi} \int_{\theta=0}^{2 \pi} \int_{t \geq 0} \exp \left(j 2 \pi t\left\langle\boldsymbol{u}_{\theta}^{\perp}, \boldsymbol{w}\right\rangle\right) \cdot \widehat{\boldsymbol{Y}}\left(t \boldsymbol{u}_{\theta}^{\perp}\right) d t d \theta
\end{aligned}
$$

$$
\begin{align*}
& =\int_{\boldsymbol{\xi} \in \mathbb{R}^{2}} \exp (j 2 \pi t\langle\boldsymbol{\xi}, \boldsymbol{w}\rangle) \cdot\left(\frac{1}{\pi\|\boldsymbol{\xi}\|_{2}}\right) \cdot \widehat{\boldsymbol{Y}}(\boldsymbol{\xi}) d \boldsymbol{\xi} \\
& =\left(\mathcal{F}_{2}^{-1}\left\{\frac{1}{\pi\|\boldsymbol{\xi}\|_{2}}\right\} * \boldsymbol{Y}\right)(\boldsymbol{w}) \tag{A.16}
\end{align*}
$$

where the third equality is derived from definition of adjoint operator, sixth equality is by coordinate transformation from polar to Cartesian, and the last equality is by convolution theorem. Hence we conclude that $\mathbb{E}_{\theta} \mathcal{L}_{\theta}^{*} \mathcal{L}_{\theta}[\boldsymbol{Y}]$ is the convolution between $\boldsymbol{Y}$ and a lowpass kernel with spectrum decay rate $\|\boldsymbol{\xi}\|_{2}^{-1}$.

When $\left\|\mathcal{L}_{0}[\boldsymbol{D}]\right\|_{L^{2}}=1$ and is a Gaussian function with deviation $r$, then $\boldsymbol{D}(\boldsymbol{w})=\frac{\sqrt{2 r \sqrt{\pi}}}{2 \pi r^{2}} \exp \left(-\frac{\|\boldsymbol{w}\|_{2}^{2}}{2 r^{2}}\right)$ with Fourier domain expression as $\mathcal{F}_{2}\{\boldsymbol{D}\}(\boldsymbol{\xi})=\sqrt{2 r \sqrt{\pi}} \exp \left(-2 \pi^{2} r^{2}\|\boldsymbol{\xi}\|_{2}^{2}\right)$. Combine with A.16, the spectrum of $\mathbb{E}_{\theta} \boldsymbol{D} * \mathcal{L}_{\theta}^{*} \mathcal{L}_{\theta}[\boldsymbol{D} * \cdot]$ becomes

$$
\begin{align*}
& \mathcal{F}_{2}\left\{\mathbb{E}_{\theta} \boldsymbol{D} * \mathcal{L}_{\theta}^{*} \mathcal{L}_{\theta}[\boldsymbol{D} * \boldsymbol{X}]\right\}(\boldsymbol{\xi}) \\
& \quad=\frac{2 r}{\sqrt{\pi}\|\boldsymbol{\xi}\|_{2}} \exp \left(-4 \pi^{2} r^{2}\|\boldsymbol{\xi}\|_{2}^{2}\right) \cdot \mathcal{F}_{2}\{\boldsymbol{X}\}(\boldsymbol{\xi}) \\
& \quad=\mathcal{F}_{2}\{\mathcal{K}\}(\boldsymbol{\xi}) \cdot \mathcal{F}_{2}\{\boldsymbol{X}\}(\boldsymbol{\xi}) \tag{A.17}
\end{align*}
$$

Plug in 2.12, when $\|\boldsymbol{\xi}\|_{2} \geq \frac{2 r}{\varepsilon}$ then clearly $\left|\mathcal{F}_{2}\{\mathcal{K}\}(\boldsymbol{\xi})\right| \leq \varepsilon$. Lastly for the other lower bound $\|\boldsymbol{\xi}\|_{2} \geq$ $\frac{1}{r}\left(\sqrt{\left|\log \left(8 r^{2} \varepsilon^{-1}\right)\right|}+0.2\right)$, we calculate

$$
\begin{align*}
\left|\mathcal{F}_{2}\{\mathcal{K}\}(\boldsymbol{\xi})\right| & \leq \frac{2 r^{2}}{0.2 \sqrt{\pi}} \cdot \exp \left(-4 \pi^{2}\left|\log \left(8 r^{2} \varepsilon^{-1}\right)\right|\right) \\
& \leq \frac{2 r^{2}}{0.2 \sqrt{\pi}} \cdot \frac{1}{8 r^{2} \varepsilon^{-1}} \leq \varepsilon \tag{A.18}
\end{align*}
$$

## A.1.5 Proof of Theorem 2.5.1

Proof. For any lines $\boldsymbol{R} \in L^{2}(\mathbb{R} \times[m])$, image $\boldsymbol{Y} \in L^{2}\left(\mathbb{R}^{2}\right)$, and any angles $\Theta=\left\{\theta_{1}, \ldots, \theta_{m}\right\}$,

$$
\begin{aligned}
\left\langle\widetilde{\boldsymbol{R}}, \mathcal{L}_{\Theta}[\boldsymbol{Y}]\right\rangle & =\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \int \widetilde{\boldsymbol{R}}_{i}(t) \mathcal{L}_{\theta_{i}}[\boldsymbol{Y}](t) d t \\
& =\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \int \widetilde{\boldsymbol{R}}_{i}(t) \int \boldsymbol{Y}\left(s \boldsymbol{u}_{\theta_{i}}+t \boldsymbol{u}_{\theta_{i}}^{\perp}\right) d s d t \\
& =\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \int \widetilde{\boldsymbol{R}}_{i}\left(\left\langle\boldsymbol{w}, \boldsymbol{u}_{\theta_{i}}^{\perp}\right\rangle\right) \boldsymbol{Y}(\boldsymbol{w}) d \boldsymbol{w} \\
& =\int\left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \widetilde{\boldsymbol{R}}_{i}\left(\left\langle\boldsymbol{w}, \boldsymbol{u}_{\theta_{i}}^{\perp}\right\rangle\right)\right) \boldsymbol{Y}(\boldsymbol{w}) d \boldsymbol{w}
\end{aligned}
$$

$$
\begin{equation*}
=\left\langle\mathcal{L}_{\Theta}^{*}[\widetilde{\boldsymbol{R}}], \boldsymbol{Y}\right\rangle . \tag{A.19}
\end{equation*}
$$

The first equality comes from the definition of inner product in lines space; the second comes from (2.2); the third uses change of variable where $\boldsymbol{w}=s \boldsymbol{u}_{\theta}+t \boldsymbol{u}_{\theta}^{\perp}$ for every $\theta$; the fourth comes from linearity; and the last equality from definition of inner product in image space.

## Appendix B

## Appendix: Short-and-Sparse <br> Deconvolution

## B. 1 Basic bounds for Bernoulli-Gaussian vectors

In this section, we prove several lemmas pertaining to the sparse random vector $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \mathrm{BG}(\theta)$.
Lemma B.1.1 (Support of $\boldsymbol{x}_{0}$ ). Let $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ and $I_{0}=\operatorname{supp}\left(\boldsymbol{x}_{0}\right) \subseteq[n]$. Suppose $n>10 \theta^{-1}$, then for any $\varepsilon \in\left(0, \frac{1}{10}\right)$, with probability at least $1-\varepsilon$ we have

$$
\begin{equation*}
\left|\left|I_{0}\right|-n \theta\right| \leq 2 \sqrt{n \theta} \log \varepsilon^{-1} \tag{B.1}
\end{equation*}
$$

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

$$
\begin{equation*}
\forall t \in[2 p] \backslash\{0\}, \quad \frac{1}{2} n \theta^{2} \leq\left|I_{0} \cap\left(I_{0}+t\right)\right| \leq 2 n \theta^{2} \tag{B.2}
\end{equation*}
$$

where $C$ is a numerical constant.

Proof. Let $\boldsymbol{x}_{0}=\boldsymbol{\omega} \cdot \boldsymbol{g} \sim_{\text {i.i.d. }} \mathrm{BG}(\theta)$, notice that the support of the Bernoulli-Gaussian vector $\boldsymbol{x}_{0}$ is almost surely equal to the support of the Bernoulli vector $\omega$. Applying Bernstein inequality Theorem B.10.4 with $\left(\sigma^{2}, R\right)=(1,1)$, then if $n \theta>10$ we have

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

For (B.2), let $J_{t}:=I_{0} \cap\left(I_{0}+t\right)$. The cardinality of $J_{t}$ is an inner product between shifts of $\omega$ :

$$
\begin{equation*}
\left|J_{t}\right|=\sum_{k \in[n]} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k-t}, \tag{B.3}
\end{equation*}
$$

and define two subset $J_{t 1} \uplus J_{t 2}=J_{t}$, as follows:

$$
\left\{\begin{array}{ll}
J_{t 1}=J_{t} \cap \mathcal{K}_{1}, & \mathcal{K}_{1}:=[n] \cap\{0, \ldots, t-1,2 t, \ldots, 3 t-1, \ldots\}  \tag{B.4}\\
J_{t 2}=J_{t} \cap \mathcal{K}_{2}, & \mathcal{K}_{2}:=[n] \cap\{t, \ldots, 2 t-1,3 t, \ldots, 4 t-1, \ldots\}
\end{array} .\right.
$$

Here, the size of sets $\mathcal{K}_{1}, \mathcal{K}_{2}$ has two-side bounds $0.4 n \leq(n-2 p) / 2 \leq\left|\mathcal{K}_{2}\right| \leq\left|\mathcal{K}_{1}\right| \leq(n+2 p) / 2 \leq 0.6 n$, thus the size of sets $J_{t 1}, J_{t 2}$ can be derived using Bernstein inequality Theorem B.10.4 with $n>C \theta^{-2} \log p$ as

$$
\begin{align*}
\mathbb{P}\left[\max _{t \in[2 p] \backslash\{0\}}\left|J_{t_{1}}\right| \geq n \theta^{2}\right] & =\mathbb{P}\left[\max _{t \in[2 p] \backslash\{0\}} \sum_{k \in \mathcal{K}_{1}} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k-t} \geq n \theta^{2}\right] \leq 2 p \cdot \mathbb{P}\left[\sum_{k \in \mathcal{K}_{1}} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k+1} \geq n \theta^{2}\right] \\
& \leq 2 p \cdot \mathbb{P}\left[\sum_{k \in \mathcal{K}_{1}} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k+1}-\mathbb{E} \sum_{k \in \mathcal{K}_{1}} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k+1} \geq n \theta^{2}-0.6 n \theta^{2}\right] \\
& \leq 4 p \cdot \exp \left(\frac{-\left(0.4 n \theta^{2}\right)^{2}}{2 \cdot 0.6 n \theta^{2}+2 \cdot 0.4 n \theta^{2}}\right)=\exp \left(\log (4 p)-0.08 n \theta^{2}\right) \leq 1 / n \tag{B.5}
\end{align*}
$$

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

$$
\begin{align*}
\mathbb{P}\left[\min _{t \in[2 p] \backslash\{0\}}\left|J_{t_{1}}\right| \leq n \theta^{2} / 4\right] & =\mathbb{P}\left[\min _{t \in[2 p] \backslash\{0\}} \sum_{k \in \mathcal{K}_{1}} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k-t} \leq n \theta^{2} / 4\right] \leq 2 p \cdot \mathbb{P}\left[\sum_{k \in \mathcal{K}_{1}} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k+1} \leq n \theta^{2} / 4\right] \\
& \leq 2 p \cdot \mathbb{P}\left[\sum_{k \in \mathcal{K}_{1}} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k+1}-\mathbb{E} \sum_{k \in \mathcal{K}_{1}} \boldsymbol{\omega}_{k} \boldsymbol{\omega}_{k+1} \leq n \theta^{2} / 4-0.4 n \theta^{2}\right] \\
& \leq 4 p \cdot \exp \left(\frac{-\left(0.15 n \theta^{2}\right)^{2}}{2 \cdot 0.6 n \theta^{2}+2 \cdot 0.15 n \theta^{2}}\right)=\exp \left(\log (4 p)-0.0015 n \theta^{2}\right) \leq 1 / n \tag{B.6}
\end{align*}
$$

The bound for $\left|J_{2}\right|$ can derived similarly to $(\boxed{\text { B.5 }}-\sqrt{\text { B.6 }}$.

Lemma B.1.2 (Norms of $\boldsymbol{x}_{0}$ ). Let $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta) \in \mathbb{R}^{n}$. If $n \geq 10 \theta^{-1}$, then for any $\varepsilon \in\left(0, \frac{1}{10}\right)$, with probability at least $1-\varepsilon$,

$$
\begin{equation*}
\left|\left\|\boldsymbol{x}_{0}\right\|_{1}-\sqrt{2 / \pi} n \theta\right| \leq 2 \sqrt{n \theta} \log \varepsilon^{-1}, \quad\left|\left\|\boldsymbol{x}_{0}\right\|_{2}^{2}-n \theta\right| \leq 3 \sqrt{n \theta} \log \varepsilon^{-1} \tag{B.7}
\end{equation*}
$$

Proof. To bound $\left\|\boldsymbol{x}_{0}\right\|_{1}$, using Bernstein inequality with $\left(\sigma^{2}, R\right)=(\theta, 1)$ and with $n \theta \geq 10$ we have

$$
\mathbb{P}\left[\left|\left\|\boldsymbol{x}_{0}\right\|_{1}-\sqrt{\frac{2}{\pi}} n \theta\right| \geq 2 \sqrt{n \theta} \log \varepsilon^{-1}\right] \leq 2 \exp \left(\frac{-4 n \theta \log ^{2} \varepsilon^{-1}}{2 n \theta+4 \sqrt{n \theta} \log \varepsilon^{-1}}\right) \leq \varepsilon
$$

Similarly for $\left\|\boldsymbol{x}_{0}\right\|_{2}^{2}$, from Gaussian moments Theorem B.10.2, we know the 2-norm $\sum_{i \in[n]} \mathbb{E}\left|x_{0 i}\right|^{4}=3 n \theta$
and $q$-norm $\sum_{i \in[n]} \mathbb{E}\left|x_{0 i}\right|^{2 p} \leq(n \theta)(2 q-1)!!\leq \frac{1}{2}(3 n \theta) 2^{q-2} q$ ! for $q \geq 3$. Let $\left(\sigma^{2}, R\right)=(3 \theta, 2)$ in Bernstein inequality form Theorem B.10.4, $n \theta \geq 10$ we have

$$
\mathbb{P}\left[\left|\left\|x_{0}\right\|_{2}^{2}-n \theta\right| \geq 3 \sqrt{n \theta} \log \varepsilon^{-1}\right] \leq 2 \exp \left(\frac{-9 n \theta \log ^{2} \varepsilon^{-1}}{2(3 n \theta)+12 \sqrt{n \theta} \log \varepsilon^{-1}}\right) \leq \varepsilon
$$

completing the proof.

Lemma B.1.3 (Norms of $\boldsymbol{x}_{0}$ subvectors). Let $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta) \in \mathbb{R}^{n}$ and $n>10$, then with probability at least $1-3 / n$, we have

$$
\begin{equation*}
\max _{\substack{U=[2 p]+j \\ j \in[n]}}\left\|\boldsymbol{P}_{U} \boldsymbol{x}_{0}\right\|_{2}^{2} \leq 2 p \theta+6(\sqrt{p \theta}+\log n) \tag{B.8}
\end{equation*}
$$

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

$$
\begin{equation*}
\max _{\substack{U=[p]+j \\ j \in[n]}}\left\|\boldsymbol{P}_{U}\left[\boldsymbol{a}_{0} * \boldsymbol{x}_{0}\right]\right\|_{2}^{2} \leq p \theta+\log n \tag{B.9}
\end{equation*}
$$

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

$$
\begin{align*}
\mathbb{P}\left[\max _{\substack{U=[2 p]+j \\
j \in[n]}}\left\|\boldsymbol{P}_{U} \boldsymbol{x}_{0}\right\|_{2}^{2} \geq 2 p \theta+6(\sqrt{p \theta}+\log n)\right] & \leq 2 n \exp \left(-\frac{36(\sqrt{p \theta}+\log n)^{2}}{6 p \theta+12(\sqrt{p \theta}+\log n)}\right) \\
& \leq 2 \exp \left(\log n-\frac{36 t^{2}}{6 t^{2}+12 t}\right) \leq \frac{2}{n} \tag{B.10}
\end{align*}
$$

For the second inequality, first we know calculate the expectation

$$
\begin{align*}
\mathbb{E}\left\|\boldsymbol{P}_{U}\left[\boldsymbol{a}_{0} * \boldsymbol{x}_{0}\right]\right\|_{2}^{2} & =\mathbb{E}\left[\boldsymbol{x}_{0}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{P}_{U} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}_{0}\right] \\
& =\theta \cdot \operatorname{tr}\left(\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{P}_{U} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right)\left\|\boldsymbol{a}_{0}\right\|_{2}^{2}+\theta \cdot \sum_{i=1}^{p-1}\left\|\boldsymbol{\iota}^{*} s_{i}\left[\boldsymbol{a}_{0}\right]\right\|_{2}^{2} \\
& =p \theta \tag{B.11}
\end{align*}
$$

Then apply Henson Wright inequality Theorem B.10.6 with $\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{P}_{U} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right\|_{F}^{2}=\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}\right\|_{F}^{2} \leq p(1+\mu p)$ and also $\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{P}_{U} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right\|_{2}=\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}\right\|_{2}^{2}=1+\mu p$, we can derive

$$
\begin{align*}
\mathbb{P}\left[\max _{\substack{U=[p]+j \\
j \in[n]}}\left\|\boldsymbol{P}_{U}\left[\boldsymbol{a}_{0} * \boldsymbol{x}_{0}\right]\right\|_{2}^{2} \geq p \theta+\log n\right] & \leq n \exp \left(-\min \left\{\frac{\log ^{2} n}{64 \theta^{2} p(1+\mu p)}, \frac{\log n}{8 \sqrt{2} \theta(1+\mu p)}\right\}\right) \\
& \leq \exp \left(\log n-\min \left\{\frac{\log ^{2} n}{128 c_{\mu}}, \frac{\log n}{32 c_{\mu}}\right\}\right) \leq \frac{1}{n} \tag{B.12}
\end{align*}
$$

when $c_{\mu}<\frac{1}{300}$.

Lemma B.1.4 (Inner product between shifted $\boldsymbol{x}_{0}$ ). Let $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta) \in \mathbb{R}^{n}$. 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:

$$
\begin{equation*}
\max _{i \neq j \in[2 p]}\left\langle s_{i}\left[\boldsymbol{x}_{0}\right], s_{j}\left[\boldsymbol{x}_{0}\right]\right\rangle \leq 6 \sqrt{n \theta^{2} \log n} \tag{B.13}
\end{equation*}
$$

and for $\boldsymbol{x}_{i}=\left|\boldsymbol{x}_{0, i}\right| \in \mathbb{R}_{+}^{n}$ the vector of magnitudes of $\boldsymbol{x}_{0}$,

$$
\begin{equation*}
\max _{i \neq j \in[2 p]}\left\langle s_{i}[\boldsymbol{x}], s_{j}[\boldsymbol{x}]\right\rangle \leq 4 n \theta^{2} \tag{B.14}
\end{equation*}
$$

Proof. We will start from proving $\overline{\mathrm{B} .14}$. 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 (B.14) can be written as summation of Gaussian r.v.s. on intersection of support set between shifts:

$$
\begin{equation*}
\left\langle s_{i}[\boldsymbol{x}], s_{j}[\boldsymbol{x}]\right\rangle=\sum_{k \in I_{0} \cap\left(I_{0}+t\right)}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k-t}\right| \tag{B.15}
\end{equation*}
$$

Define $J_{t}:=I_{0} \cap\left(I_{0}+t\right)=J_{t 1} \uplus J_{t 2}$ same as (B.4). Notice that both $\sum_{k \in J_{t 1}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k-t}\right|$ and $\sum_{k \in J_{t 2}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k-t}\right|$ are sum of independent r.v.s.. We are left to consider the upper bound of $\sum_{j \in J_{t i}}\left|\boldsymbol{g}_{j}\right|\left|\boldsymbol{g}_{j}^{\prime}\right|$ where $\boldsymbol{g}, \boldsymbol{g}^{\prime}$ are independent Gaussian vectors. We condition on the following event

$$
\begin{equation*}
\mathcal{E}_{J}:=\left\{\forall t \in[2 p] \backslash\{0\}, n \theta^{2} / 4 \leq\left|J_{t 1}\right|,\left|J_{t 2}\right| \leq n \theta^{2}\right\}, \tag{B.16}
\end{equation*}
$$

which holds w.p. at least $1-2 / n$ from Theorem B.1.1. Since $\sum_{j \in J_{t 1}}\left|\boldsymbol{g}_{j}\right|\left|\boldsymbol{g}_{j}^{\prime}\right| \leq\left\|\boldsymbol{g}_{J_{t 1}}\right\|_{2}\left\|\boldsymbol{g}_{J_{t 1}}^{\prime}\right\|_{2}$, we use Gaussian concentration Theorem B.10.3 and union bound to obtain

$$
\begin{align*}
\mathbb{P}\left[\max _{t \in[2 p] \backslash\{0\}} \sum_{j \in J_{t 1}}\left|\boldsymbol{g}_{j} \boldsymbol{g}_{j}^{\prime}\right|>2\left|J_{t 1}\right|\right] & \leq 2 p \cdot \mathbb{P}\left[\left\|\boldsymbol{g}_{J_{t 1}}\right\|_{2}\left\|\boldsymbol{g}_{J_{t 1}}^{\prime}\right\|_{2}-\mathbb{E}\left\|\boldsymbol{g}_{J_{t 1}}\right\|_{2}\left\|\boldsymbol{g}_{J_{t 1}}^{\prime}\right\|_{2}>\left|J_{t 1}\right|\right] \\
& \leq 4 p \cdot \mathbb{P}\left[\left\|\boldsymbol{g}_{J_{t 1}}\right\|_{2}-\mathbb{E}\left\|\boldsymbol{g}_{J_{t 1}}\right\|_{2}>\sqrt{\left|J_{t 1}\right|} / 3\right] \\
& \leq 4 p \exp \left(-\left(\left|J_{t 1}\right| / 9\right) / 2\right) \leq 4 p \exp \left(-n \theta^{2} / 72\right) \leq 1 / n \tag{B.17}
\end{align*}
$$

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

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

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

$$
\begin{equation*}
\left\langle s_{i}[\boldsymbol{x}], s_{j}[\boldsymbol{x}]\right\rangle=\sum_{k \in J_{t 1}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k-t}^{\prime}\right|+\sum_{k \in J_{t 2}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k-t}^{\prime}\right| \leq 2\left(\left|J_{t_{1}}\right|+\left|J_{t_{2}}\right|\right) \leq 4 n \theta^{2} . \tag{B.18}
\end{equation*}
$$

For (B.13) similarly condition on event $\mathcal{E}_{J}$, using Bernstein inequality Theorem B.10.4 with $\left(\sigma^{2}, R\right)=(1,1)$ :

$$
\begin{equation*}
\mathbb{P}\left[\max _{t \in[2 p] \backslash\{0\}}\left|\sum_{j \in J_{t 1}} \boldsymbol{g}_{j} \boldsymbol{g}_{j}^{\prime}\right|>3 \sqrt{n \theta^{2} \log n}\right] \leq p \cdot \exp \left(\frac{-9 n \theta^{2} \log n}{2\left|J_{t 1}\right|+6 \sqrt{n \theta^{2} \log n}}\right) \leq p \cdot \exp \left(\frac{-9 n \theta^{2} \log n}{3 n \theta^{2}}\right) \leq \frac{1}{n} \tag{B.19}
\end{equation*}
$$

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

$$
\begin{equation*}
\left|\left\langle s_{i}\left[\boldsymbol{x}_{0}\right], s_{j}\left[s_{0}\right]\right\rangle\right| \leq\left|\sum_{k \in J_{t 1}} \boldsymbol{g}_{k} \boldsymbol{g}_{k-t}^{\prime}\right|+\left|\sum_{k \in J_{t 2}} \boldsymbol{g}_{k} \boldsymbol{g}_{k-t}^{\prime}\right| \leq 6 \sqrt{n \theta^{2} \log n} \tag{B.20}
\end{equation*}
$$

Finally, both B.18, B.20 holds simultaneously with probability at least

$$
\begin{equation*}
1-2 / n-1 / n-1 / n=1-4 / n \tag{B.21}
\end{equation*}
$$

Lemma B.1.5 (Convolution of $\boldsymbol{x}_{0}$ ). Given $\boldsymbol{y}=\boldsymbol{x}_{0} * \boldsymbol{a}_{0}$ where $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta) \in \mathbb{R}^{n}$ 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:

$$
\begin{equation*}
\left\|\boldsymbol{C}_{\boldsymbol{y}} \ell\right\|_{2}^{2} \leq 3(1+\mu p) n \theta \tag{B.22}
\end{equation*}
$$

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

$$
\begin{equation*}
\left\|\boldsymbol{P}_{J} \boldsymbol{C}_{\boldsymbol{y}} \iota\right\|_{2}^{2} \leq 14|J|(1+\mu p)(p \theta+\log n) \tag{B.23}
\end{equation*}
$$

Proof. Given any $\boldsymbol{a} \in \mathbb{S}^{p-1}$, write $\boldsymbol{\beta}=\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \iota \boldsymbol{a}$ where $|\boldsymbol{\beta}| \leq 2 p$. Apply $\left\|\boldsymbol{x}_{0}\right\|_{2}^{2} \leq 2 n \theta$ from Theorem B.1.2 by choosing $\varepsilon=1 / n$, also $\left|\left\langle s_{i}\left[\boldsymbol{x}_{0}\right], s_{j}\left[\boldsymbol{x}_{0}\right]\right\rangle\right| \leq 6 \sqrt{n \theta^{2} \log n}$ fromTheorem B.1.4 we get:

$$
\begin{aligned}
\left\|\boldsymbol{C}_{\boldsymbol{y}} \boldsymbol{\iota}\right\|_{2}^{2} & =\left\|\boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\beta}\right\|_{2}^{2} \leq\|\boldsymbol{\beta}\|_{2}^{2}\left\|\boldsymbol{x}_{0}\right\|_{2}^{2}+\sum_{i \neq j \in[ \pm p]}\left|\beta_{i} \beta_{j}\left\langle s_{i}\left[\boldsymbol{x}_{0}\right], s_{j}\left[\boldsymbol{x}_{0}\right]\right\rangle\right| \\
& \leq\|\boldsymbol{\beta}\|_{2}^{2}\left\|\boldsymbol{x}_{0}\right\|_{2}^{2}+\|\boldsymbol{\beta}\|_{1}^{2} \max _{i \neq j \in[ \pm p]}\left|\left\langle s_{i}\left[\boldsymbol{x}_{0}\right], s_{j}\left[\boldsymbol{x}_{0}\right]\right\rangle\right| \\
& \leq\|\boldsymbol{\beta}\|_{2}^{2} \cdot 2 n \theta+p\|\boldsymbol{\beta}\|_{2}^{2} \cdot 6 \sqrt{n \theta^{2} \log n} \leq 3\|\boldsymbol{\beta}\|_{2}^{2} n \theta
\end{aligned}
$$

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 $\left\|\boldsymbol{P}_{J} \boldsymbol{C}_{\boldsymbol{y}} \iota \boldsymbol{\iota}\right\|_{2}^{2}$. Simply apply Theorem B.1.3 and utilize norm bound of $\|\boldsymbol{\beta}\|_{2}^{2}$, with probability at least $1-2 / n$ we have:

$$
\left\|\boldsymbol{P}_{J} \boldsymbol{C}_{\boldsymbol{y}} \boldsymbol{\iota} \boldsymbol{a}\right\|_{2}^{2}=\sum_{i \in J}\left|\left\langle s_{i}\left[\boldsymbol{x}_{0}\right], \boldsymbol{\beta}\right\rangle\right|^{2} \leq|J| \max _{\substack{U=[2 p]+j \\ j \in[n]}}\left\|\boldsymbol{P}_{U} \boldsymbol{x}_{0}\right\|_{2}^{2}\|\boldsymbol{\beta}\|_{2}^{2} \leq|J| \cdot 14(p \theta+\log n) \cdot\|\boldsymbol{\beta}\|_{2}^{2}
$$

Finally apply Theorem B.2.4 and Gershgorin disc theorem obtain

$$
\begin{equation*}
\|\boldsymbol{\beta}\|_{2}^{2}=\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \iota \boldsymbol{\iota}\right\|_{2}^{2} \leq\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota}\right\|_{2}^{2}=\sigma_{\max }(\boldsymbol{M}) \leq 1+\mu p . \tag{B.24}
\end{equation*}
$$

Remark B.1.6. When $\boldsymbol{a}_{0}$ is a basis vector $\boldsymbol{e}_{0}$, the result of Theorem B.1.5 gives upper bound of $\left\|\boldsymbol{C}_{\boldsymbol{x}_{0}}\right\|_{2}<3 n \theta$, whose lower bound can be derived similarly with $\left\|\boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\right\|_{2} \geq \frac{2}{3} n \theta$

## B. 2 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 Theorem B.2.2 we prove that measuring the distance from $\boldsymbol{a}$ to subspace $\mathcal{S}_{\boldsymbol{\tau}}$ in terms of $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}$ gives a seminorm. We then use this distance to characterize a region $\mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ around the subspace $\mathcal{S}_{\boldsymbol{\tau}}$. Later, in Theorem B.2.4 we illustrate the relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\beta}=\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}$. Finally in Theorem B.2.5 and Theorem B.2.6 controls the magnitude of $\boldsymbol{\alpha}_{\boldsymbol{\tau}}$ and $\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}$ near $\mathcal{S}_{\boldsymbol{\tau}}$.

Definition B.2.1 (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 $\operatorname{SCC}\left(c_{\mu}\right)$ with constant $c_{\mu}$, if

$$
\begin{equation*}
\theta \in\left[\frac{1}{p}, \frac{c_{\mu}}{4 \max \{|\boldsymbol{\tau}|, \sqrt{p}\}}\right] \cdot \frac{1}{\log ^{2} \theta^{-1}}, \quad \mu \cdot \max \left\{|\boldsymbol{\tau}|^{2}, p^{2} \theta^{2}\right\} \cdot \log ^{2} \theta^{-1} \leq \frac{c_{\mu}}{4}, \tag{B.25}
\end{equation*}
$$

where $p=3 p_{0}-2$.

Lemma B.2.2 $\left(d_{\alpha}\right.$ is a seminorm). For every solution subspace $\mathcal{S}_{\boldsymbol{\tau}}$, the function $d_{\alpha}\left(\cdot, \mathcal{S}_{\boldsymbol{\tau}}\right): \mathbb{R}^{p} \rightarrow \mathbb{R}_{+}$defined as

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)=\inf \left\{\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \mid \boldsymbol{a}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}\right\} \tag{B.26}
\end{equation*}
$$

is a seminorm, and for all $\boldsymbol{a} \in \mathcal{S}_{\boldsymbol{\tau}}, d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)=0$.
Proof. It is immediate from definition that $d\left(\cdot, \mathcal{S}_{\boldsymbol{\tau}}\right)$ is nonnegative and $\mathcal{S}_{\boldsymbol{\tau}} \subseteq\left\{\boldsymbol{a}: d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)=0\right\}$. Subadditivity can be shown from simple norm inequalities and our definition of $d_{\alpha}$, for all $\boldsymbol{a}_{1}$, $\boldsymbol{a}_{2}$ we have

$$
\begin{aligned}
d_{\alpha}\left(\boldsymbol{a}_{1}+\boldsymbol{a}_{2}, \mathcal{S}_{\boldsymbol{\tau}}\right) & =\inf \left\{\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \mid \boldsymbol{a}_{1}+\boldsymbol{a}_{2}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}\right\} \\
& =\inf \left\{\left\|\boldsymbol{\alpha}_{1 \boldsymbol{\tau}^{c}}+\boldsymbol{\alpha}_{2 \boldsymbol{\tau}^{c}}\right\|_{2} \mid \boldsymbol{a}_{1}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}_{1}, \quad \boldsymbol{a}_{2}=\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}_{2}\right\} \\
& \leq \inf \left\{\left\|\boldsymbol{\alpha}_{1 \boldsymbol{\tau}^{c}}\right\|_{2}+\left\|\boldsymbol{\alpha}_{2 \boldsymbol{\tau}^{c}}\right\|_{2} \mid \boldsymbol{a}_{1}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}_{1}, \quad \boldsymbol{a}_{2}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{a_{0}} \boldsymbol{\alpha}_{2}\right\} \\
& =\inf \left\{\left\|\boldsymbol{\alpha}_{1 \boldsymbol{\tau}^{c}}\right\|_{2} \mid \boldsymbol{a}_{1}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}_{1}\right\}+\inf \left\{\left\|\boldsymbol{\alpha}_{2 \boldsymbol{\tau}^{c}}\right\|_{2} \mid \boldsymbol{a}_{2}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}_{2}\right\} \\
& =d_{\alpha}\left(\boldsymbol{a}_{1}, \mathcal{S}_{\boldsymbol{\tau}}\right)+d_{\alpha}\left(\boldsymbol{a}_{2}, \mathcal{S}_{\boldsymbol{\tau}}\right)
\end{aligned}
$$

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

$$
\begin{aligned}
d_{\alpha}\left(c \cdot \boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) & =\inf \left\{\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}^{\prime}\right\|_{2} \mid c \cdot \boldsymbol{a}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}^{\prime}\right\}=\inf \left\{\left\|c \cdot \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \mid \boldsymbol{a}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}\right\} \\
& =|c| \cdot \inf \left\{\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \mid \boldsymbol{a}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}\right\}=|c| \cdot d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)
\end{aligned}
$$

which completes the proof that $d_{\alpha}$ is a seminorm.
Definition B.2.3 (Widened subspace). For subspace $\mathcal{S}_{\tau}$ let

$$
\begin{equation*}
\mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right):=\left\{\boldsymbol{a} \in \mathbb{S}^{p-1} \mid d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \gamma\right\} \tag{B.27}
\end{equation*}
$$

denote its widening by $\gamma$, in the seminorm $d_{\alpha}$.
Our analysis works with a specific choice of width $\gamma\left(c_{\mu}\right)$, which depends on the problem parameters $\boldsymbol{a}_{0}, \theta,|\tau|$ and a constant $c_{\mu}$, via

$$
\begin{equation*}
\gamma\left(c_{\mu}\right)=\frac{c_{\mu}}{4 \log ^{2} \theta^{-1}} \min \left\{\frac{1}{\sqrt{|\boldsymbol{\tau}|}}, \frac{1}{\sqrt{\mu p}}, \frac{1}{\mu p \sqrt{\theta}|\boldsymbol{\tau}|}\right\} \tag{B.28}
\end{equation*}
$$

Lemma B.2.4 (Properties of $\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota} \boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}$ ). Let $\boldsymbol{M}=\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota} \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}$, with $\boldsymbol{a}_{0} \in \mathbb{S}^{p_{0}-1} \mu$-shift coherent. The diagonal
entries of $\boldsymbol{M}$ satisfy

$$
\begin{cases}\boldsymbol{M}_{i i}=1 & i \in\left[-p_{0}+1, p_{0}-1\right]=\left[ \pm p_{0}\right]  \tag{B.29}\\ 0 \leq \boldsymbol{M}_{i i} \leq 1 & i \in\left[-2 p_{0}+2,-p_{0}\right] \cup\left[p_{0}, 2 p_{0}-2\right] \\ \boldsymbol{M}_{i i}=0 & \text { otherwise },\end{cases}
$$

and the off-diagonal entries satisfy

$$
\begin{cases}\left|\boldsymbol{M}_{i j}\right| \leq \mu & 0<|i-j|<p_{0}, \quad\left\{i \in\left[-p_{0}+1, p_{0}-1\right]\right\} \cup\left\{j \in\left[-p_{0}+1, p_{0}-1\right]\right\}  \tag{B.30}\\ \left|\boldsymbol{M}_{i j}\right|<1 & \left\{i, j \in\left[-2 p_{0}+2,-p_{0}\right]\right\} \cup\left\{i, j \in\left[p_{0}, 2 p_{0}-2\right]\right\} \\ 0 & \text { otherwise. }\end{cases}
$$

Furthermore, let $\boldsymbol{\tau} \subset\left[ \pm p_{0}\right]$, and $\boldsymbol{\tau}^{c}=\left[ \pm 2 p_{0}-1\right] \backslash \boldsymbol{\tau}$. The singular values of submatrix $\boldsymbol{\iota}_{\boldsymbol{\tau}}^{*} \boldsymbol{M} \boldsymbol{\iota}_{\boldsymbol{\tau}}$ can be bounded as:

$$
\left\{\begin{array}{l}
1-\mu|\boldsymbol{\tau}| \leq \sigma_{\min }\left(\iota_{\boldsymbol{\tau}}^{*} \boldsymbol{M} \iota_{\boldsymbol{\tau}}\right) \leq \sigma_{\max }\left(\iota_{\boldsymbol{\tau}}^{*} \boldsymbol{M} \iota_{\boldsymbol{\tau}}\right) \leq 1+\mu|\boldsymbol{\tau}|  \tag{B.31}\\
\sigma_{\max }\left(\boldsymbol{\iota}_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{M} \iota_{\boldsymbol{\tau}}\right) \leq \mu \sqrt{p|\boldsymbol{\tau}|} \\
\sigma_{\max }\left(\boldsymbol{\iota}_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{M} \boldsymbol{\iota}_{\boldsymbol{\tau}^{c}}\right) \leq 1+\mu p
\end{array}\right.
$$

Proof. Recall the definition of $\iota$, which selects the entries $\left\{-p_{0}+1, \ldots, 2 p_{0}-2\right\}$. The entrywise properties of $\boldsymbol{M}$ can be derived by carefully counting the entries of the shifted support. The submatrix $\boldsymbol{M}$ on support
$\left\{-2 p_{0}+2, \ldots, 2 p_{0}-2\right\}$ has an upper bound to be characterized as follows:

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

As for the singular values, notice that the first and second inequalities consider submatrix not containing $\boldsymbol{J}$ since $\boldsymbol{\tau} \subseteq\left[ \pm p_{0}\right]$; thus the first inequality can be derived with Gershgorin disc theorem directly, and the second inequality with the upper bound with its Frobenius norm:

$$
\begin{equation*}
\sigma_{\max }\left(\iota_{\boldsymbol{\tau}_{c}^{c}}^{*} \boldsymbol{M} \boldsymbol{\iota}_{\boldsymbol{\tau}}\right) \leq \mu \sqrt{\left(2 p_{0}-1\right)|\boldsymbol{\tau}|}<\mu \sqrt{p|\boldsymbol{\tau}|} . \tag{B.33}
\end{equation*}
$$

Finally by recalling $p=3 p_{0}-2>2 p_{0}-1$. The last inequality is direct from bound of $\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}$ :

$$
\begin{equation*}
\sigma_{\max }\left(\iota_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{M} \iota_{\boldsymbol{\tau}^{c}}\right) \leq\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \iota^{*} \boldsymbol{\iota}_{\boldsymbol{a}_{0}}\right\|_{2}=\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*}\right\|_{2}=\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}\right\|_{2} \leq 1+\mu p \tag{B.34}
\end{equation*}
$$

where the third equality is derived via commutativity of convolution.
Lemma B.2.5 (Shift space vectors in widened subspace). Let $\left(\boldsymbol{a}_{0}, \theta,|\boldsymbol{\tau}|\right)$ satisfy the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. Then for every $\boldsymbol{a} \in \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$, every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{a}=\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}$ and $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \gamma\left(c_{\mu}\right)$ has

$$
\begin{equation*}
\left|\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2}-1\right| \leq c_{\mu} ; \tag{B.35}
\end{equation*}
$$

moreover, $\boldsymbol{\beta}=\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota} \boldsymbol{a}$ satisfies
$1-3 c_{\mu} \leq\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2} \leq 1+\frac{c_{\mu}}{|\boldsymbol{\tau}| \log ^{2} \theta^{-1}}, \quad\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{\infty} \leq \frac{c_{\mu}}{\sqrt{|\boldsymbol{\tau}|} \log ^{2} \theta^{-1}}, \quad\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \frac{c_{\mu}}{|\boldsymbol{\tau}| \theta \log \theta^{-1}} \min \left\{\sqrt{\theta}, \gamma\left(c_{\mu}\right)\right\}$.

Proof. Write $-1 / \log \theta=\theta_{\log }$ and $\gamma=\gamma\left(c_{\mu}\right)$ for convenience. First, by using bounds on $\gamma$ in (B.28) and $\mu|\boldsymbol{\tau}|<1$ we obtain:

$$
\left\{\begin{array}{l}
\gamma \cdot \sqrt{1+\mu p} \leq \gamma(1+\sqrt{\mu p}) \leq c_{\mu} \theta_{\log }^{2} / 2  \tag{B.37}\\
\gamma \cdot \sqrt{1+\mu^{2} p} \leq \gamma\left(1+\sqrt{\mu^{2} p}\right) \leq \frac{c_{\mu} \theta_{\log }^{2}}{4}\left(\frac{1}{\sqrt{|\boldsymbol{\tau}|}}+\sqrt{\mu}\right) \leq \frac{c_{\mu} \theta_{\log }^{2}}{2 \sqrt{|\boldsymbol{\tau}|}} \\
\gamma \cdot \mu \sqrt{p|\boldsymbol{\tau}|} \leq \gamma \cdot \sqrt{\mu p} \cdot \sqrt{\mu|\boldsymbol{\tau}|} \leq c_{\mu} \theta_{\log }^{2} / 4
\end{array}\right.
$$

Let $\boldsymbol{a}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}$ with $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}<\gamma$. Utilize properties of $\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}$ from Theorem B.2.4 and $\mu|\boldsymbol{\tau}|<c_{\mu} / 4$ and (B.37), we have:

$$
\begin{align*}
\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2} & \geq\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}_{\boldsymbol{\tau}}\right\|_{2}^{-1}\left(\|\boldsymbol{a}\|_{2}-\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\right) \geq\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}_{\boldsymbol{\tau}}\right\|_{2}^{-1}\left(1-\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right\|_{2}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\right) \\
& \geq \frac{1}{\sqrt{1+\mu|\boldsymbol{\tau}|}}(1-\gamma \cdot \sqrt{1+\mu p}) \geq \frac{1-c_{\mu} / 2}{\sqrt{1+c_{\mu} / 4}} \geq 1-c_{\mu} \tag{B.38}
\end{align*}
$$

and similarly, the upper bound can be derived as:

$$
\begin{align*}
\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2} & \leq \sigma_{\min }^{-1}\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}_{\boldsymbol{\tau}}\right)\left(\|\boldsymbol{a}\|_{2}+\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\right) \leq \sigma_{\min }^{-1}\left(\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}_{\boldsymbol{\tau}}\right)\left(1+\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right\|_{2}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\right) \\
& \leq \frac{1}{\sqrt{1-\mu|\boldsymbol{\tau}|}}(1+\gamma \cdot \sqrt{1+\mu p}) \leq \frac{1+c_{\mu} / 2}{\sqrt{1-c_{\mu} / 4}} \leq 1+c_{\mu} . \tag{B.39}
\end{align*}
$$

The bound of $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}$ can be simply obtained using $\mu|\boldsymbol{\tau}|<c_{\mu} / 4$ and $\gamma$ bound from (B.37) as:

$$
\begin{align*}
\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2} & \leq \sigma_{\max }^{2}\left(\iota_{\tau}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}\right) \leq 1+\mu|\boldsymbol{\tau}| \leq 1+\frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}  \tag{B.40}\\
\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2} & \geq\left(\sigma_{\min }\left(\iota_{\boldsymbol{\tau}}^{*} \boldsymbol{M} \iota_{\boldsymbol{\tau}}\right)\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2}-\sigma_{\max }\left(\iota_{\tau}^{*} \boldsymbol{M} \boldsymbol{\iota}_{\boldsymbol{\tau}^{c}}\right)\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\right)^{2} \\
& \geq\left((1-\mu|\boldsymbol{\tau}|)\left(1-c_{\mu}\right)-\mu \sqrt{p|\boldsymbol{\tau}|} \cdot \gamma\right)^{2} \geq 1-3 c_{\mu} . \tag{B.41}
\end{align*}
$$

As for the upper bound of and $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{\infty}$, follow from (B.37), we have:

$$
\begin{align*}
\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{\infty} & \leq\left\|\boldsymbol{\iota}_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{M} \boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{\infty}+\left\|\boldsymbol{\iota}_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{M} \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{\infty} \leq \mu \sqrt{|\boldsymbol{\tau}|}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2}+\sqrt{1+\mu^{2} p}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \\
& \leq \frac{c_{\mu} \theta_{\log }^{2}\left(1+c_{\mu}\right)}{4|\boldsymbol{\tau}|}+\gamma \cdot \sqrt{1+\mu^{2} p} \leq \frac{c_{\mu} \theta_{\log }^{2}}{\sqrt{|\boldsymbol{\tau}|}} \tag{B.42}
\end{align*}
$$

the bound for $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}$ requires two inequalities, we know

$$
\begin{equation*}
\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq\left\|\boldsymbol{\iota}_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{M} \boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2}+\left\|\boldsymbol{\iota}_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{M} \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \mu \sqrt{p|\boldsymbol{\tau}|}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2}+(1+\mu p)\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \tag{B.43}
\end{equation*}
$$

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

$$
\begin{align*}
(\overline{\mathrm{B} .43} & \leq \frac{\mu \sqrt{p \theta}|\boldsymbol{\tau}|^{3 / 2}}{\sqrt{\theta}|\boldsymbol{\tau}|}\left(1+c_{\mu}\right)+\frac{\sqrt{\theta|\boldsymbol{\tau}|} \cdot \sqrt{|\boldsymbol{\tau}|} \gamma}{\sqrt{\theta}|\boldsymbol{\tau}|}+\frac{\mu p \sqrt{\theta}|\boldsymbol{\tau}| \gamma}{\sqrt{\theta}|\boldsymbol{\tau}|} \\
& \leq \frac{2 c_{\mu} \theta_{\log }^{2}+c_{\mu} \theta_{\log }^{3}+c_{\mu} \theta_{\log }^{2}}{4 \sqrt{\theta}|\boldsymbol{\tau}|} \leq \frac{c_{\mu} \theta_{\log }^{2}}{\sqrt{\theta}|\boldsymbol{\tau}|} \tag{B.44}
\end{align*}
$$

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

$$
\begin{align*}
(\overline{\text { B.43) }} & \leq \frac{\gamma}{\theta|\boldsymbol{\tau}|} \cdot \frac{\mu \sqrt{p} \theta|\boldsymbol{\tau}|^{3 / 2}}{\gamma}\left(1+c_{\mu}\right)+\gamma+\mu p \gamma \\
& \leq \frac{\gamma}{\theta|\boldsymbol{\tau}|} \cdot \frac{4 \mu \sqrt{p} \theta|\boldsymbol{\tau}|^{3 / 2}}{c_{\mu} \theta_{\log }^{2}} \cdot \max \{\sqrt{|\boldsymbol{\tau}|}, \sqrt{\mu p}, \mu p \sqrt{\theta}|\boldsymbol{\tau}|\}+\frac{\gamma}{\theta|\boldsymbol{\tau}|} \cdot \theta|\boldsymbol{\tau}|+\frac{\gamma}{\theta|\boldsymbol{\tau}|} \cdot \mu p \theta|\boldsymbol{\tau}| \\
& \leq \frac{\gamma}{\theta|\boldsymbol{\tau}|} \cdot\left(\frac{4}{c_{\mu} \theta_{\log }^{2}} \cdot \max \left\{\mu|\boldsymbol{\tau}|^{2} \cdot \sqrt{p} \theta, \mu(p \theta)|\boldsymbol{\tau}| \cdot \sqrt{\mu|\boldsymbol{\tau}|}, \mu \sqrt{p \theta}|\boldsymbol{\tau}|^{3 / 2} \cdot \mu p \theta|\boldsymbol{\tau}|\right\}+\frac{c_{\mu} \theta_{\log }^{2}}{4}+\frac{c_{\mu} \theta_{\log }^{2}}{4}\right) \\
& \leq \frac{\gamma}{\theta|\boldsymbol{\tau}|}\left(\frac{c_{\mu} \theta_{\log }}{4}+\frac{c_{\mu} \theta_{\log }^{2}}{4}+\frac{c_{\mu} \theta_{\log }^{2}}{4}\right) \leq \frac{c_{\mu} \theta_{\log } \gamma}{\theta|\boldsymbol{\tau}|} \tag{B.45}
\end{align*}
$$

which completes the proof.

Corollary B.2.6 $\left(\mid\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}, \boldsymbol{x}_{\left.0, \boldsymbol{\tau}^{c}\right\rangle}\right\rangle\right.$ is small). Given $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ in $\mathbb{R}^{n}$ and $|\boldsymbol{\tau}|, c_{\mu}$ such that $\left(\boldsymbol{a}_{0}, \theta,|\boldsymbol{\tau}|\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. Write $\lambda=c_{\lambda} / \sqrt{|\boldsymbol{\tau}|}$ with some $c_{\lambda} \geq 1 / 5$, then if $c_{\mu} \leq \frac{c_{\lambda}}{25}$,

$$
\begin{equation*}
\mathbb{P}\left[\left|\sum_{i \in \boldsymbol{\tau}^{c}} \boldsymbol{\beta}_{i} \boldsymbol{x}_{0 i}\right|>\frac{\lambda}{10}\right] \leq 2 \theta, \quad \mathbb{P}\left[\left|\sum_{i} \boldsymbol{\beta}_{i} \boldsymbol{x}_{0 i}\right|>\frac{\lambda}{10}\right] \leq \theta|\boldsymbol{\tau}|+2 \theta \tag{B.46}
\end{equation*}
$$

Proof. We bound tail probability of the first result with Gaussian moments Theorem B.10.2 and Bernstein inequality Theorem B.10.4. Via Hölder's inequality, $\sum_{i \in \boldsymbol{\tau}^{c}} \mathbb{E}\left(\beta_{i} x_{i}\right)^{q}=\mathbb{E} x_{0}^{q}\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{q}^{q} \leq \theta(q-1)!!\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{\infty}^{q-2}$, thus

$$
\begin{equation*}
\mathbb{P}\left[\left|\sum_{i \in \boldsymbol{\tau}^{c}} \boldsymbol{\beta}_{i} \boldsymbol{x}_{0 i}\right|>\lambda / 10\right] \leq 2 \exp \left(\frac{-(\lambda / 10)^{2}}{2 \theta\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}+2(\lambda / 10)\left\|\boldsymbol{\beta}_{\tau^{c}}\right\|_{\infty}}\right) \tag{B.47}
\end{equation*}
$$

Write $\theta_{\log }=-\frac{1}{\log \theta}$, Theorem B.2.5 imples when $c_{\mu} \leq \frac{c \lambda}{25}$, we have $\theta\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \leq \frac{c_{\mu}^{2} \theta_{\log }^{2}}{|\boldsymbol{\tau}|^{2}} \leq \frac{\theta_{\log \lambda^{2}}}{625}$ and $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{\infty} \leq$
$\frac{c_{\mu} \theta_{\mathrm{log}}}{\sqrt{|\tau|}} \leq \frac{\theta_{\log \lambda} \lambda}{25}$, therefore,

$$
\begin{equation*}
\text { B.47) } \leq 2 \exp \left(\frac{-\lambda^{2} / 100}{2 \theta_{\log } \lambda^{2} / 625+2\left(\theta_{\log } \lambda / 25\right) \cdot(\lambda / 10)}\right) \leq 2 \exp (\log \theta) \leq 2 \theta \tag{B.48}
\end{equation*}
$$

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

$$
\begin{align*}
\mathbb{P}\left[\left|\sum_{i} \boldsymbol{\beta}_{i} \boldsymbol{x}_{0 i}\right|>\frac{\lambda}{10}\right] & \leq \mathbb{P}\left[\left|\boldsymbol{\beta}_{\boldsymbol{\tau}}^{*} \boldsymbol{x}_{\boldsymbol{\tau}}\right|+\left|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{x}_{\boldsymbol{\tau}^{c}}\right|>\lambda / 10\right] \\
& \leq \mathbb{P}\left[\boldsymbol{x}_{\boldsymbol{\tau}} \neq \mathbf{0}\right]+\mathbb{P}\left[\boldsymbol{x}_{\boldsymbol{\tau}}=\mathbf{0}\right] \cdot \mathbb{P}\left[\left|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}^{*} \boldsymbol{x}_{\boldsymbol{\tau}^{c}}\right|>\lambda / 10\right] \\
& \leq \theta|\boldsymbol{\tau}|+2 \theta \tag{B.49}
\end{align*}
$$

Corollary B.2.7 $\left(\left|\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash(0)}, \boldsymbol{x}_{0, \boldsymbol{\tau} \backslash(0)}\right\rangle\right|\right.$ 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 $\left(\boldsymbol{a}_{0}, \theta,|\boldsymbol{\tau}|\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$, then if $c_{\mu} \leq \frac{1}{10}$, for any $\boldsymbol{a}$ such that $\left|\boldsymbol{\beta}_{(1)}\right| \leq$ $\frac{\lambda}{4 \log \theta^{-1}}$, we have

$$
\begin{equation*}
\mathbb{P}\left[\left|\sum_{i \in \boldsymbol{\tau} \backslash(0)} \boldsymbol{\beta}_{i} \boldsymbol{x}_{0 i}\right|>\frac{2 \lambda}{5}\right] \leq 2 \theta \tag{B.50}
\end{equation*}
$$

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 $s^{\prime}=\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash\{0\}}, \boldsymbol{x}_{\boldsymbol{\tau} \backslash\{0\}}\right\rangle$. Firstly most of the entries of $\boldsymbol{x}_{\boldsymbol{\tau}}$ would be zero since via Bernstein inequality with $\theta|\boldsymbol{\tau}|<0.1$ :

$$
\begin{equation*}
\mathbb{P}\left[\sum_{i \in \boldsymbol{\tau}} \boldsymbol{\omega}_{i}>\log \theta^{-1}\right] \leq \mathbb{P}\left[\sum_{i \in \boldsymbol{\tau}} \boldsymbol{\omega}_{i}>\theta|\boldsymbol{\tau}|+0.9 \log \theta^{-1}\right] \leq \exp \left(\frac{-0.9^{2} \log ^{2} \theta^{-1}}{2\left(\theta|\boldsymbol{\tau}|+0.9 \log \theta^{-1} / 3\right)}\right) \leq \theta \tag{B.51}
\end{equation*}
$$

thus with probability at least $1-\theta$, we can write $s^{\prime}$ as a Gaussian r.v. with variation bounded as $\mathbb{E} s^{\prime 2} \leq$ $\mathbb{E}\left[\sum_{i=1}^{\log \theta^{-1}} \boldsymbol{\beta}_{i} \boldsymbol{g}_{i}\right]^{2}=\log \theta^{-1} \boldsymbol{\beta}_{(1)}^{2}$, then via Gaussian tail bound Theorem B.10.1.

$$
\begin{equation*}
\mathbb{P}\left[\left|s^{\prime}\right|>0.4 \lambda\right] \leq \mathbb{P}\left[|g|>\frac{0.4 \lambda}{\sqrt{\log \theta^{-1}}\left|\boldsymbol{\beta}_{(1)}\right|}\right]+\mathbb{P}\left[\sum_{i \in \boldsymbol{\tau}} \boldsymbol{\omega}_{i}>\log \theta^{-1}\right] \leq \frac{2}{\sqrt{2 \pi}} \exp \left(-1.2 \log \theta^{-1}\right)+\theta \leq 2 \theta \tag{B.52}
\end{equation*}
$$

## B. 3 Euclidean gradient as soft-thresholding in shift space

In this section, we will study the Euclidean gradient (3.41), by deriving bounds showing that the $\chi$ operator approximates a soft-thresholding function in shift space Theorem B.3.2 and Theorem B.3.4). Furthermore, we will show the operator $\boldsymbol{\chi}\left[\boldsymbol{\beta}_{i}\right]$ is monotone in $\left|\boldsymbol{\beta}_{i}\right|$ from Theorem B.3.3. A figure of visualized $\chi$ operator is shown in Figure B.1.

Expectation of $\chi$ operator. To understand the $\chi$ operator, we shall first consider a simple case-when $\boldsymbol{x}_{0}$ is highly sparse. By definition of $\boldsymbol{\beta}$ from (3.38) we can see that $\boldsymbol{\beta}$ has a short support of size at most $2 p-1$, when $\boldsymbol{x}_{0}$ has support entries separated by at least $2 p$, the entries of vector $\boldsymbol{\chi}[\boldsymbol{\beta}]_{i}$ become sum of independent random variables as:

$$
\boldsymbol{\chi}[\boldsymbol{\beta}]_{i}=\left\langle s_{-i}\left[\boldsymbol{x}_{0}\right], \mathcal{S}_{\lambda}\left[\boldsymbol{x}_{0} * \check{\boldsymbol{\beta}}\right]\right\rangle \underbrace{=}_{\boldsymbol{x}_{0} \text { sep. }}\left\langle s_{-i}\left[\boldsymbol{x}_{0}\right], \mathcal{S}_{\lambda}\left[\boldsymbol{\beta}_{i} s_{-i}\left[\boldsymbol{x}_{0}\right]\right]\right\rangle=\sum_{j \in \operatorname{supp}\left(\boldsymbol{x}_{0}\right)} \boldsymbol{g}_{j} \cdot \mathcal{S}_{\lambda}\left[\boldsymbol{g}_{j} \cdot \boldsymbol{\beta}_{i}\right]
$$

where $\left(\boldsymbol{g}_{j}\right)_{j \in[n]}$ are standard Gaussian r.v.s.
The following lemma describes the behavior of the summands in the above expression:
Lemma B.3.1 (Gaussian smoothed soft-thresholding). Let $g \sim \mathcal{N}(0,1)$. Then for every $b, s \in \mathbb{R}$ and $\lambda>0$,

$$
\begin{equation*}
\mathbb{E}_{g}\left[g \mathcal{S}_{\lambda}[b \cdot g+s]\right]=b\left(1-\operatorname{erf}_{b}(\lambda, s)\right) \tag{B.53}
\end{equation*}
$$

where

$$
\begin{equation*}
\operatorname{erf}_{b}(\lambda, s)=\frac{1}{2} \operatorname{erf}\left(\frac{\lambda+s}{\sqrt{2}|b|}\right)+\frac{1}{2} \operatorname{erf}\left(\frac{\lambda-s}{\sqrt{2}|b|}\right) \tag{B.54}
\end{equation*}
$$

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

$$
\begin{equation*}
\sigma \mathcal{S}_{\nu_{2}^{\prime} \lambda}[b] \leq \sigma \mathbb{E}_{g}\left[g \mathcal{S}_{\lambda}[b \cdot g]\right] \leq \sigma \mathcal{S}_{\nu_{1}^{\prime}(\varepsilon) \lambda}[b]+\varepsilon \tag{B.55}
\end{equation*}
$$

where $\nu_{1}^{\prime}(\varepsilon)=1 /(2 \sqrt{-\log \varepsilon})$ and $\nu_{2}^{\prime}=\sqrt{2 / \pi}$.
Proof. Wlog assume $b>0$. Write $f$ as the pdf of standard Gaussian distribution. With integral by parts:

$$
\int_{-\infty}^{t} t^{\prime} f\left(t^{\prime}\right) d t^{\prime}=-f(t), \quad \int_{-\infty}^{t} t^{\prime 2} f\left(t^{\prime}\right) d t^{\prime}=\frac{1}{2} \operatorname{erf}\left(\frac{t}{\sqrt{2}}\right)-t f(t)
$$

Integrating, we obtain

$$
\mathbb{E}\left[g \mathcal{S}_{\lambda}[b \cdot g+s]\right]=\int_{t \geq \frac{\lambda-s}{b}}\left(b t^{2}-(\lambda-s) t\right) f(t) d t+\int_{t \leq-\frac{\lambda+s}{b}}\left(b t^{2}+(\lambda+s) t\right) f(t) d t
$$

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

$$
\int_{t \geq \frac{L}{b}}\left(b t^{2}-L t\right) f(t) d t=b\left[\frac{1}{2}-\frac{1}{2} \operatorname{erf}\left(\frac{L}{\sqrt{2} b}\right)+\frac{L}{b} f\left(\frac{L}{b}\right)\right]-L f\left(\frac{L}{b}\right)=\frac{b}{2}-\frac{b}{2} \operatorname{erf}\left(\frac{L}{\sqrt{2} b}\right)
$$

and similarly for the second summand, which gives

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

For $b<0$, alternatively we have

$$
\mathbb{E}\left[g S_{\lambda}[-|b| \cdot g+s]\right]=-\mathbb{E}\left[g S_{\lambda}[|b| \cdot g-s]=-|b|\left(1-\operatorname{erf}_{b}(\lambda,-s)\right)=b\left(1-\operatorname{erf}_{b}(\lambda, s)\right)\right.
$$

To show B.55, via definition of error function, for $x>0$, we know:

$$
\begin{equation*}
\min \left\{1-\varepsilon, \frac{1-\varepsilon}{\sqrt{\log (1 / \varepsilon)}} x\right\} \leq \operatorname{erf}(x)=\frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^{2}} d t \leq \frac{2 x}{\sqrt{\pi}} \tag{B.56}
\end{equation*}
$$

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

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

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

$$
\operatorname{erf}(x) \geq \frac{\operatorname{erf}(T)-\operatorname{erf}(0)}{T-0} x+\operatorname{erf}(0) \geq \frac{1-\varepsilon}{\sqrt{\log (1 / \varepsilon)}} x
$$

Lastly plug B.56 into B.53 and apply condition $|b| \leq 1$ and $\varepsilon<1 / 4$ we have

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

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 $x_{0}$ is sufficiently sparse (when Theorem B.2.1 holds). Recall that $\chi: \mathbb{R}^{n} \rightarrow \mathbb{R}^{n}$ is defined as

$$
\begin{equation*}
\boldsymbol{\chi}[\boldsymbol{\beta}]=\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{\beta}\right] \tag{B.57}
\end{equation*}
$$

The following lemma bounds its expectation:

Lemma B.3.2 (Expectation of $\boldsymbol{\chi}(\boldsymbol{\beta})$ ). Let $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ and $\lambda>0$, then for every $\boldsymbol{a} \in \mathbb{S}^{p-1}$ and every $i \in[n]$,


Figure B.1: 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 $2 p$ separated. Left: the yellow line is the function $\boldsymbol{\beta}_{i} \rightarrow \boldsymbol{\beta}_{i}\left(1-\operatorname{erf}_{\boldsymbol{\beta}_{\boldsymbol{i}}}(\lambda, 0)\right)$ derived from (B.53), and the blue/red lines are its upper/lower bound B.55) utilized in the analysis respectively. Right: functions of $\boldsymbol{\beta}_{i} \rightarrow \boldsymbol{\beta}_{i}\left(1-\operatorname{erf}_{\boldsymbol{\beta}_{i}}(\lambda, 0)\right)$ with different $\lambda$, the section of function of $\boldsymbol{\beta}_{i}>\nu_{2} \lambda$ are close to linear.
define the operator $\chi$ as in (B.57), then

$$
\begin{equation*}
n^{-1} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}=\theta \boldsymbol{\beta}_{i}\left(1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)\right) \tag{B.58}
\end{equation*}
$$

where $\boldsymbol{s}_{i}=\sum_{\ell \neq i} \boldsymbol{\beta}_{\ell} \boldsymbol{x}_{0 \ell}$. Suppose $\left(\boldsymbol{a}_{0}, \theta,|\boldsymbol{\tau}|\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$ and $\lambda=c_{\lambda} / \sqrt{|\boldsymbol{\tau}|}$ for some $c_{\lambda}>1 / 5$ and $\sigma_{i}=\operatorname{sign}\left(\boldsymbol{\beta}_{i}\right)$, then there exists some numerical constant $\bar{c}$ such that if $c_{\mu} \leq \bar{c}$ then for every $\boldsymbol{a} \in \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ and every $i \in[n]$, B.58) has upper bound

$$
\sigma_{i} n^{-1} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \leq \sigma_{i} n^{-1} \overline{\mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}}:= \begin{cases}4 \theta^{2}|\boldsymbol{\tau}|\left|\boldsymbol{\beta}_{i}\right| & \left|\boldsymbol{\beta}_{i}\right|<\nu_{1} \lambda  \tag{B.59}\\ \theta\left(\left|\boldsymbol{\beta}_{i}\right|-\nu_{1} \lambda / 2\right) & \left|\boldsymbol{\beta}_{i}\right| \geq \nu_{1} \lambda\end{cases}
$$

and lower bound

$$
\begin{equation*}
\sigma_{i} n^{-1} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \geq \sigma_{i} n^{-1} \underline{\mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}}=: \theta \mathcal{S}_{\nu_{2} \lambda}\left[\left|\boldsymbol{\beta}_{i}\right|\right] \tag{B.60}
\end{equation*}
$$

where $\nu_{1}=1 /\left(2 \sqrt{\log \theta^{-1}}\right), \nu_{2}=\sqrt{2 / \pi}$.
This lemma shows the expectation of $\boldsymbol{\chi}[\boldsymbol{\beta}]_{i}$ acts like a shrinkage operation on $\left|\boldsymbol{\beta}_{i}\right|$ : for large $\left|\boldsymbol{\beta}_{i}\right|$, it acts like a soft thresholding operation, and for small $\left|\boldsymbol{\beta}_{i}\right|$, it reduces $\left|\boldsymbol{\beta}_{i}\right|$ by multiplying a very small number $4 \theta|\boldsymbol{\tau}| \ll 1$. We rigorously prove this segmentation of $\chi$ operator as follows:

Proof. First, since $s_{i}\left[\boldsymbol{x}_{0}\right] \equiv{ }_{d} s_{j}\left[\boldsymbol{x}_{0}\right]$,

$$
\boldsymbol{\chi}[\boldsymbol{\beta}]_{i}=\boldsymbol{e}_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{\beta}\right]=\left\langle s_{-i}\left[\boldsymbol{x}_{0}\right], \mathcal{S}_{\lambda}\left[\boldsymbol{x}_{0} * \check{\boldsymbol{\beta}}\right]\right\rangle \equiv_{d}\left\langle s_{-j}\left[\boldsymbol{x}_{0}\right], \mathcal{S}_{\lambda}\left[s_{i-j}\left[\boldsymbol{x}_{0}\right] * \check{\boldsymbol{\beta}}\right]\right\rangle=\boldsymbol{\chi}\left[s_{j-i}[\boldsymbol{\beta}]\right]_{j}
$$

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:

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

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

$$
\begin{equation*}
Z_{j}(\boldsymbol{\beta})=\boldsymbol{x}_{j} \mathcal{S}_{\lambda}\left[\boldsymbol{\beta}_{0} \boldsymbol{x}_{j}+\sum_{\ell \in[ \pm p] \backslash 0} \boldsymbol{\beta}_{\ell} \boldsymbol{x}_{j+\ell}\right], \tag{B.61}
\end{equation*}
$$

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 $s_{0}=\sum_{\ell \neq 0} \boldsymbol{\beta}_{\ell} \boldsymbol{x}_{\ell}$, which is independent of $\boldsymbol{x}_{0}$. From Theorem B.3.1. we can conclude

$$
\begin{equation*}
n^{-1} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{0}=\mathbb{E}_{\boldsymbol{x}_{0}, s_{0}} \boldsymbol{x}_{0} \mathcal{S}_{\lambda}\left[\boldsymbol{\beta}_{0} \boldsymbol{x}_{0}+s_{0}\right]=\theta \boldsymbol{\beta}_{0}\left(1-\mathbb{E}_{\boldsymbol{s}_{0}} \operatorname{erf}_{\boldsymbol{\beta}_{0}}\left(\lambda, s_{0}\right)\right) \tag{B.62}
\end{equation*}
$$

so that B.58 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 $\mathbb{E} \boldsymbol{x}_{0} \mathcal{S}_{\lambda}\left[0 \cdot \boldsymbol{x}_{0}+s_{0}\right]=0$, we have

$$
\begin{align*}
\mathbb{E} Z(\boldsymbol{\beta}) & \leq \boldsymbol{\beta}_{0} \sup _{\beta \in\left[0, \boldsymbol{\beta}_{0}\right]} \frac{d}{d \beta} \mathbb{E}_{\boldsymbol{x}_{0}, \boldsymbol{s}_{0}} \boldsymbol{x}_{0} \mathcal{S}_{\lambda}\left[\beta \boldsymbol{x}_{0}+\boldsymbol{s}_{0}\right]=\theta \boldsymbol{\beta}_{0} \sup _{\beta \in\left[0, \boldsymbol{\beta}_{0}\right]} \frac{d}{d \beta} \int_{\left|\beta g+s_{0}\right|>\lambda} g\left(\beta g+s_{0}-\operatorname{sign}\left(\beta g+\boldsymbol{s}_{0}\right) \cdot \lambda\right) d \mu(g) d \mu\left(\boldsymbol{s}_{0}\right) \\
& \left.=\theta \boldsymbol{\beta}_{0} \sup _{\beta \in\left[0, \boldsymbol{\beta}_{0}\right]} \mathbb{E}_{g, \boldsymbol{s}_{0}}\left[g^{2} \mathbf{1}_{\left\{\left|\beta g+\boldsymbol{s}_{0}\right|>\lambda\right\}}\right] \leq \theta \boldsymbol{\beta}_{0} \sup _{\beta \in\left[0, \boldsymbol{\beta}_{0}\right]} \mathbb{E}_{g, \boldsymbol{s}_{0}}\left[g^{2}\left(\mathbf{1}_{\left\{|\beta g|>\frac{9 \lambda}{10}\right\}}+\mathbf{1}_{\left\{\left|s_{0}\right|>\lambda\right.}\right\}\right)\right] \\
& \leq \theta \boldsymbol{\beta}_{0}\left(\left(\mathbb{E} g^{6}\right)^{1 / 3} \mathbb{P}\left[\left|\boldsymbol{\beta}_{0} g\right|>(9 \lambda / 10)\right]^{2 / 3}+\mathbb{P}\left[\left|\boldsymbol{s}_{0}\right|>\lambda / 10\right]\right) \tag{B.63}
\end{align*}
$$

We bound the tail probability of $s_{0}$ using Theorem B.2.6 where

$$
\begin{equation*}
\mathbb{P}\left[\left|s_{0}\right|>\lambda / 10\right] \leq \mathbb{P}\left[\left|\sum_{i} \boldsymbol{\beta}_{i} \boldsymbol{x}_{i}\right|>\lambda / 10\right] \leq \theta|\boldsymbol{\tau}|+2 \theta \leq 3 \theta|\boldsymbol{\tau}| . \tag{B.64}
\end{equation*}
$$

On the other hand, the first term in (B.63) can be derived by pdf of Gaussian r.v. Theorem B.10.1 as:

$$
\begin{equation*}
\left(\mathbb{E} g^{6}\right)^{1 / 3} \mathbb{P}\left[\left|\boldsymbol{\beta}_{0} g\right|>(9 \lambda / 10)\right]^{2 / 3} \leq \sqrt[3]{15}\left(\frac{10 \boldsymbol{\beta}_{0}}{9 \lambda \sqrt{2 \pi}}\right)^{2 / 3} \exp \left(-\frac{\lambda^{2}}{4 \boldsymbol{\beta}_{0}^{2}}\right) \leq \frac{3}{2}\left(\frac{\boldsymbol{\beta}_{0}}{\lambda}\right)^{2 / 3} \exp \left(-\frac{\lambda^{2}}{4 \boldsymbol{\beta}_{0}^{2}}\right) \tag{B.65}
\end{equation*}
$$

Combine $\overline{\text { B.48, }}$ (65), when $\boldsymbol{\beta}_{0}<\nu_{1} \lambda$, we know $e^{-\frac{\lambda^{2}}{4 \boldsymbol{\beta}_{0}^{2}}} \leq e^{\log \theta} \leq \theta|\boldsymbol{\tau}|$. The first type of upper bound $\mathbb{E} Z$ is derived as

$$
\begin{equation*}
\forall \boldsymbol{\beta}_{0} \in\left[0, \nu_{1} \lambda\right], \quad \mathbb{E} Z(\boldsymbol{\beta}) \leq \theta \boldsymbol{\beta}_{0}\left(\frac{3}{2} \nu_{1}^{2 / 3} \exp \left(-\frac{\lambda^{2}}{4 \boldsymbol{\beta}_{0}^{2}}\right)+3 \theta|\boldsymbol{\tau}|\right) \leq 4 \theta^{2}|\boldsymbol{\tau}| \boldsymbol{\beta}_{0} \tag{B.66}
\end{equation*}
$$

(2). The second type of upper bound can be derived directly from Theorem B.3.1

$$
\begin{align*}
\mathbb{E} Z(\boldsymbol{\beta}) & \leq \mathbb{E}_{\boldsymbol{x}_{0}} \mathbb{E}_{\boldsymbol{s}_{0}} \boldsymbol{x}_{0} \mathcal{S}_{\lambda}\left[\boldsymbol{\beta}_{0} \boldsymbol{x}_{0}+\boldsymbol{s}_{0}\right] \leq \mathbb{E}_{\boldsymbol{x}_{0}} \boldsymbol{x}_{0} \mathcal{S}_{\lambda}\left[\boldsymbol{\beta}_{0} \boldsymbol{x}_{0}\right]+\mathbb{E}_{\boldsymbol{x}_{0}}\left|\boldsymbol{x}_{0}\right| \mathbb{E}_{\boldsymbol{s}_{0}}\left|\boldsymbol{s}_{0}\right| \\
& \leq \theta \cdot\left(\mathcal{S}_{\nu_{1}^{\prime} \lambda}\left[\boldsymbol{\beta}_{0}\right]+\varepsilon+\sqrt{2 / \pi} \cdot \mathbb{E}\left|\boldsymbol{s}_{0}\right|\right) \tag{B.67}
\end{align*}
$$

where $\mathbb{E}|s|$ can be bounded with $\|\boldsymbol{\beta}\|_{2}$ and $\theta|\boldsymbol{\tau}|<c_{\mu} \theta_{\log }$ from Theorem B.2.5. When $c_{\mu}<\frac{1}{10}$, observe that

$$
\begin{equation*}
\mathbb{E}|s| \leq \sqrt{\sum_{\ell} \mathbb{E} \boldsymbol{x}_{\ell}^{2} \boldsymbol{\beta}_{\ell}^{2}} \leq \sqrt{\theta}\left(\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}+\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}\right) \leq \sqrt{\theta}\left(1+c_{\mu}\right)+\frac{c_{\mu} \theta_{\log }}{|\boldsymbol{\tau}|} \leq \frac{2 c_{\mu} \theta_{\log }}{\sqrt{|\boldsymbol{\tau}|}} \tag{B.68}
\end{equation*}
$$

Now choose $\varepsilon=\theta \leq \frac{c_{\mu} \theta_{\mathrm{log}}}{|\boldsymbol{\tau}|}$, so that $\nu_{1}^{\prime}=\nu_{1}=\frac{\sqrt{\theta_{\mathrm{log}}}}{2}$ in B.67. Since $c_{\mu}<\frac{c_{\lambda}}{25}$ we gain

$$
\begin{align*}
\mathbb{E} Z(\boldsymbol{\beta}) & \leq \theta\left(\mathcal{S}_{\nu_{1} \lambda}\left[\boldsymbol{\beta}_{0}\right]+\frac{c_{\mu} \theta_{\log }}{|\boldsymbol{\tau}|}+\sqrt{\frac{2}{\pi}} \cdot \frac{2 c_{\mu} \theta_{\log }}{\sqrt{|\boldsymbol{\tau}|}}\right) \leq \theta\left(\mathcal{S}_{\nu_{1} \lambda}\left[\boldsymbol{\beta}_{0}\right]+\frac{3 c_{\mu} \theta_{\log }}{\sqrt{|\boldsymbol{\tau}|}}\right) \\
& \leq \theta\left(\mathcal{S}_{\nu_{1} \lambda}\left[\boldsymbol{\beta}_{0}\right]+\frac{\sqrt{\theta_{\log }}}{5} \lambda\right) \leq \theta\left(\mathcal{S}_{\nu_{1} \lambda}\left[\boldsymbol{\beta}_{0}\right]+\frac{1}{2} \nu_{1} \lambda\right) \tag{B.69}
\end{align*}
$$

(3). Combine both B.66 and B.69, we can thus conclude that

$$
\mathbb{E} Z(\boldsymbol{\beta}):=\overline{\mathbb{E} Z(\boldsymbol{\beta})} \leq \begin{cases}4 \theta^{2}|\boldsymbol{\tau}| \boldsymbol{\beta}_{0} & \boldsymbol{\beta}_{0} \leq \nu_{1} \lambda  \tag{B.70}\\ \theta\left(\boldsymbol{\beta}_{0}-\frac{\nu_{1}}{2} \lambda\right) & \boldsymbol{\beta}_{0}>\nu_{1} \lambda\end{cases}
$$

2. (Lower bound of $\mathbb{E} Z$ ) 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 $s_{0}$, we have

$$
\begin{align*}
\mathbb{E} Z(\boldsymbol{\beta}) & =\mathbb{E}_{\boldsymbol{s}_{0}} \mathbb{E}_{\boldsymbol{x}_{0}} \boldsymbol{x}_{0} \mathcal{S}_{\lambda}\left[\boldsymbol{\beta}_{0} \boldsymbol{x}_{0}+\boldsymbol{s}_{0}\right]=\theta \cdot \mathbb{E}_{\boldsymbol{s}_{0}}\left[\boldsymbol{\beta}_{0}-\frac{\boldsymbol{\beta}_{0}}{2} \cdot \operatorname{erf}\left(\frac{\lambda-\boldsymbol{s}_{0}}{\sqrt{2}\left|\boldsymbol{\beta}_{0}\right|}\right)-\frac{\boldsymbol{\beta}_{0}}{2} \cdot \operatorname{erf}\left(\frac{\lambda+\boldsymbol{s}_{0}}{\sqrt{2}\left|\boldsymbol{\beta}_{0}\right|}\right)\right] \\
& \geq \theta\left(\boldsymbol{\beta}_{0}-\boldsymbol{\beta}_{0} \cdot \operatorname{erf}\left(\frac{\lambda}{\sqrt{2}\left|\boldsymbol{\beta}_{0}\right|}\right)\right) \geq \theta \cdot \mathcal{S}_{\nu_{2}^{\prime} \lambda}\left[\boldsymbol{\beta}_{0}\right]=: \underline{\mathbb{E} Z(\boldsymbol{\beta})} . \tag{B.71}
\end{align*}
$$

The proof of $\boldsymbol{\beta}_{0}<0$ is in the same vein. For cases of $i \neq 0$, since $\boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \equiv_{d} \boldsymbol{\chi}\left[s_{-i}[\boldsymbol{\beta}]\right]_{0}$, 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. $\left|\boldsymbol{\beta}_{i}\right|$. The monotonicity is clear in Figure B.1; it is demonstrated rigorously with the following lemma:

Lemma B.3.3 (Monotonicity of $\mathbb{E} \boldsymbol{\chi}(\boldsymbol{\beta}))$. Suppose $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ in $\mathbb{R}^{n}$, and $|\boldsymbol{\tau}|, c_{\mu}$ such that $\left(\boldsymbol{a}_{0}, \theta,|\boldsymbol{\tau}|\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $\left|\mathbb{E}[\boldsymbol{\chi}[\boldsymbol{\beta}]]_{i}\right|$ is monotone increasing in $\left|\boldsymbol{\beta}_{i}\right|$. In other words, if $\left|\boldsymbol{\beta}_{i}\right|>\left|\boldsymbol{\beta}_{j}\right|$ then

$$
\begin{equation*}
\sigma_{i} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \geq \sigma_{j} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{j} \tag{B.72}
\end{equation*}
$$

where $\sigma_{i}=\operatorname{sign}\left(\boldsymbol{\beta}_{i}\right)$.

The proof first operate simple calculus and then followed by studying cases of $\left|\boldsymbol{\beta}_{i}\right|-\left|\boldsymbol{\beta}_{j}\right|$ 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 Theorem B.3.2 we can write $(n \theta)^{-1} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}=\boldsymbol{\beta}_{i}\left(1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)\right)$. Consider $t \in[0,1]$ and define $\ell(t)=t \boldsymbol{\beta}_{i}-t \boldsymbol{\beta}_{j}$. Write the random variable $\boldsymbol{s}_{i j}=\sum_{\ell \neq i, j} \boldsymbol{\beta}_{\ell} \boldsymbol{x}_{\ell}$. Define $h$ as a function of $t$ such that

$$
\begin{align*}
h(t) & =\mathbb{E}_{x, \boldsymbol{s}_{i j}}\left[\left((1-t) \boldsymbol{\beta}_{i}+t \boldsymbol{\beta}_{j}\right)\left(1-\operatorname{erf}_{(1-t) \boldsymbol{\beta}_{i}+t \boldsymbol{\beta}_{j}}\left(\lambda,\left((1-t) \boldsymbol{\beta}_{j}+t \boldsymbol{\beta}_{i}\right) x+\boldsymbol{s}_{i j}\right)\right)\right] \\
& =\mathbb{E}_{x, \boldsymbol{s}_{i j}}\left[\left(\boldsymbol{\beta}_{i}-\ell(t)\right)\left(1-\operatorname{erf}_{\boldsymbol{\beta}_{i}-\ell(t)}\left(\lambda, x \cdot\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+\boldsymbol{s}_{i j}\right)\right)\right] \tag{B.73}
\end{align*}
$$

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 $h^{\prime}(t)<0$ for all $t \in[0,1]$. Write $f$ as pdf of standard Gaussian r.v. where

$$
\operatorname{erf}_{\beta}\left(\lambda, s_{i j}\right)=\int_{0}^{\frac{\lambda+s_{i j}}{\beta}} f(z) d z+\int_{0}^{\frac{\lambda-s_{i j}}{\beta}} f(z) d z
$$

and use chain rule:

$$
\begin{aligned}
h^{\prime}(t)= & \mathbb{E}_{x, \boldsymbol{s}_{i j}}\left[\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right)\left(1-\operatorname{erf}_{\boldsymbol{\beta}_{i}-\ell(t)}\left(\lambda, x \cdot\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+\boldsymbol{s}_{i j}\right)\right)\right. \\
& -\left(\boldsymbol{\beta}_{i}-\ell(t)\right) \cdot \frac{d}{d t}\left(\frac{\lambda+x \cdot\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+\boldsymbol{s}_{i j}}{\boldsymbol{\beta}_{i}-\ell(t)}\right) \cdot f\left(\frac{\lambda+x \cdot\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+\boldsymbol{s}_{i j}}{\boldsymbol{\beta}_{i}-\ell(t)}\right) \\
& \left.-\left(\boldsymbol{\beta}_{i}-\ell(t)\right) \cdot \frac{d}{d t}\left(\frac{\lambda-x \cdot\left(\boldsymbol{\beta}_{j}+\ell(t)\right)-\boldsymbol{s}_{i j}}{\boldsymbol{\beta}_{i}-\ell(t)}\right) \cdot f\left(\frac{\lambda-x \cdot\left(\boldsymbol{\beta}_{j}+\ell(t)\right)-\boldsymbol{s}_{i j}}{\boldsymbol{\beta}_{i}-\ell(t)}\right)\right] \\
= & \left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right) \mathbb{E}_{x, \boldsymbol{s}_{i j}}\left[1-\operatorname{erf}_{\boldsymbol{\beta}_{i}-\ell(t)}\left(\lambda, x \cdot\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+\boldsymbol{s}_{i j}\right)\right.
\end{aligned}
$$

$$
\begin{gather*}
+\left(\frac{\lambda+x\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+\boldsymbol{s}_{i j}}{\boldsymbol{\beta}_{i}-\ell(t)}+x\right) \cdot f \underbrace{\left(\frac{\lambda+x\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+\boldsymbol{s}_{i j}}{\boldsymbol{\beta}_{i}-\ell(t)}\right)}_{z_{\lambda_{+}}} \\
+\left(\frac{\lambda-x\left(\boldsymbol{\beta}_{j}+\ell(t)\right)-\boldsymbol{s}_{i j}}{\boldsymbol{\beta}_{i}-\ell(t)}-x\right) \cdot f \underbrace{\left(\frac{\lambda-x\left(\boldsymbol{\beta}_{j}+\ell(t)\right)-\boldsymbol{s}_{i j}}{\boldsymbol{\beta}_{i}-\ell(t)}\right)}_{z_{\lambda_{-}}}] \\
=\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right) \mathbb{E}_{x, \boldsymbol{s}_{i j}}\left[1-\int_{0}^{z_{\lambda_{+}}} f(z) d z-\int_{0}^{z_{\lambda_{-}}} f(z) d z+\left(z_{\lambda_{+}}+x\right) f\left(z_{\lambda_{+}}\right)+\left(z_{\lambda_{-}}-x\right) f\left(z_{\lambda_{-}}\right)\right] . \tag{B.74}
\end{gather*}
$$

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

$$
\left\{\begin{array}{l}
\mu_{+-}:=\mathbb{E}_{x, s_{i j} \mid z_{\lambda_{+}} \leq 0}\left[\int_{0}^{z_{\lambda_{+}}} f(z) d z-z_{\lambda_{+}} f\left(z_{\lambda_{+}}\right)\right]=\mathbb{E}_{x, s \mid z_{\lambda_{+} \leq 0}}\left[-\int_{0}^{-z_{\lambda_{+}}} f(z) d z-z_{\lambda_{+}} f\left(z_{\lambda_{+}}\right)\right] \leq 0 \\
\mu_{++}:=\mathbb{E}_{x, s_{i j} \mid z_{\lambda_{+}>0}}\left[\int_{0}^{z_{\lambda_{+}}} f(z) d z-z_{\lambda_{+}} f\left(z_{\lambda_{+}}\right)\right] \leq \min \left\{\frac{1}{2}, \frac{1}{\sqrt{2 \pi}} \mathbb{E}_{x, s_{i j} \mid z_{\lambda_{+}}>0} z_{\lambda_{+}}\right\}
\end{array}\right.
$$

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

$$
\begin{cases}-\int_{0}^{v} f(z) d z+v \cdot f(v)=0 & v=0 \\ \frac{d}{d v}\left\{-\int_{0}^{v} f(z) d z+v \cdot f(v)\right\}=-f(v)+f(v)+v \cdot f^{\prime}(v)<0 & v>0\end{cases}
$$

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

$$
\left\{\begin{array}{l}
\mu_{--}:=\mathbb{E}_{x, s_{i j} \mid z_{\lambda_{-} \leq 0}}\left[\int_{0}^{z_{\lambda_{-}}} f(z) d z-z_{\lambda_{-}} f\left(z_{\lambda_{-}}\right)\right] \leq 0 \\
\mu_{-+}:=\mathbb{E}_{x, s_{i j} \mid z_{\lambda_{-}>0}}\left[\int_{0}^{z_{\lambda_{-}}} f(z) d z-z_{\lambda_{-}} f\left(z_{\lambda_{-}}\right)\right] \leq \min \left\{\frac{1}{2}, \frac{1}{\sqrt{2 \pi}} \mathbb{E}_{x, s_{i j} \mid z_{\lambda_{-}}>0} z_{\lambda_{-}}\right\}
\end{array}\right.
$$

then combine every term to $\mathrm{B.74}$ using tower property and from assumption $\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}<0$ we obtain

$$
\begin{align*}
\text { B.74 } & \leq\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right)\left(1-\mathbb{P}\left[z_{\lambda_{+}}>0\right] \cdot \mu_{++}-\mathbb{P}\left[z_{\lambda_{-}}>0\right] \cdot \mu_{-+}+\mathbb{E}_{x, \boldsymbol{s}_{i j}}\left[x\left(f\left(z_{\lambda_{+}}\right)-f\left(z_{\lambda_{-}}\right)\right)\right]\right) \\
& \leq\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right)\left(1-\min \left\{\frac{\mathbb{P}\left[z_{\lambda_{+}}>0\right]}{2}, \frac{\mathbb{E}\left|z_{\lambda_{+}}\right|}{\sqrt{2 \pi}}\right\}-\min \left\{\frac{\mathbb{P}\left[z_{\lambda_{-}}>0\right]}{2}, \frac{\mathbb{E}\left|z_{\lambda_{-}}\right|}{\sqrt{2 \pi}}\right\}-\frac{\theta}{\sqrt{2 \pi}} \cdot \mathbb{E}|g|\right) \tag{B.75}
\end{align*}
$$

where $g$ is standard Gaussian r.v..
2. (Cases of varying $\left.\boldsymbol{\beta}_{i}, \boldsymbol{\beta}_{j}\right)$ Let $c_{\lambda}<\frac{1}{4}$. Suppose $\boldsymbol{\beta}_{i}-\ell(t) \leq \frac{1}{4 \sqrt{|\boldsymbol{\tau}|}}$. Recall that $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2} \geq 1-3 c_{\mu}$. We are going to show there is at least one of the entry $\boldsymbol{\beta}_{*} \in\left\{\boldsymbol{\beta}_{r}\right\}_{r \in \boldsymbol{\tau} \neq i, j} \uplus\left\{\boldsymbol{\beta}_{j}+\ell(t)\right\}$ is greater than $\frac{0.85}{\sqrt{|\boldsymbol{\tau}|}}$. First, if both $i, j \notin \boldsymbol{\tau}$, the lower bound is immediate since $\boldsymbol{\beta}_{*}^{2}=\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{\infty}^{2}>\frac{1-3 c_{\mu}}{|\boldsymbol{\tau}|}$. 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 $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash\{i, j\}}\right\|_{\infty}^{2}<\frac{1-3 c_{\mu}}{|\boldsymbol{\tau}|}$, then we know via norm inequalities,

$$
\begin{equation*}
\left(\boldsymbol{\beta}_{i}+\boldsymbol{\beta}_{j}\right)^{2}>\boldsymbol{\beta}_{i}^{2}+\boldsymbol{\beta}_{j}^{2}>\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}-(|\boldsymbol{\tau}|-1)\left\|\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash\{i, j\}}\right\|_{\infty}^{2} \geq \frac{1-3 c_{\mu}}{|\boldsymbol{\tau}|} \tag{B.76}
\end{equation*}
$$

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

$$
\begin{equation*}
\boldsymbol{\beta}_{*}=\boldsymbol{\beta}_{j}+\ell(t)=\left(\boldsymbol{\beta}_{i}+\boldsymbol{\beta}_{j}\right)-\left(\boldsymbol{\beta}_{i}-\ell(t)\right) \geq \frac{\sqrt{1-3 c_{\mu}}}{\sqrt{|\boldsymbol{\tau}|}}-\frac{1}{4 \sqrt{|\boldsymbol{\tau}|}} \geq \frac{0.72}{\sqrt{|\boldsymbol{\tau}|}} \tag{B.77}
\end{equation*}
$$

In this case, adopt result from Theorem B.2.6 such that $\mathbb{P}\left[\left|\sum \boldsymbol{\beta}_{\ell} x_{\ell}\right|>\lambda / 10\right] \leq 3 \theta|\boldsymbol{\tau}| \leq .01$, we have

$$
\begin{align*}
\mathbb{P}\left[z_{\lambda_{-}}>0\right]=\mathbb{P}\left[z_{\lambda_{+}}>0\right] & =1-\mathbb{P}\left[x\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+s_{i j}<-\lambda\right] \\
& \leq 1-\mathbb{P}\left[\boldsymbol{x}_{*} \boldsymbol{\beta}_{*}<-11 \lambda / 10\right] \cdot \mathbb{P}\left[x\left(\boldsymbol{\beta}_{j}+\ell(t)\right)+\boldsymbol{s}_{i j}-\boldsymbol{x}_{*} \boldsymbol{\beta}_{*}<\lambda / 10\right] \\
& \leq 1-\theta \cdot \mathbb{P}\left[\boldsymbol{g}_{*} \cdot \frac{0.72}{\sqrt{|\boldsymbol{\tau}|}}<\frac{-11 c_{\lambda}}{10 \sqrt{|\boldsymbol{\tau}|}}\right] \cdot\left(1-\mathbb{P}\left[\sum \boldsymbol{\beta}_{\ell} \boldsymbol{x}_{\ell}>\frac{\lambda}{10}\right]\right) \\
& \leq 1-\theta \cdot \mathbb{P}\left[0.72 \cdot \boldsymbol{g}_{*} \leq-1.1 \cdot 0.25\right] \cdot\left(1-3 c_{\mu}\right) \\
& \leq 1-0.35 \theta . \tag{B.78}
\end{align*}
$$

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

$$
\begin{align*}
\mathbb{E}_{x, \boldsymbol{s}_{i j}}\left|z_{\lambda_{-}}\right|=\mathbb{E}_{x, \boldsymbol{s}_{i j}}\left|z_{\lambda_{+}}\right| & \leq \mathbb{E}_{x, \boldsymbol{s}_{i j}} \frac{\lambda+\left|x\left(\boldsymbol{\beta}_{j}+\ell(t)\right)-\boldsymbol{s}_{i j}\right|}{\boldsymbol{\beta}_{i}-\ell(t)} \leq 1+4 \sqrt{|\boldsymbol{\tau}|} \cdot\left(\mathbb{E}_{x, \boldsymbol{s}_{i j}}\left|x\left(\boldsymbol{\beta}_{j}+\ell(t)\right)-\boldsymbol{s}_{i j}\right|^{2}\right)^{1 / 2} \\
& \leq 1+4 \sqrt{|\boldsymbol{\tau}| \theta}\|\boldsymbol{\beta}\|_{2} \leq 1+4 \sqrt{|\boldsymbol{\tau}| \theta}\left(1+c_{\mu}+\frac{c_{\mu}}{\sqrt{\theta}|\boldsymbol{\tau}|}\right) \leq 1.2 \tag{B.79}
\end{align*}
$$

Combine (B.75, (B.78 we have

$$
\begin{equation*}
h^{\prime}(t) \leq\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right)\left(1-2 \cdot \frac{(1-0.35 \theta)}{2}-\frac{\theta}{\sqrt{2 \pi}} \cdot \sqrt{\frac{2}{\pi}}\right) \leq 0.03 \theta\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right)<0 \tag{B.80}
\end{equation*}
$$

and combine B.75, $\overline{\text { B.79 }}$ and $\theta<c_{\mu}$ we have

$$
\begin{equation*}
h^{\prime}(t) \leq\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right)\left(1-2 \cdot \frac{1.2}{\sqrt{2 \pi}}-\frac{\theta}{\sqrt{2 \pi}} \cdot \sqrt{\frac{2}{\pi}}\right) \leq 0.03\left(\boldsymbol{\beta}_{j}-\boldsymbol{\beta}_{i}\right)<0 \tag{B.81}
\end{equation*}
$$

which proves the monotonicity.

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

Corollary B.3.4 (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 $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $n \geq C p^{5} \theta^{-2} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-3 / n$,
for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ and every $i \in[n]$, we have:

$$
\begin{equation*}
\left|n^{-1} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}-n^{-1} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}\right| \leq c \theta / p^{3 / 2}, \tag{B.82}
\end{equation*}
$$

Proof. See Appendix B.9.1

## B. 4 Euclidean Hessian as logic function in shift space

We can express the (pseudo) curvature (3.45) in direction $\boldsymbol{v} \in \mathbb{S}^{p-1}$ in terms of the correlation $\gamma=C_{\boldsymbol{a}_{0}}^{*} \boldsymbol{v}$ between $\boldsymbol{v}$ and $\boldsymbol{a}_{0}$, giving

$$
\boldsymbol{v}^{*} \widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a}) \boldsymbol{v}=-\gamma^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{P}_{I} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \gamma,
$$

where

$$
\begin{equation*}
I(\boldsymbol{a})=\operatorname{supp}\left(\mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \iota \boldsymbol{a}\right]\right)=\left\{\left.i \in[n]| | \boldsymbol{x}_{0} * \breve{\boldsymbol{\beta}}\right|_{i}>\lambda\right\} . \tag{B.83}
\end{equation*}
$$

The $i$-th diagonal entry of $\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{P}_{I(a)} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}}$ is

$$
\begin{equation*}
-\boldsymbol{e}_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{e}_{i}=-\left\|\boldsymbol{P}_{I(\boldsymbol{a})} \check{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{e}_{i}\right\|_{2}^{2}=-\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2} \tag{B.84}
\end{equation*}
$$

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 Theorem B.4.2 and Theorem B.4.3. whose figure of visualized $\left\|\boldsymbol{P}_{I(\boldsymbol{a})}{ }^{s_{-i}}\left[\boldsymbol{x}_{0}\right]\right\|_{2}$ is shown in Figure B.1. Lastly, we also prove the off-diagonal terms $\boldsymbol{e}_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{e}_{j}$ of Hessian is likely inconsequential in calculation of curvature in Theorem B.4.4

Expectation of Hessian diagonals. We expect the Hessian to have stronger negative component in the $s_{i}\left[\boldsymbol{a}_{0}\right]$ direction as $\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}$ 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 $2 p-1$ samples, then by implementing the definition of support from (B.83), we can derive

$$
\begin{equation*}
-\left\|P_{I(a)}^{s_{-i}}\left[x_{0}\right]\right\|_{2}^{2}=-\sum_{j \in I_{0}} x_{0 j}^{2} \mathbf{1}_{\left\{\left|\sum_{\ell} \beta_{\ell} x_{0(\ell+j-i)}\right|>\lambda\right\}} \underbrace{=}_{\text {sep. }}-\sum_{j \in I_{0}} g_{j}^{2} \mathbf{1}_{\left\{\left|\boldsymbol{\beta}_{i} g_{j}\right|>\lambda\right\}}, \tag{B.85}
\end{equation*}
$$

where $\mathbf{1}$ is the indicator function and $\boldsymbol{g}_{j}$ are independent standard Gaussian r.v.s.. In expectation, the summands in (B.85) acts like a smoothed logic function on entry $\boldsymbol{\beta}_{i}$ :

Lemma B.4.1 (Gaussian smoothed indicator). Let $g \sim \mathcal{N}(0,1)$, then for any $b, s \in \mathbb{R}$ and $\lambda>0$.

$$
\begin{equation*}
\mathbb{E}_{g}\left[g^{2} \mathbf{1}_{\{|b \cdot g+s|>\lambda\}}\right]=1-\operatorname{erf}_{b}(\lambda, s)+f_{b}(\lambda, s) \tag{B.86}
\end{equation*}
$$

where

$$
\begin{equation*}
f_{b}(\lambda, s)=\frac{1}{\sqrt{2 \pi}}\left[\left(\frac{\lambda+s}{|b|}\right) e^{-\frac{(\lambda+s)^{2}}{2 b^{2}}}+\left(\frac{\lambda-s}{|b|}\right) e^{-\frac{(\lambda-s)^{2}}{2 b^{2}}}\right] . \tag{B.87}
\end{equation*}
$$

Proof. The proof can be derived via same calculation of integrals in Theorem B.3.1.

Although the definition B.86 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:

$$
\mathbb{E} g^{2} \mathbf{1}_{\{|b \cdot g|>\lambda\}} \approx \begin{cases}1, & |b| \approx 1  \tag{B.88}\\ 1 / 2, & |b| \approx \sqrt{2 / \pi} \lambda \\ 0, & |b| \approx 0\end{cases}
$$

Relate $(\overline{\mathrm{B} .88})$ to $\overline{\mathrm{B} .85}$, when $\left|\boldsymbol{\beta}_{i}\right|$ is close to 1 then we expect $-\frac{1}{n \theta}\left\|\boldsymbol{P}_{I} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}$ to be close to -1 , and it increases to 0 as $\left|\boldsymbol{\beta}_{i}\right|$ decreases, suggests that the Euclidean Hessian at point $\boldsymbol{a}$ has stronger negative component at $s_{i}\left[\boldsymbol{a}_{0}\right]$ direction if $\left|\left\langle\boldsymbol{a}, s_{i}\left[\boldsymbol{a}_{0}\right]\right\rangle\right|$ is larger. See Figure B. 2 for a numerical example. This phenomenon can be extend beyond the idealistic separating case as follows:

Lemma B.4.2 (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 (B.83), 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 $i \in[n]$ :

$$
\begin{equation*}
n^{-1} \mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}=\theta\left[1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)+\mathbb{E}_{\boldsymbol{s}_{i}} f_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)\right] \tag{B.89}
\end{equation*}
$$

Proof. Write $\boldsymbol{x}_{0}$ as $\boldsymbol{x}$. Observe that $\boldsymbol{y} * \breve{\boldsymbol{a}}=\boldsymbol{x}_{0} * \breve{\boldsymbol{\beta}}=\sum_{\ell} \boldsymbol{\beta}_{\ell} s_{-\ell}\left[\boldsymbol{x}_{0}\right]$. Thus for any $j \in[n]$ and $i \in[ \pm p]$ :

$$
\begin{equation*}
(\boldsymbol{y} * \breve{\boldsymbol{a}})_{j-i}=\left(\boldsymbol{\beta}_{i} s_{-i}[\boldsymbol{x}]+\sum_{\ell \neq i} \boldsymbol{\beta}_{\ell} s_{-\ell}[\boldsymbol{x}]\right)_{j-i}=\boldsymbol{\beta}_{i} \boldsymbol{x}_{j}+\sum_{\ell \neq i} \boldsymbol{\beta}_{\ell} \boldsymbol{x}_{j+\ell-i}=: \boldsymbol{\beta}_{i} \boldsymbol{x}_{j}+\boldsymbol{s}_{j} \tag{B.90}
\end{equation*}
$$

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 $j \in[n]$.


Figure B.2: A numerical example for $\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}$. We provide a figure to illustrate the expectation of $-\frac{1}{n \theta}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}$ when entries of $\boldsymbol{x}_{0}$ are $2 p$-separated, as a function plot of $\boldsymbol{\beta}_{i} \rightarrow 1-\operatorname{erf}_{\boldsymbol{\beta}_{i}}(\lambda, 0)+f_{\boldsymbol{\beta}_{i}}(\lambda, 0)$ from (B.86) with different $\lambda$. When $\left|\boldsymbol{\beta}_{i}\right| \approx \nu_{2} \lambda$ where $\nu_{2}=\sqrt{2 / \pi}$, then the its function value is close to 0.5 . If $\left|\boldsymbol{\beta}_{i}\right|$ is much larger then $\lambda$ its value grow to 1 , implies there is a negative curvature at $s_{i}\left[\boldsymbol{a}_{0}\right]$ direction. Similarly if $\left|\boldsymbol{\beta}_{i}\right|$ is much smaller then $\lambda$ the function value is 0 thus the curvature is positive in $s_{i}\left[\boldsymbol{a}_{0}\right]$ direction.

Rewrite the random variable using (B.83) as

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

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

$$
\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}=n \theta \cdot \mathbb{E} \boldsymbol{g}_{0}^{2} \mathbf{1}_{\left\{\left|\boldsymbol{\beta}_{i} \boldsymbol{g}_{0}+\boldsymbol{s}_{0}\right|>\lambda\right\}}=n \theta \cdot \mathbb{E}\left(1-\operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)+f_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)\right)
$$

where $\boldsymbol{s}_{i}=\sum_{\ell \neq i} \boldsymbol{x}_{0 i} \boldsymbol{\beta}_{i}$ with $\boldsymbol{x}_{0 i} \sim_{\text {i.i.d. }} \mathrm{BG}(\theta)$, 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 $\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}$ will be close enough to its expected value.

Corollary B.4.3 (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 $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $n \geq C p^{4} \theta^{-1} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-3 / n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ and every $i \in[n]$, we have:

$$
\begin{equation*}
\left|n^{-1}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}-n^{-1} \mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}\right| \leq c \theta / p \tag{B.91}
\end{equation*}
$$

Proof. See Appendix B.9.2

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 B.4.4 (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 $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$, where $\left|\boldsymbol{\beta}_{(1)}\right| \leq \frac{1}{4 \log \theta^{-1}} \lambda$ and every $i \neq j \in[ \pm p] \backslash\{(0)\}$, we have

$$
\begin{equation*}
\left|s_{i}\left[\boldsymbol{x}_{0}\right]^{*}\right| \boldsymbol{P}_{I(\boldsymbol{a})}\left|s_{j}\left[\boldsymbol{x}_{0}\right]\right|<8 n \theta^{3} \tag{B.92}
\end{equation*}
$$

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 $s^{\prime}=\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash\{0, i, j\}}, \boldsymbol{x}_{\boldsymbol{\tau} \backslash\{0, i, j\}}\right\rangle$. Firstly via Theorem B.2.7 we have $\mathbb{P}\left[\left|s^{\prime}\right|>0.4 \lambda\right] \leq 2 \theta$; also define $s=\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau}^{c} \backslash\{0, i, j\}}, \boldsymbol{x}_{\boldsymbol{\tau}^{c} \backslash\{0, i, j\}}\right\rangle$, and base on Theorem B.2.6 we have $\mathbb{P}[|s|>\lambda / 10] \leq 2 \theta$. Expand the $(-i,-j)$-th cross term with $\theta<0.1$ we have:

$$
\begin{align*}
\mathbb{E}\left|s_{-i}[\boldsymbol{x}]^{*}\right| \boldsymbol{P}_{I(\boldsymbol{a})}\left|s_{-j}[\boldsymbol{x}]\right| & =\mathbb{E} \sum_{k \in[n]}\left|\boldsymbol{x}_{k+i} \boldsymbol{x}_{k+j}\right| \mathbf{1}_{\left\{\left|\boldsymbol{\beta}_{0} \boldsymbol{x}_{k}+\boldsymbol{\beta}_{i} \boldsymbol{x}_{k+i}+\boldsymbol{\beta}_{j} \boldsymbol{x}_{k+j}+s+s^{\prime}\right|>\lambda\right\}} \\
& =n \theta^{2} \cdot \mathbb{E}\left|\boldsymbol{g}_{i} \boldsymbol{g}_{j}\right| \mathbf{1}_{\left\{\left|\boldsymbol{\beta}_{0} \boldsymbol{x}_{0}+\boldsymbol{\beta}_{i} \boldsymbol{g}_{i}+\boldsymbol{\beta}_{j} \boldsymbol{g}_{j}+s+s^{\prime}\right|>\lambda\right\}} \\
& \leq n \theta^{2} \cdot \mathbb{E}\left[\left|\boldsymbol{g}_{i} \boldsymbol{g}_{j}\right|\left(2 \mathbf{1}_{\left\{\left|\boldsymbol{\beta}_{i} \boldsymbol{g}_{i}\right|>\lambda / 4\right\}}+\mathbb{P}\left[\boldsymbol{x}_{0} \neq 0\right]+\mathbb{P}[|s|>0.1 \lambda]+\mathbb{P}\left[\left|s^{\prime}\right|>0.4 \lambda\right]\right)\right] \\
& \leq n \theta^{2} \cdot\left(\exp \left(-\log ^{2} \theta^{-1}\right)+\theta+2 \theta+2 \theta\right) \\
& \leq 6 n \theta^{3} . \tag{B.93}
\end{align*}
$$

Write (B.92) as two summation of independent random variables with $t=j-i$ by separating sum into two sets $J_{t 1}, J_{t 2}$ defined in (B.4) with both $\left|J_{t 1}\right|,\left|J_{t 2}\right|<n \theta^{2}$ with probability at least $1-2 / n$ from Theorem B.1.1

$$
\mathbb{E}\left|s_{-i}[\boldsymbol{x}]^{*}\right| \boldsymbol{P}_{I(\boldsymbol{a})}\left|s_{-j}[\boldsymbol{x}]\right|=\sum_{(k-i) \in I(\boldsymbol{a})}\left|\boldsymbol{x}_{k}\right|\left|\boldsymbol{x}_{k+t}\right|=\sum_{(k-i) \in I(\boldsymbol{a}) \cap J_{t 1}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k+t}\right|+\sum_{(k-i) \in I(\boldsymbol{a}) \cap J_{t 2}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k+t}\right|
$$

whose first summands can be upper bounded w.h.p. via Bernstein inequality Theorem B.10.4 with $\left(\sigma^{2}, R\right)=$ $(1,1)$ and writes $\mathcal{C}:=\cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right) \cap\left\{\boldsymbol{a}| | \boldsymbol{\beta}_{(1)} \left\lvert\, \leq \frac{1}{4 \log \theta^{-1}} \lambda\right.\right\}$, then we have

$$
\mathbb{P}\left[\max _{\substack{i \neq j \in[ \pm p \backslash \backslash\{0\} \\ \boldsymbol{a} \in \mathcal{C}}}\left(\sum_{(k-i) \in I(\boldsymbol{a}) \cap J_{t 1}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k+t}\right|-\mathbb{E} \sum_{(k-i) \in I(\boldsymbol{a}) \cap J_{t 1}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k+t}\right|\right) \geq n \theta^{3}\right]
$$

$$
\begin{align*}
& \mathbb{P}\left[\max _{i \neq j \in[ \pm p] \backslash 0\}}\left(\sum_{(k-i) \in \cap J_{t 1}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k+t}\right|-\mathbb{E} \sum_{(k-i) \cap J_{t 1}}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k+t}\right|\right) \geq n \theta^{3}\right] \\
& \leq 4 p^{2} \cdot \exp \left(\frac{-n^{2} \theta^{6}}{2\left|J_{t 1}\right|+2 n \theta^{3}}\right) \leq \exp \left(8 \log p-\frac{-n^{2} \theta^{6}}{3 n \theta^{2}}\right) \leq \exp \left(-\frac{n \theta^{4}}{10}\right) \leq \frac{1}{n} \tag{B.94}
\end{align*}
$$

when $n=C \theta^{-4} \log p$ with $C>10^{4}$ and $\theta \log ^{2} \theta^{-1} \geq 1 / p$. Thus for all $i \neq j \in[ \pm p] \backslash\{0\}$ and $\boldsymbol{a}$ satisfies our condition of lemma, from (B.93) and B.94 we can conclude :

$$
\left|s_{-i}[\boldsymbol{x}]^{*}\right| \boldsymbol{P}_{I(\boldsymbol{a})}\left|s_{-j}[\boldsymbol{x}]\right| \leq \sum_{I(\boldsymbol{a}) \cap J_{t 1}} \mathbb{E}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k+t}\right|+\sum_{I(\boldsymbol{a}) \cap J_{t 2}} \mathbb{E}\left|\boldsymbol{g}_{k}\right|\left|\boldsymbol{g}_{k+t}\right|+2 n \theta^{3} \leq 8 n \theta^{3}
$$

which holds with probability at least $1-2 / n-2 \cdot 1 / n=1-4 / n$ base on Theorem B.1.1 and (B.94).

## B. 5 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 Theorem B.5.2 that is smooth and well approximates $\ell^{1}$ when the proper smoothing parameter $\delta$ is chosen from the result of Theorem B.5.6. In particular we ask the regularizer $\rho_{\delta}(x)$ to be uniformly bounded to $|x|$ by $\delta / 2$ :

$$
\begin{equation*}
\forall x \in \mathbb{R}, \quad\left|\rho_{\delta}(x)-|x|\right| \leq \delta / 2 \tag{B.95}
\end{equation*}
$$

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

$$
\begin{align*}
&\left\|\operatorname{prox}_{\lambda^{1}}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}]-\operatorname{prox}_{\lambda \rho_{\delta}}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}]\right\|_{2} \rightarrow 0,  \tag{B.96}\\
&\left\|\nabla \varphi_{\ell^{1}}(\boldsymbol{a})-\nabla \varphi_{\rho_{\delta}}(\boldsymbol{a})\right\|_{2} \rightarrow 0,  \tag{B.97}\\
&\left\|\widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a})-\nabla^{2} \varphi_{\rho_{\delta}}(\boldsymbol{a})\right\|_{2} \rightarrow 0 . \tag{B.98}
\end{align*}
$$

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

Calculus of $\varphi_{\rho}$. The marginal minimizer over $x$ in (3.8) can be expressed in terms of the proximal operator [BC11] of $\rho$ at point $\check{a} * y$ :

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

Plugging in, we obtain

$$
\begin{equation*}
\varphi_{\rho}(\boldsymbol{a})=\lambda \rho\left(\operatorname{prox}_{\lambda \rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]\right)+\frac{1}{2}\left\|\check{\boldsymbol{a}} * \boldsymbol{y}-\operatorname{prox}_{\lambda \rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]\right\|_{2}^{2}-\frac{1}{2}\|\check{\boldsymbol{a}} * \boldsymbol{y}\|_{2}^{2}+\frac{1}{2}\|\boldsymbol{y}\|_{2}^{2} \tag{B.99}
\end{equation*}
$$

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

$$
\begin{equation*}
\varphi_{\rho}(\boldsymbol{a})=\epsilon(\lambda \rho)(\check{\boldsymbol{a}} * \boldsymbol{y})-\frac{1}{2}\|\check{\boldsymbol{a}} * \boldsymbol{y}\|_{2}^{2}+\frac{1}{2}\|\boldsymbol{y}\|_{2}^{2} \tag{B.100}
\end{equation*}
$$

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 B.5.1 (Derivative of Moreau envelope, [BC11], Prop.12.29). Let $f$ be a proper lower semicontinuous convex function and $\lambda>0$ then the Moreau envelope $\epsilon(\lambda f)(\boldsymbol{z})=\lambda f\left(\operatorname{prox}_{\lambda f}[\boldsymbol{z}]\right)+\frac{1}{2}\left\|\boldsymbol{z}-\operatorname{prox}_{\lambda f}[\boldsymbol{z}]\right\|_{2}^{2}$ 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

$$
\begin{align*}
\nabla \varphi_{\rho}(\boldsymbol{a}) & =-\iota^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}} \operatorname{prox}_{\lambda \rho}\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \iota \boldsymbol{a}\right]  \tag{B.101}\\
\nabla^{2} \varphi_{\rho}(\boldsymbol{a}) & =-\iota^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}} \nabla \operatorname{prox}_{\lambda \rho}\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \iota \boldsymbol{a}\right] \breve{\boldsymbol{C}}_{\boldsymbol{y}} \iota \tag{B.102}
\end{align*}
$$

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

$$
\begin{align*}
& \operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})=-\boldsymbol{P}_{\boldsymbol{a}^{\perp}} \boldsymbol{\iota}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}} \operatorname{prox}_{\lambda \rho}\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{\iota} \boldsymbol{a}\right]  \tag{B.103}\\
& \operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a})=-\boldsymbol{P}_{\boldsymbol{a}^{\perp}}\left(\boldsymbol{\iota}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}} \nabla \operatorname{prox}_{\lambda \rho}\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{\iota} \boldsymbol{a}\right] \breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{\iota}-\left\langle\nabla \varphi_{\rho}(\boldsymbol{a}), \boldsymbol{a}\right\rangle \boldsymbol{I}\right) \boldsymbol{P}_{\boldsymbol{a}^{\perp}} \tag{B.104}
\end{align*}
$$

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 B.5.2 ( $\delta$-smoothed $\ell^{1}$ function). We call an additively separable function $\rho(\boldsymbol{x})=\sum_{i=1}^{n} \rho_{i}\left(\boldsymbol{x}_{i}\right): \mathbb{R}^{n} \rightarrow \mathbb{R}$, 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 $x \in \mathbb{R}$ :

$$
\begin{equation*}
\left|\rho_{i}(x)-|x|+c\right| \leq \delta / 2 \tag{B.105}
\end{equation*}
$$

The proximal operator of the $\ell^{1}$ norm is the entrywise soft thresholding function $\mathcal{S}_{\lambda}$; the proximal operator

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

Table B.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 $\rho(0)-|x| \leq \delta$, satisfies the condition B.105. 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.
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 B.5.3 ( $\sqrt{\delta}$-smoothed soft threshold). An odd function $\mathcal{S}_{\lambda}^{\delta}[\cdot]: \mathbb{R} \rightarrow \mathbb{R}$ 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

$$
\begin{equation*}
0 \leq \operatorname{sign}(z)\left(\mathcal{S}_{\lambda}^{\delta}[z]-\mathcal{S}_{\lambda}[z]\right) \leq \sqrt{\lambda \delta}, \quad \forall z \in \mathbb{R} \tag{B.106}
\end{equation*}
$$

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

$$
\begin{equation*}
\left|\nabla \mathcal{S}_{\lambda}^{\delta}[z]-\nabla \mathcal{S}_{\lambda}[z]\right| \leq \sqrt{\lambda \delta} / B, \quad| | z|-\lambda| \geq B \tag{B.107}
\end{equation*}
$$

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 B.5.4 (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

$$
\begin{equation*}
\boldsymbol{x}_{z}:=\operatorname{prox}_{\lambda \rho}[\boldsymbol{z}]=\underset{\boldsymbol{x} \in \mathbb{R}^{n}}{\operatorname{argmin}} \lambda \rho(\boldsymbol{x})+\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_{2}^{2} \tag{B.108}
\end{equation*}
$$

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 $\lambda \nabla \rho_{i}\left(\boldsymbol{x}_{z i}\right)+\boldsymbol{x}_{z i}-\boldsymbol{z}_{i}=0$, implies

$$
\begin{equation*}
\boldsymbol{x}_{z i}=\left(\operatorname{Id}+\lambda \nabla \rho_{i}\right)^{-1}\left(\boldsymbol{z}_{i}\right) . \tag{B.109}
\end{equation*}
$$

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 $\nabla \boldsymbol{x}_{z i}$ has the form

$$
\begin{equation*}
\nabla \boldsymbol{x}_{z i}=\nabla_{i}\left(\mathrm{Id}+\lambda \nabla \rho_{i}\right)^{-1}\left(\boldsymbol{z}_{i}\right)=\frac{1}{\lambda \nabla^{2} \rho_{i}\left(\boldsymbol{x}_{z i}\right)+1}<1 . \tag{B.110}
\end{equation*}
$$

Notice that since $\nabla^{2} \rho_{i}(x)$ is monotone decreasing when $x \geq 0$, hence $\nabla \boldsymbol{x}_{z i}$ is monotone increasing in $\boldsymbol{z}_{i} \geq 0$.
Now we are left to show that (B.106) and B.107) hold, and since 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 $\left|\rho_{i}(x)-|x|\right| \leq \delta / 2$ we are going to show

$$
\begin{equation*}
\left|\nabla \rho_{i}(x)\right| \leq 1 \quad \forall x \in \mathbb{R}, \quad \nabla \rho_{i}(x) \geq 1-\sqrt{\delta / \lambda} \quad \forall x \geq \sqrt{\lambda \delta} . \tag{B.111}
\end{equation*}
$$

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 (B.106), when $0 \leq \boldsymbol{z}_{i} \leq \lambda$, then $\mathcal{S}_{\lambda}\left[\boldsymbol{z}_{i}\right]=0$ and $\boldsymbol{x}_{z i} \leq \sqrt{\lambda \delta}$ since if $\boldsymbol{x}_{z i}>\sqrt{\lambda \delta}$, by (B.111):

$$
\lambda \nabla \rho_{i}\left(\boldsymbol{x}_{z i}\right)+\boldsymbol{x}_{z i}>\lambda(1-\sqrt{\delta / \lambda})+\sqrt{\lambda \delta}=\lambda \geq \boldsymbol{z}_{i}
$$

then $\boldsymbol{x}_{z i}$ violate the stationary condition in (B.109, resulting $0 \leq \boldsymbol{x}_{z i}-\mathcal{S}_{\lambda}\left[\boldsymbol{z}_{i}\right] \leq \sqrt{\lambda \delta}$ whenever $0 \leq \boldsymbol{z}_{i} \leq \lambda$. Likewise in the case of $\boldsymbol{z}_{i} \geq \lambda$ where $\mathcal{S}_{\lambda}\left[\boldsymbol{z}_{i}\right]=\boldsymbol{z}_{i}-\lambda$, B.111 provides:

$$
\begin{cases}\forall \boldsymbol{x}_{z i}>\boldsymbol{z}_{i}-\lambda+\sqrt{\lambda \delta}, & \lambda \nabla \rho_{i}\left(\boldsymbol{x}_{z i}\right)+\boldsymbol{x}_{z i}>\lambda(1-\sqrt{\delta / \lambda})+\boldsymbol{z}_{i}-\lambda+\sqrt{\lambda \delta}=\boldsymbol{z}_{i} \\ \forall \boldsymbol{x}_{z i}<\boldsymbol{z}_{i}-\lambda, & \lambda \nabla \rho_{i}\left(\boldsymbol{x}_{z i}\right)+\boldsymbol{x}_{z i}<\lambda+\boldsymbol{z}_{i}-\lambda=\boldsymbol{z}_{i}\end{cases}
$$

again violates B.109 and therefore B.106 holds for all $\boldsymbol{z}_{i} \in \mathbb{R}$.
Lastly B.107) is a direct result of B.106). For all $\boldsymbol{z}_{i} \leq \lambda-B$, recall that $\nabla \boldsymbol{x}_{z i}$ is monotone increasing in $\boldsymbol{z}_{i}$ :

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

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

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

implies B.107 holds.

Approximate geometry of $\varphi_{\rho}$ using $\varphi_{\ell^{1}} \quad$ Based on B.103-B.104 and denote $\breve{\boldsymbol{C}}_{\boldsymbol{y}} \iota \boldsymbol{a}=\check{\boldsymbol{a}} * \boldsymbol{y}$, the only differences of Riemannian gradient and Hessian between $\varphi_{\rho}$ and $\varphi_{\ell^{1}}$ comes from the difference of prox $\lambda_{\lambda \rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]$ and $\operatorname{prox}_{\lambda\|\cdot\|_{1}}[\check{\boldsymbol{a}} * \boldsymbol{y}]$. Thus for the purpose of obtaining good geometric approximation of $\varphi_{\rho}$ with that of objective $\varphi_{\ell^{1}}$, we may apply both Theorem B.5.3 and Theorem B.5.4. together suggest if $\rho$ is a $\delta$-smoothed $\ell^{1}$ function, then the $i$-th entry of $\operatorname{prox}_{\lambda \rho}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}]$ will be $\sqrt{\lambda \delta}$-close to the authentic soft thresholding function $\mathcal{S}_{\lambda}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}]_{i}$, and its gradient $\nabla \operatorname{prox}_{\lambda \rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]$ is $\sqrt{\lambda \delta} / B$-close to $\nabla \mathcal{S}_{\lambda}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}]$ 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 $\check{\boldsymbol{a}} * \boldsymbol{y}$ will be close to $\pm \lambda$.

Lemma B.5.5 (Gradients discontinuity entries). For each $\boldsymbol{a} \in \mathbb{S}^{p-1}$, let

$$
\begin{equation*}
J_{B}(\boldsymbol{a}):=\left\{i \mid\left(\check{\boldsymbol{C}}_{\boldsymbol{y}} \iota \boldsymbol{a}\right)_{i} \in[-\lambda-B,-\lambda+B] \cup[\lambda-B, \lambda+B]\right\} . \tag{B.112}
\end{equation*}
$$

Suppose the subspace dimension is at most $k$ and signal $\boldsymbol{y}$ satisfies Theorem B.2.1 Let $\lambda=c_{\lambda} / \sqrt{k}$ and $B \leq$ $c^{\prime} \lambda \theta^{2} / p \log n$ for some $c_{\lambda}, c^{\prime} \in(0,1)$, then there is a numerical constant $C>0$ such that if $n \geq C p^{5} \theta^{-2} \log p$, then with probability at least $1-3 / n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$, we have

$$
\begin{equation*}
\left|J_{B}(\boldsymbol{a})\right| \leq \frac{24 c^{\prime} n \theta^{2}}{p \log n} \tag{B.113}
\end{equation*}
$$

Proof. See Appendix B.9.3.
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 B.5.6 $\left(\varphi_{\ell^{1}}\right.$ approximates $\left.\varphi_{\rho}\right)$. Suppose $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \mathrm{BG}(\theta)$ in $\mathbb{R}^{n}$, and $k, c_{\mu}$ such that $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. Let $\rho \in \mathbb{R}^{n} \rightarrow \mathbb{R}$ be a $\delta$-smoothed $\ell^{1}$ function with

$$
\begin{equation*}
\lambda=\frac{c_{\lambda}}{\sqrt{k}}, \quad \delta \leq \frac{c^{\prime 4} \theta^{8}}{p^{2} \log ^{2} n} \lambda \tag{B.114}
\end{equation*}
$$

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

$$
\begin{equation*}
\left\|\boldsymbol{\iota}_{[ \pm p]}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \operatorname{prox}_{\lambda \ell^{1}}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}]-\boldsymbol{\iota}_{[ \pm p]}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \operatorname{prox}_{\lambda \rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]\right\|_{2} \leq c^{\prime} n \theta^{4} \tag{B.115}
\end{equation*}
$$

(2). The gradient has norm difference

$$
\begin{equation*}
\left\|\nabla \varphi_{\ell^{1}}(\boldsymbol{a})-\nabla \varphi_{\rho}(\boldsymbol{a})\right\|_{2} \leq c^{\prime} n \theta^{4} \tag{B.116}
\end{equation*}
$$

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

$$
\begin{equation*}
\forall \boldsymbol{v} \in \mathbb{S}^{p-1}, \quad\left|\boldsymbol{v}^{*}\left(\widetilde{\operatorname{Hess}}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})-\operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a})\right) \boldsymbol{v}\right| \leq 200 c^{\prime} n \theta^{2} \tag{B.117}
\end{equation*}
$$

Proof. 1. (Coefficients) From Theorem B.5.4 the proximal $\delta$-smoothed $\ell^{1}$ function satisfies

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

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

$$
\begin{equation*}
\left\|\boldsymbol{\iota}_{[ \pm p]}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}[\check{\boldsymbol{a}} * \boldsymbol{y}]-\boldsymbol{\iota}_{[ \pm p]}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}^{\delta}[\check{\boldsymbol{a}} * \boldsymbol{y}]\right\|_{2} \leq \sqrt{\lambda \delta n} \cdot\left\|\boldsymbol{\iota}_{[ \pm p]}^{*} \check{\boldsymbol{C}}_{\boldsymbol{x}_{0}}\right\|_{2} \tag{B.118}
\end{equation*}
$$

Apply Theorem B.1.5 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

$$
\left\|\boldsymbol{\iota}_{[ \pm p]}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}}\right\|_{2} \leq\left\|\boldsymbol{\iota}_{[ \pm p]}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{e}_{0}}^{*}\right\|_{2} \leq \sqrt{3(1+2 \mu p) n \theta} \leq \sqrt{3 n \theta}
$$

and we gain

$$
\overline{\mathrm{B} .118} \leq \sqrt{\lambda \delta n} \cdot \sqrt{3 n \theta} \leq n \sqrt{3 \lambda \theta \delta} \leq c^{\prime} n \theta^{4} .
$$

2. (Gradient) From definition of Riemannian gradient B.103) and apply similar norm bound of B.118, and condition on the following events of Theorem B.1.5 holds, obtain

$$
\begin{equation*}
\left\|\nabla \varphi_{\ell^{1}}(\boldsymbol{a})-\nabla \varphi_{\rho}(\boldsymbol{a})\right\|_{2} \leq \sqrt{\lambda \delta n} \cdot\left\|\iota^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}}\right\|_{2} \leq n \sqrt{3 \lambda \theta(1+\mu p) \delta} \leq c^{\prime} n \theta^{4} \tag{B.119}
\end{equation*}
$$

3. (Hessian) For every realization of $J_{B}(\boldsymbol{a})$ from $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$, base on Theorem B.5.5, condition
on the event such that

$$
\begin{equation*}
B \leq \frac{c^{\prime} \lambda \theta^{2}}{p \log n}, \quad|J| \leq \frac{24 c^{\prime} n \theta^{2}}{p \log n} \tag{B.120}
\end{equation*}
$$

and rewrite $J_{B}(\boldsymbol{a})$ as $J$. Also condition on the event using Theorem B.1.5 and $(1+\mu p) \theta \log \theta^{-1}<1$

$$
\begin{equation*}
\left\|\boldsymbol{\iota}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}}\right\|_{2} \leq \sqrt{3 n}, \quad\left\|\iota^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{P}_{J}\right\|_{2} \leq \sqrt{8|J| p \log n} \tag{B.121}
\end{equation*}
$$

then the difference of Hessian $\overline{B .104}$, in direction $\boldsymbol{v} \in \mathbb{S}^{p-1}$ can be bounded as

$$
\begin{align*}
& \left|\boldsymbol{v}^{*}\left(\widetilde{\operatorname{Hess}}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})-\operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a})\right) \boldsymbol{v}\right| \\
\leq & \left|\boldsymbol{v}^{*} \iota^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}}\left(\boldsymbol{P}_{I(\boldsymbol{a})}-\operatorname{diag}\left[\nabla \mathcal{S}_{\lambda}^{\delta}\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \iota \boldsymbol{a}\right]\right]\right) \breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{\iota} \boldsymbol{v}\right|+\left\|\nabla \varphi_{\ell^{1}}(\boldsymbol{a})-\nabla \varphi_{\rho}(\boldsymbol{a})\right\|_{2} \tag{B.122}
\end{align*}
$$

where $I(\boldsymbol{a})$ is defined in B.83). Let $\boldsymbol{D}=\boldsymbol{P}_{I(\boldsymbol{a})}-\operatorname{diag}\left[\nabla \mathcal{S}_{\lambda}^{\delta}\left[\check{\boldsymbol{C}}_{\boldsymbol{y}} \iota \boldsymbol{a}\right]\right]$ and notice that $\boldsymbol{D}$ is a diagonal matrix, which suggests B.122 can be decomposed using

$$
\left(\boldsymbol{P}_{J}+\boldsymbol{P}_{J^{c}}\right) \boldsymbol{D}\left(\boldsymbol{P}_{J}+\boldsymbol{P}_{J^{c}}\right)=\boldsymbol{P}_{J} \boldsymbol{D} \boldsymbol{P}_{J}+\boldsymbol{P}_{J^{c}} \boldsymbol{D} \boldsymbol{P}_{J^{c}}
$$

where, from with property of $\sqrt{\delta}$-smoothed $\ell^{1}$ function Theorem B.5.4

$$
\max _{j}\left|\boldsymbol{P}_{J} \boldsymbol{D} \boldsymbol{P}_{J}\right|_{j j} \leq 1, \quad \max _{j}\left|\boldsymbol{P}_{J^{c}} \boldsymbol{D} \boldsymbol{P}_{J^{c}}\right|_{j j} \leq \sqrt{\lambda \delta} / B
$$

Finally, once again apply $\delta$ bound from $(\bar{B} .114)$ and bounds for $B,|J|, \boldsymbol{y}$ from $(\bar{B} .120)-(\overline{B .121}$, we gain

$$
\begin{aligned}
\text { (B.122) } & \leq\left\|\iota^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{P}_{J}\right\|_{2}^{2}+\frac{\sqrt{\lambda \delta}}{B}\left\|\boldsymbol{\iota}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}}\right\|_{2}^{2}+\left\|\nabla \varphi_{\ell^{1}}(\boldsymbol{a})-\nabla \varphi_{\rho}(\boldsymbol{a})\right\|_{2} \\
& \leq 8|J| p \log n+\frac{3 n \sqrt{\lambda \delta}}{B}+c^{\prime} n \theta^{2} \\
& \leq 8 \cdot \frac{24 c^{\prime} n \theta^{2}}{p \log n} \cdot p \log n+\frac{3 n\left(c^{\prime 4} \lambda^{2} \theta^{8} / p^{2} \log ^{2} n\right)^{1 / 2}}{c^{\prime} \lambda \theta^{2} / p \log p}+c^{\prime} n \theta^{2} \\
& \leq 200 c^{\prime} n \theta^{2}
\end{aligned}
$$

where all above result holds with probability at least $1-10 / n$ from Theorem B.5.5 and Theorem B.1.5.

## B. 6 Analysis of geometry

In this section we prove major geometrical result in Theorem 3.4.1. This lemma consists of three parts of geometry of $\varphi_{\rho}$; including the negative curvature region Theorem B.6.2. large gradient region Theorem B.6.4. strong convexity region near shift Theorem B.6.6. and retraction to subspace Theorem B.6.8, which are
respectively base on geometric properties of $\varphi_{\ell^{1}}$ in Theorem B.6.1. Theorem B.6.3. Theorem B.6.5 and Theorem B.6.7. 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 B.3: 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 $s_{i}\left[\boldsymbol{a}_{0}\right]$ and $s_{j}\left[\boldsymbol{a}_{0}\right]$. When $\boldsymbol{a}$ lies near middle of two shifts (light green region) such that $\left|\boldsymbol{\beta}_{i}\right| \approx\left|\boldsymbol{\beta}_{j}\right|$, then there exists a negative curvature direction in subspace $\mathcal{S}_{\{i, j\}}$. When $\boldsymbol{a}$ leans closer to one of the shifts $s_{i}\left[\boldsymbol{a}_{0}\right]$ (blue green region), its negative gradient direction points at that nearest shift. When $\boldsymbol{a}$ is in the neighborhood of the shift $s_{i}\left[\boldsymbol{a}_{0}\right]$ (dark green region) such that $\left|\boldsymbol{\beta}_{i}\right| \ll \lambda$, it will be strongly convex at $\boldsymbol{a}$, and the unique minimizer within the convex region will be close to $s_{i}\left[\boldsymbol{a}_{0}\right]$. Finally, the negative gradient will be pointing back toward the subspace $\mathcal{S}_{\{i, j\}}$ if near boundary (grey region).


Figure B.4: 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 $\boldsymbol{v}^{*} \operatorname{Hess}[\varphi](\boldsymbol{a}) \boldsymbol{v}$ is negative. In large gradient region, the norm of Riemannian gradient $\|\operatorname{grad}[\varphi](\boldsymbol{a})\|_{2}$ will be strictly greater then 0 and pointing at the nearest shift. Finally there is a convex region near all shifts such that Hess $[\varphi](\boldsymbol{a})$ is positive semidefinite.

## B.6.1 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 $s_{(0)}\left[\boldsymbol{a}_{0}\right], s_{(1)}\left[\boldsymbol{a}_{0}\right]$. We will first demonstrate the pseudo negative curvature of $\varphi_{\ell^{1}}$ in Theorem B.6.1. then show $\varphi_{\rho}$ approximates $\varphi_{\ell^{1}}$ in terms of Hessian in Theorem B.6.2 when $\rho$ is properly defined as in Appendix B. 5

Lemma B.6.1 (Negative curvature for $\left.\varphi_{\ell^{1}}\right)$. Suppose that $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ in $\mathbb{R}^{n}$, and $k, c_{\mu}$ such that $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $C, c, c^{\prime}, \bar{c}>0$ such that if $n>C p^{5} \theta^{-2} \log p$, and $c_{\mu} \leq \bar{c}$, then with probability at least $1-c^{\prime} / n$ the following holds at every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ satisfying $\left|\boldsymbol{\beta}_{(1)}\right| \geq \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|$ : for $\boldsymbol{v} \in \mathcal{S}_{\{(0),(1)\}} \cap \mathbb{S}^{p-1} \cap \boldsymbol{a}^{\perp}$,

$$
\begin{equation*}
\boldsymbol{v}^{*} \widetilde{\operatorname{Hess}}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a}) \boldsymbol{v} \leq-\operatorname{cn} \theta \lambda \tag{B.123}
\end{equation*}
$$

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

$$
\begin{equation*}
0.79 \geq \frac{\left|\boldsymbol{\beta}_{(0)}\right|}{\sqrt{\boldsymbol{\beta}_{(0)}^{2}+\boldsymbol{\beta}_{(1)}^{2}}}\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2} \geq\left|\boldsymbol{\beta}_{(0)}\right| \geq\left|\boldsymbol{\beta}_{(1)}\right| \geq \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right| \geq \frac{4}{5} \cdot \frac{\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}}{\sqrt{|\boldsymbol{\tau}|}} \geq \frac{0.79}{\sqrt{|\boldsymbol{\tau}|}} \tag{B.124}
\end{equation*}
$$

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

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

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

$$
\begin{equation*}
\gamma_{(0)}^{2} \geq \frac{(1-\mu) \cdot 16}{25+16} \geq 0.385 \tag{B.125}
\end{equation*}
$$

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

$$
\begin{equation*}
\boldsymbol{v}^{*} \widetilde{\operatorname{Hess}}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a}) \boldsymbol{v}=\boldsymbol{v}^{*} \widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a}) \boldsymbol{v}-\left\langle\nabla \varphi_{\ell^{1}}(\boldsymbol{a}), \boldsymbol{a}\right\rangle=-\boldsymbol{\gamma}^{*} \boldsymbol{\iota}_{J}^{*} \boldsymbol{M} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{M} \boldsymbol{\iota}_{J} \boldsymbol{\gamma}+\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \tag{B.126}
\end{equation*}
$$

expand the first term of B.126 we obtain

$$
\begin{aligned}
&-\gamma^{*} \iota_{J}^{*} \boldsymbol{M} \breve{C}_{\boldsymbol{x}} \boldsymbol{P}_{I(a)} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{M} \iota_{J} \gamma \\
&=-\gamma^{*} \iota_{J}^{*} \boldsymbol{M}\left(\boldsymbol{P}_{(0)}+\boldsymbol{P}_{(1)}+\boldsymbol{P}_{J^{c}}\right) \breve{C}_{\boldsymbol{x}} \boldsymbol{P}_{I(a)} \breve{\boldsymbol{C}}_{\boldsymbol{x}}\left(\boldsymbol{P}_{(0)}+\boldsymbol{P}_{(1)}+\boldsymbol{P}_{J^{c}}\right) \boldsymbol{M} \iota_{J} \gamma
\end{aligned}
$$

$$
\begin{align*}
\leq & -\sum_{i \in J}\left\|\boldsymbol{P}_{I(a)} \breve{C}_{\boldsymbol{x}} \boldsymbol{e}_{i}\right\|_{2}^{2}\left(e_{i}^{*} \boldsymbol{M} \iota_{J} \gamma\right)^{2}+2 \sum_{\substack{(i, j) \in\left\{J, J^{c}\right\} \\
(i, j)=((0),(1))}}\left|e_{i}^{*} \breve{C}_{\boldsymbol{x}} \boldsymbol{P}_{I(a)} \breve{C}_{\boldsymbol{x}} \boldsymbol{e}_{j}\right|\left|\left(e_{i}^{*} \boldsymbol{M} \iota_{J} \gamma\right)\left(e_{j}^{*} \boldsymbol{M} \iota_{J} \gamma\right)\right| \\
\leq & -\sum_{i \in J}\left\|\boldsymbol{P}_{I(a)} \breve{C}_{\boldsymbol{x}} \boldsymbol{e}_{i}\right\|_{2}^{2}\left(\left|\gamma_{i}\right|-\mu\right)^{2} \\
& +2 \max _{i \neq j \in[ \pm p]}\left|e_{i}^{*} \breve{C}_{\boldsymbol{x}} \boldsymbol{P}_{I(a)} \breve{C}_{\boldsymbol{x}} \boldsymbol{e}_{j}\right|\left(\left\|\iota_{J}^{*} M \iota_{J} \gamma\right\|_{1}\left\|\iota_{J{ }_{c}}^{*} M \iota_{J} \gamma\right\|_{1}+\left(\left|\gamma_{(0)}\right|+\mu\right)\left(\left|\gamma_{(1)}\right|+\mu\right)\right) \tag{B.127}
\end{align*}
$$

Consider the following events

$$
\left\{\begin{array}{l}
\mathcal{E}_{\text {cross }}:=\left\{\forall \boldsymbol{a} \in \mathbb{S}^{p-1}, \max _{i \neq j \in[ \pm p]}\left|e_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{e}_{j}\right|<4 n \theta^{2}\right\}  \tag{B.128}\\
\mathcal{E}_{\text {ncurv }}:=\left\{\forall \boldsymbol{a} \in \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right), \min _{i \in J}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}]\right\|_{2}^{2} \geq n \theta\left(1-\mathbb{E}_{s_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)+\mathbb{E}_{\boldsymbol{s}_{i}}\left(\lambda, s_{i}\right)\right)-\frac{c_{\mu} n \theta}{p}\right\}
\end{array}\right.
$$

and from Theorem B.2.4 we know

$$
\left\|\iota_{J}^{*} \boldsymbol{M} \iota_{J} \boldsymbol{\gamma}\right\|_{1} \leq\|\boldsymbol{\gamma}\|_{1}+2 \mu \leq 1.5, \quad\left\|\iota_{J c}^{*} \boldsymbol{M} \iota_{J} \boldsymbol{\gamma}\right\|_{1} \leq \mu p\|\gamma\|_{1} \leq 1.5 \mu p,
$$

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

$$
\begin{align*}
& -\boldsymbol{\gamma}^{*} \iota_{J}^{*} \boldsymbol{M} \check{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{P}_{I(\boldsymbol{a})} \check{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{M} \iota_{\boldsymbol{\iota}} \boldsymbol{\gamma} \\
& \quad \leq \underbrace{-n \theta \cdot \sum_{i \in J}\left(\left|\boldsymbol{\gamma}_{i}\right|-\mu\right)^{2}\left(1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)+\mathbb{E}_{\boldsymbol{s}_{i}} f_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)\right)}_{g_{1}(\boldsymbol{\beta})}+(18 \mu p+8) n \theta^{2}+\frac{2 c_{\mu} n \theta}{\sqrt{|\boldsymbol{\tau}|}} \tag{B.129}
\end{align*}
$$

Meanwhile, for the latter term of (B.126), consider the following event $\mathcal{E}_{\bar{\chi}}$ where we write $\sigma_{i}=\operatorname{sign}\left(\boldsymbol{\beta}_{i}\right)$ as:
and use both $\|\boldsymbol{\beta}\|_{1} \leq \frac{c_{\mu} p}{\sqrt{|\boldsymbol{\tau}|}},\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \leq \frac{c_{\mu}}{\theta|\boldsymbol{\tau}|^{2}}$. On this event we have

$$
\begin{align*}
\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] & \leq n \theta \cdot \sum_{i \in \boldsymbol{\tau}} \boldsymbol{\beta}_{i}^{2}\left(1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)\right)+4 n \theta^{2}|\boldsymbol{\tau}|\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}+\frac{c_{\mu} n \theta}{p}\|\boldsymbol{\beta}\|_{1} \\
& \leq \underbrace{n \theta \cdot \sum_{i \in \boldsymbol{\tau}} \boldsymbol{\beta}_{i}^{2}\left(1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)\right)}_{g_{2}(\boldsymbol{\beta})}+\frac{5 c_{\mu} n \theta}{\sqrt{|\boldsymbol{\tau}|}} \tag{B.131}
\end{align*}
$$

2. (Lower bound $\mathbb{E} f_{\mathcal{\beta}_{i}}$ ) Combine the first term from each of the B.129) and (B.131). Use $\mu \leq c_{\mu} \leq \frac{1}{300}$ and
B.125) to obtain $\left(\left|\gamma_{(0)}\right|-\mu\right)^{2}>0.38$, we have

$$
\begin{align*}
\frac{1}{n \theta}\left(g_{1}(\boldsymbol{\beta})+g_{2}(\boldsymbol{\beta})\right) \leq & -\sum_{i \in J}\left[\left(\left|\gamma_{i}\right|-\mu\right)^{2}-\boldsymbol{\beta}_{i}^{2}\right]\left(1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)\right) \\
& +\sum_{i \in \boldsymbol{\mathcal { T }} \backslash J} \boldsymbol{\beta}_{i}^{2}\left(1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)\right)-0.38 \sum_{i \in J} \mathbb{E}_{\boldsymbol{s}_{i}} f_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right), \tag{B.132}
\end{align*}
$$

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

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

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

$$
\left\{\begin{array}{l}
f^{\prime}\left(\beta_{*}\right)=0 \Longrightarrow 2 \lambda \beta_{*}^{2}=3 \lambda\left(\lambda^{2}+\frac{9 c_{\mu}}{|\tau|}\right) \\
f^{\prime \prime}\left(\beta_{*}\right)=\frac{1}{\left|\beta_{*}\right|^{3}}\left(4 \lambda-\frac{12 \lambda}{\beta_{*}^{2}}\left(\lambda^{2}+\frac{9 c_{\mu}}{|\tau|}\right)\right)=\frac{1}{\left|\beta_{*}\right|^{3}}\left(4 \lambda-\frac{12}{3 / 2} \lambda\right)<0,
\end{array},\right.
$$

then combine with regional condition (B.124), and also apply assumption $c_{\lambda} \leq \frac{1}{3}$ and $c_{\mu} \leq \frac{1}{300}$, we gain

$$
\begin{align*}
0.38 \sum_{i \in J} \mathbb{E}_{\boldsymbol{s}_{i}} f_{\mathcal{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right) & \geq 0.3 \min _{\beta=\frac{0.79}{\sqrt{|T|}, 0.79}} f(\beta) \\
& \geq 0.3 \min \left\{\frac{2 c_{\lambda}}{0.79}-\frac{c_{\lambda}^{3}+9 c_{\mu} c_{\lambda}}{0.79^{3}}, \lambda\left(\frac{2}{0.79}-\frac{c_{\lambda}^{2}+9 c_{\mu}}{0.79^{3}}\right)\right\} \\
& \geq 0.3 \min \left\{2 c_{\lambda}, 2 \lambda\right\} \geq 0.6 \lambda . \tag{B.133}
\end{align*}
$$

3. (Upper bound $\left.\mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}\right)$ When $\boldsymbol{\beta}_{(0)}^{2}=\left(\left|\gamma_{(0)}\right|-\mu\right)^{2}-\eta$ for some $\eta>0$. With monotonicity Theorem B.3.3. which implies:

$$
\begin{equation*}
\left(1-\mathbb{E}_{\boldsymbol{s}_{(0)}} \operatorname{erf}_{\boldsymbol{\beta}_{(0)}}\left(\lambda, \boldsymbol{s}_{(0)}\right)\right) \geq\left(1-\mathbb{E}_{\boldsymbol{s}_{(1)}} \operatorname{erf}_{\boldsymbol{\beta}_{(1)}}\left(\lambda, \boldsymbol{s}_{(1)}\right)\right) \geq\left(1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)\right), \tag{B.134}
\end{equation*}
$$

then combine ( $\overline{\text { B.133 }}-\left(\overline{\text { B.134 }}\right.$ and use $\mu \leq \frac{c_{\mu}}{4 \sqrt{|\tau|}}$ from Theorem B.2.5

$$
\begin{aligned}
\overline{(\overline{B .132)} \leq} & -\underbrace{\left(\left(\left|\boldsymbol{\gamma}_{(0)}\right|^{2}-\mu\right)^{2}-\boldsymbol{\beta}_{(0)}^{2}-\eta\right)}_{=0}\left(1-\mathbb{E}_{\boldsymbol{s}_{(0)}} \operatorname{erf}_{\boldsymbol{\beta}_{(0)}}\left(\lambda, \boldsymbol{s}_{(0)}\right)\right) \\
& +\left(\sum_{i \in \boldsymbol{\tau} \backslash(0)} \boldsymbol{\beta}_{i}^{2}-\left(\left|\gamma_{(1)}\right|-\mu\right)^{2}-\eta\right) \underbrace{\left(1-\mathbb{E}_{\boldsymbol{s}_{(1)}} \operatorname{erf}_{\boldsymbol{\beta}_{(1)}}\left(\lambda, \boldsymbol{s}_{(1)}\right)\right)}_{<1}-0.38 \sum_{i \in J} \mathbb{E}_{\boldsymbol{s}_{i}} f_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)
\end{aligned}
$$

[^12]\[

$$
\begin{align*}
& \leq\left(\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}-\|\gamma\|_{2}^{2}+2 \mu\|\boldsymbol{\gamma}\|_{1}\right)-0.6 \lambda \\
& \leq \frac{2 c_{\mu}}{\sqrt{|\boldsymbol{\tau}|}}-0.6 \lambda \tag{B.135}
\end{align*}
$$
\]

On the other hand, when $\boldsymbol{\beta}_{(0)}^{2} \geq\left(\left|\gamma_{(0)}\right|-\mu\right)^{2}>0.38$, combining (B.133)-(B.134) gives:

$$
\begin{align*}
\text { (B.132) } \leq & \left(\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}-\|\boldsymbol{\gamma}\|_{2}^{2}+2 \mu\|\boldsymbol{\gamma}\|_{1}\right)+\left(\left(\left|\boldsymbol{\gamma}_{(0)}\right|-\mu\right)^{2}-\boldsymbol{\beta}_{(0)}^{2}\right) \mathbb{E}_{\boldsymbol{s}_{(0)}} \operatorname{erf}_{\boldsymbol{\beta}_{(0)}}\left(\lambda, \boldsymbol{s}_{(0)}\right) \\
& +\left(\left(\left|\boldsymbol{\gamma}_{(1)}\right|-\mu\right)^{2}-\sum_{i \in \boldsymbol{\tau} \backslash(0)} \beta_{i}^{2}\right) \mathbb{E}_{\boldsymbol{s}_{(1)}} \operatorname{erf}_{\boldsymbol{\beta}_{(1)}}\left(\lambda, \boldsymbol{s}_{(1)}\right)-0.38 \sum_{i \in J} \mathbb{E}_{\boldsymbol{s}_{i}} f_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right) \\
\leq & \left(\frac{c_{\mu}}{\sqrt{|\boldsymbol{\tau}|}}+4 \mu\right)+\left(\gamma_{(1)}^{2}-\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}+\boldsymbol{\beta}_{(0)}^{2}\right) \mathbb{E}_{\boldsymbol{s}_{(1)}} \operatorname{erf}_{\boldsymbol{\beta}_{(1)}}\left(\lambda, \boldsymbol{s}_{(1)}\right)-0.6 \lambda \tag{B.136}
\end{align*}
$$

where Theorem B.3.2 provides the upper bound for $\mathbb{E}_{\boldsymbol{s}_{(1)}} \operatorname{erf}_{\boldsymbol{\beta}_{(1)}}\left(\lambda, \boldsymbol{s}_{(1)}\right)$ as

$$
\begin{align*}
\mathbb{E}_{\boldsymbol{s}_{(1)}} \operatorname{erf}_{\boldsymbol{\beta}_{(1)}}\left(\lambda, \boldsymbol{s}_{(1)}\right) & =1-\frac{1}{n \theta \boldsymbol{\beta}_{(1)}} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{(1)} \leq 1-\frac{\sigma_{(1)}}{n \theta\left|\boldsymbol{\beta}_{(1)}\right|} \underline{\mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]}{ }_{(1)}=1-\frac{1}{\left|\boldsymbol{\beta}_{(1)}\right|}\left(\left|\boldsymbol{\beta}_{(1)}\right|-\sqrt{\frac{2}{\pi}} \lambda\right) \\
& \leq \sqrt{\frac{2}{\pi}} \cdot \frac{\lambda}{\left|\boldsymbol{\beta}_{(1)}\right|} \tag{B.137}
\end{align*}
$$

then calculate the constant for the second term in B.136) by writing $\kappa=\left|\frac{\gamma_{(1)}}{\boldsymbol{\gamma}_{(0)}}\right|=\left|\frac{\boldsymbol{\beta}_{(0)}}{\boldsymbol{\beta}_{(1)}}\right| \leq \frac{5}{4}$, which provides $\gamma_{(1)}^{2} \leq \frac{(1+\mu) \kappa^{2}}{\kappa^{2}+1}$ and $\boldsymbol{\beta}_{(0)}^{2} \leq \frac{\left\|\boldsymbol{\beta}_{7}\right\|_{2}^{2} \kappa^{2}}{\kappa^{2}+1}$ 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 $\frac{\left(\boldsymbol{\gamma}_{(1)}^{2}-1\right)+c_{\mu}+\boldsymbol{\beta}_{(0)}^{2}}{\left|\boldsymbol{\beta}_{(1)}\right|} \leq-\frac{\kappa}{\left(\kappa^{2}+1\right)\left|\boldsymbol{\beta}_{(0)}\right|}+\kappa\left|\boldsymbol{\beta}_{(0)}\right|+\frac{\mu+c_{\mu}}{0.3} \leq \frac{\kappa^{2}-1}{\sqrt{\kappa^{2}+1}}+\kappa\left(\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}-1\right)+4.2 c_{\mu} \leq 0.36+6 c_{\mu}$,
and finally combine $\overline{\text { B.137 }}-\sqrt{\text { B.138 }}$, follow from and use $c_{\lambda} \leq \frac{1}{3}$ :

$$
\begin{align*}
(\overline{\mathrm{B} .132} & \leq \frac{2 c_{\mu}}{\sqrt{|\boldsymbol{\tau}|}}+\sqrt{\frac{2}{\pi}}\left(\gamma_{(1)}^{2}-1+c_{\mu}+\boldsymbol{\beta}_{(0)}^{2}\right) \frac{\lambda}{\left|\boldsymbol{\beta}_{(1)}\right|}-0.6 \lambda \\
& \leq \frac{2 c_{\mu}}{\sqrt{|\boldsymbol{\tau}|}}+\sqrt{\frac{2}{\pi}}\left(0.36 \lambda+\frac{6 c_{\mu} c_{\lambda}}{0.3}\right)-0.6 \lambda \\
& \leq \frac{4 c_{\mu}}{\sqrt{|\boldsymbol{\tau}|}}-0.3 \lambda \tag{B.139}
\end{align*}
$$

3. (Collect all results) Combine the components of pseudo Hessian B.129, B.131) with bounds for $g_{1}+g_{2}$ from B.135 and B.139, and use Lemma B.2.5 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 $c_{\lambda} \geq \frac{1}{5}$, we can obtain:

$$
\boldsymbol{v}^{*} \widetilde{\operatorname{Hess}}_{\varphi_{\ell^{1}}}[\boldsymbol{a}] \boldsymbol{v} \leq g_{1}(\boldsymbol{\beta})+g_{2}(\boldsymbol{\beta})+\frac{7 c_{\mu} n \theta}{\sqrt{|\boldsymbol{\tau}|}}+(18 \mu p+8) n \theta^{2}
$$

$$
\begin{align*}
& \leq n \theta \cdot\left(\frac{4 c_{\mu}}{\sqrt{|\boldsymbol{\tau}|}}-0.3 \lambda\right)+n \theta \cdot \frac{7 c_{\mu}}{\sqrt{|\boldsymbol{\tau}|}}+n \theta \cdot \frac{6.5 c_{\mu}}{|\boldsymbol{\tau}|} \\
& \leq \frac{n \theta}{\sqrt{|\boldsymbol{\tau}|}}(0.059-0.06) \leq-0.001 n \theta \lambda \tag{B.140}
\end{align*}
$$

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

$$
\begin{equation*}
1-\underbrace{\mathbb{P}\left[\mathcal{E}_{\text {cross }}^{c}\right]}_{\text {Theorem B.1.4 }}-\underbrace{\mathbb{P}\left[\mathcal{E}_{\text {ncurv }}^{c}\right]}_{\text {Theorem B.4.3 }}-\underbrace{\mathbb{P}\left[\mathcal{E}_{\bar{x}}^{c}\right]}_{\text {Theorem B.3.4 }} . \tag{B.141}
\end{equation*}
$$

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

Corollary B.6.2 (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 $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $C, c, c^{\prime}, c^{\prime \prime}, \bar{c}>0$ such that if $\rho$ is $\delta$-smoothed $\ell^{1}$ function where $\delta \leq c^{\prime \prime} \lambda \theta^{8} / p^{2} \log ^{2} n$, $n>C p^{5} \theta^{-2} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-c^{\prime} / n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ satisfying $\left|\boldsymbol{\beta}_{(1)}\right| \geq \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|:$ for $\boldsymbol{v} \in \mathcal{S}_{\{(0),(1)\}} \cap \mathbb{S}^{p-1} \cap \boldsymbol{a}^{\perp}$,

$$
\begin{equation*}
\boldsymbol{v}^{*} \widetilde{\operatorname{Hess}}\left[\varphi_{\rho}\right](\boldsymbol{a}) \boldsymbol{v} \leq-\operatorname{cn} \theta \lambda \tag{B.142}
\end{equation*}
$$

Proof. Choose $\boldsymbol{v} \in \mathbb{S}^{p-1}$ according to Theorem B.6.1 and B.117 from Theorem B.5.6 with constant multiplier $\delta$ satisfies $c^{\prime \prime 1 / 4}<10^{-3} c$, we gain

$$
\begin{equation*}
\boldsymbol{v}^{*} \operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a}) \boldsymbol{v} \leq-c n \theta \lambda+200 c^{\prime} n \theta^{2} \leq-c n \theta \lambda / 2 \tag{B.143}
\end{equation*}
$$

## B.6.2 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 Theorem B.6.3, and the $\varphi_{\rho}$ version in Theorem B.6.4 when $\rho$ is properly defined as in Appendix B.5.

Lemma B.6.3 (Large gradient for $\varphi_{\ell^{1}}$ ). Suppose that $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ in $\mathbb{R}^{n}$, and $k, c_{\mu}$ such that $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $C, c^{\prime}, c, \bar{c}>0$, such that if $n>C p^{5} \theta^{-2} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-c^{\prime} / n$,
for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ satisfying $\frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|>\left|\boldsymbol{\beta}_{(1)}\right|>\frac{1}{4 \log \theta^{-1}} \lambda$,

$$
\begin{equation*}
\left\langle\boldsymbol{\sigma}_{(0)} \iota^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right],-\operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})\right\rangle \geq c n \theta\left(\log ^{-2} \theta^{-1}\right) \lambda^{2} \tag{B.144}
\end{equation*}
$$

where $\boldsymbol{\sigma}_{i}=\operatorname{sign}\left(\boldsymbol{\beta}_{i}\right)$.
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 $\iota^{*} s_{i}\left[\boldsymbol{a}_{0}\right]$ as a function of $\boldsymbol{\beta}_{i}$ as $d_{\mathbb{S}}\left(\boldsymbol{a}, \pm \boldsymbol{\iota}^{*} s_{i}\left[\boldsymbol{a}_{0}\right]\right)=$ $\cos ^{-1}\left(\boldsymbol{\beta}_{i}\right)$, then by triangle inequality we have:

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

Use the regional condition $\left|\boldsymbol{\beta}_{(1)}\right| \geq \frac{\theta_{10 g}}{4} \lambda$ and since $\mu|\boldsymbol{\tau}|^{3 / 2}<\frac{c_{\lambda}}{100} \theta_{\log }$ from Theorem B.2.1 implies

$$
\begin{equation*}
\left|\boldsymbol{\beta}_{(0)}\right| \leq 1-\frac{\boldsymbol{\beta}_{(1)}^{2}}{2}\left(1-\frac{4 \mu \sqrt{|\tau|}}{\theta_{\log } c_{\lambda}}\right) \leq 1-0.49 \boldsymbol{\beta}_{(1)}^{2}=: \beta_{\mathrm{ub}} \tag{B.145}
\end{equation*}
$$

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

$$
\begin{equation*}
\left|\boldsymbol{\beta}_{(0)}\right| \geq \frac{5}{4}\left|\boldsymbol{\beta}_{(1)}\right|=: \beta_{\mathrm{lb}} . \tag{B.146}
\end{equation*}
$$

Also since $\boldsymbol{\beta}=\boldsymbol{M} \boldsymbol{\alpha}$, based on properties of $\boldsymbol{M}$ from Theorem B.2.4 When $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2} \leq 1+c_{\mu}$ and $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq$ $\gamma \leq \frac{c_{\mu} \theta_{\text {log }}^{2}}{4 \mu \sqrt{p}|\tau|}$, we gain:

$$
\begin{align*}
& \boldsymbol{\beta}_{(0)}=\boldsymbol{\alpha}_{(0)}+\boldsymbol{e}_{(0)}^{*} \boldsymbol{M} \boldsymbol{\alpha}_{\backslash(0)} \\
\Longrightarrow & \left|\boldsymbol{\alpha}_{(0)}-\boldsymbol{\beta}_{(0)}\right| \leq \mu \sqrt{|\boldsymbol{\tau}|}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2}+\mu \sqrt{p}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \frac{c_{\mu} \theta_{\text {log }}^{2}\left(1+c_{\mu}\right)}{4|\boldsymbol{\tau}|}+\mu \sqrt{p} \gamma \leq \frac{c_{\mu} \theta_{\text {og }}^{2}}{|\boldsymbol{\tau}|} . \tag{B.147}
\end{align*}
$$

and therefore $\left|\boldsymbol{\alpha}_{(0)}\right| \leq\left|\boldsymbol{\beta}_{(0)}\right|+\frac{c_{\mu} \theta_{\text {og }}^{2}}{|\boldsymbol{\tau}|} \leq 1-.49\left(\frac{\theta_{\text {log }}}{4} \lambda\right)^{2}+\frac{c_{\mu} \theta_{\text {og }}^{2}}{|\boldsymbol{\tau}|}<1$.
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:

$$
h\left(\boldsymbol{\beta}_{i}\right):= \begin{cases}\boldsymbol{\beta}_{i}^{2}-\frac{\nu_{1} \lambda}{2}\left|\boldsymbol{\beta}_{i}\right| & \left|\boldsymbol{\beta}_{i}\right| \geq \nu_{1} \lambda \\ \frac{1}{2} \boldsymbol{\beta}_{i}^{2} & \left|\boldsymbol{\beta}_{i}\right| \leq \nu_{1} \lambda\end{cases}
$$

then Theorem B.10.7 tells us since $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash(0)}\right\|_{\infty} \leq \boldsymbol{\beta}_{(1)}$ :

$$
\begin{aligned}
\sum_{i \in \boldsymbol{\tau} \backslash(0)} h\left(\boldsymbol{\beta}_{i}\right) & \leq\left\|\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash(0)}\right\|_{2}^{2}\left(1-\frac{\nu_{1} \lambda \boldsymbol{\beta}_{(1)}}{2 \boldsymbol{\beta}_{(1)}^{2}}\right) \leq\left(1+\frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}-\boldsymbol{\beta}_{(0)}^{2}\right)\left(1-\frac{\nu_{1} \lambda}{2 \boldsymbol{\beta}_{(1)}}\right) \\
& \leq\left(1-\frac{\nu_{1} \lambda}{2 \boldsymbol{\beta}_{(1)}}\right)\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)+\frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|},
\end{aligned}
$$

then condition on the following event using Theorem B.3.4.

$$
\mathcal{E}_{\bar{\chi}}:=\left\{\boldsymbol{\beta}_{i} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \leq\left\{\begin{array}{ll}
n \theta \cdot h\left(\boldsymbol{\beta}_{i}\right)+\frac{c_{\mu} \theta}{p^{3 / 2}}\left|\boldsymbol{\beta}_{i}\right|, & \forall i \in \boldsymbol{\tau} \backslash(0) \\
n \theta \cdot 4 \boldsymbol{\beta}_{i}^{2} \theta|\boldsymbol{\tau}|+\frac{c_{\mu} \theta}{p^{3 / 2}}\left|\boldsymbol{\beta}_{i}\right|, & \forall i \in \boldsymbol{\tau}^{c}
\end{array}\right\}\right.
$$

which provides the upper bound of $\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]$ by applying $5 p>\log ^{8 / 3}\left(p \log ^{2} p\right)>\left(\theta_{\log }^{2}\right)^{4 / 3}$ from lower bound of $\theta$ from Theorem B.2.1, $\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \frac{c_{\mu} \theta_{\log }}{\sqrt{\theta}|\boldsymbol{\tau}|}$ from Theorem B.2.5, $|\boldsymbol{\tau}| \leq \sqrt{p}$ from lemma assumption and let $c_{\mu}<\frac{1}{100}$ :

$$
\begin{align*}
\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] & \leq \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)} \boldsymbol{\beta}_{(0)}+\sum_{i \in \boldsymbol{\tau} \backslash(0)} \boldsymbol{\beta}_{i} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}+\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}, \boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\rangle \\
& \leq \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)} \boldsymbol{\beta}_{(0)}+n\left(\theta \sum_{i \in \boldsymbol{\tau} \backslash(0)} h\left(\boldsymbol{\beta}_{i}\right)+4 \theta^{2}|\boldsymbol{\tau}|\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}+\frac{c_{\mu} \theta}{p^{3 / 2}}\left(\sqrt{|\boldsymbol{\tau}|}\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}+\sqrt{p}\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}\right)\right) \\
& \leq \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)} \boldsymbol{\beta}_{(0)}+n\left(\theta \cdot \eta\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)+\theta \cdot \frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}+\frac{4 \theta^{2}|\boldsymbol{\tau}| c_{\mu}^{2} \theta_{\log }^{2}}{\theta|\boldsymbol{\tau}|^{2}}+c_{\mu} \theta\left(\frac{1+c_{\mu}}{p^{3 / 4}|\boldsymbol{\tau}|}+\frac{c_{\mu} \theta_{\log }}{p \sqrt{\theta}|\boldsymbol{\tau}|}\right)\right) \\
& \leq \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)} \boldsymbol{\beta}_{(0)}+n \theta\left(\eta\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)+\frac{6 c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\right) \tag{B.148}
\end{align*}
$$

where $\eta=1-\frac{\nu_{1} \lambda}{2 \boldsymbol{\beta}_{(1)}}$.
3. (Align the gradient with $\boldsymbol{\iota}^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right]$ ) Base on the definition $\boldsymbol{\beta}$, since $\boldsymbol{\beta}_{(0)}=\left\langle\boldsymbol{a}, \iota^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right]\right\rangle$, we can expect that the negative gradient is likely aligned with direction toward one of the candidate solution $\pm \boldsymbol{\iota}^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right]$. Wlog assume that both $\boldsymbol{\beta}_{(0)}, \boldsymbol{\beta}_{(1)}$ are positive, then expand the gradient and use incoherent property for $\boldsymbol{a}_{0}$ Theorem B.2.4 we have:

$$
\begin{align*}
\left\langle\boldsymbol{\iota}^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right],-\operatorname{grad}_{\varphi_{\ell_{1}}}[\boldsymbol{a}]\right\rangle & =\left\langle\boldsymbol{\iota}^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right], \boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{\chi}[\boldsymbol{\beta}]-\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{\alpha}\right)\right\rangle \\
& \geq\left(\boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)}-\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{\alpha}_{(0)}\right)-\mu\left\|\boldsymbol{\chi}[\boldsymbol{\beta}] \backslash(0)-\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{\alpha}_{\backslash(0)}\right\|_{1} \tag{B.149}
\end{align*}
$$

where $\backslash(0)$ is an abbreviation of the complement set $\left[ \pm 2 p_{0}\right] \backslash(0)$. The latter part of $\mathrm{B.149}$ has an upper bound using bounds of $\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]<\frac{3 n \theta}{2},\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2}<\frac{n \theta \gamma_{2}}{20}$ from (B.184), and $\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau} \backslash(0)}\right\|_{2} \leq n \theta\left\|\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash(0)}\right\|_{2}$ in
event $\mathcal{E}_{\bar{\chi}}$, we obtain:

$$
\begin{align*}
& \mu\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\backslash(0)}-\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{\alpha}_{\backslash(0)}\right\|_{1} \\
& \leq \mu\left(\sqrt{|\boldsymbol{\tau}|}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau} \backslash(0)}\right\|_{2}+\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \sqrt{|\boldsymbol{\tau}|}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau} \backslash(0)}\right\|_{2}+\sqrt{p}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2}+\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \sqrt{p}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\right) \\
& \leq n \theta \cdot\left[\mu \sqrt{|\boldsymbol{\tau}|}\left(\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}-\left|\boldsymbol{\beta}_{(0)}\right|\right)+\mu \sqrt{|\boldsymbol{\tau}|}\left(\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}}\right\|_{2}-\left|\boldsymbol{\alpha}_{(0)}\right|\right)+\frac{1}{20} \mu \sqrt{p} \gamma_{2}+\frac{3}{2} \mu \sqrt{p} \gamma_{2}\right] \\
& \leq n \theta \cdot \frac{c_{\mu} \theta_{\log }^{2}}{4|\boldsymbol{\tau}|}\left[2\left(1+c_{\mu}\right)-\left|\boldsymbol{\beta}_{(0)}\right|-\left|\boldsymbol{\alpha}_{(0)}\right|+\left(\frac{1}{20}+\frac{3}{2}\right) c_{\mu}\right] \\
& \leq n \theta \cdot \frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\left(0.5+c_{\mu}-0.5 \boldsymbol{\beta}_{(0)}\right) . \tag{B.150}
\end{align*}
$$

On the other hand, the former term of (B.149) possesses a lower bound using $\bar{B} .147)-(\overline{B .148}), \chi[\boldsymbol{\beta}]_{(0)}>$ $n \theta\left(\boldsymbol{\beta}_{(0)}-\frac{\nu_{1}}{2} \lambda-\frac{c_{\mu}}{p}\right) \geq n \theta\left(\boldsymbol{\beta}_{(0)}-0.51 \nu_{1} \lambda\right)$ and $\boldsymbol{\alpha}_{(0)} \leq 1$ :

$$
\begin{align*}
& \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)}-\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{\alpha}_{(0)} \\
& \geq\left(1-\boldsymbol{\alpha}_{(0)} \boldsymbol{\beta}_{(0)}\right) \boldsymbol{\chi}[\boldsymbol{\beta}]_{(0)}-n \theta \cdot\left[\eta\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)+\frac{6 c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\right] \boldsymbol{\alpha}_{(0)} \\
& \geq n \theta \underbrace{\left(1-\left(\boldsymbol{\beta}_{(0)}+\frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\right) \boldsymbol{\beta}_{(0)}\right)\left(\boldsymbol{\beta}_{(0)}-0.51 \nu_{1} \lambda\right)}_{(a)}-n \theta \underbrace{\left[\eta\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)\left(\boldsymbol{\beta}_{(0)}+\frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\right)+\frac{6 c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|} \boldsymbol{\alpha}_{(0)}\right]}_{(b)} \\
& \geq n \theta[\underbrace{\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)\left(\boldsymbol{\beta}_{(0)}-0.51 \nu_{1} \lambda\right)-\frac{c_{\mu} \theta_{\log }^{2} \boldsymbol{\beta}_{(0)}^{2}}{|\boldsymbol{\tau}|}}_{(a)} \underbrace{-\left(1-\boldsymbol{\beta}_{(0)}^{2}\right) \eta \boldsymbol{\beta}_{(0)}-\eta \frac{c_{\mu} \theta_{\log }\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)}{|\boldsymbol{\tau}|}-\frac{6 c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}}_{(b)}] \\
& \geq n \theta\left[\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)\left((1-\eta) \boldsymbol{\beta}_{(0)}-0.51 \nu_{1} \lambda\right)-\frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\left((1-\eta) \boldsymbol{\beta}_{(0)}^{2}+7\right)\right], \tag{B.151}
\end{align*}
$$

combine B .149 with $\mathrm{B} .150-\mathrm{B} .151)$ and $\eta>0$, we have
$\overline{\text { B. } 149} \geq n \theta\left[\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)\left((1-\eta) \boldsymbol{\beta}_{(0)}-0.51 \nu_{1} \lambda\right)-\frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\left((1-\eta) \boldsymbol{\beta}_{(0)}^{2}+7\right)\right]-n \theta \cdot \frac{c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\left(0.5+c_{\mu}-0.5 \boldsymbol{\beta}_{(0)}\right)$

$$
\begin{equation*}
\geq n \theta[\underbrace{\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)\left(\frac{\nu_{1} \lambda}{2 \boldsymbol{\beta}_{(1)}} \boldsymbol{\beta}_{(0)}-0.51 \nu_{1} \lambda\right)}_{f(\beta)}-\frac{8 c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}] \tag{B.152}
\end{equation*}
$$

4. (Lower bound of $f(\beta)$ ) Given a fixed $\boldsymbol{\beta}_{(1)}$, the cubic function $f\left(\boldsymbol{\beta}_{(0)}\right)$ 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 $\boldsymbol{\beta}_{(0)} \in\left\{\beta_{\mathrm{lb}}, \beta_{\mathrm{ub}}\right\}$ from $\sqrt{\text { B.145 }}-\sqrt{\text { B.146 }}$,
we can observe that

$$
\boldsymbol{\beta}_{(0)} \in\left[\beta_{\mathrm{lb}}, \beta_{\mathrm{ub}}\right]=\left[\frac{5}{4} \boldsymbol{\beta}_{(1)}, 1-0.49 \boldsymbol{\beta}_{(1)}^{2}\right] \subseteq\left[1.02 \boldsymbol{\beta}_{(1)}, 1\right]
$$

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 $\nu_{1} \lambda<\frac{\sqrt{\theta_{\text {og }}}}{2 \sqrt{|\boldsymbol{\tau}|}} \leq \frac{1}{2 \sqrt{2}}$, since $|\boldsymbol{\tau}| \geq 2$, we have:

$$
\begin{equation*}
f\left(\beta_{\mathrm{lb}}\right) \geq\left(1-\beta_{\mathrm{lb}}^{2}\right)\left(\frac{\nu_{1} \lambda}{2 \boldsymbol{\beta}_{(1)}} \beta_{\mathrm{lb}}-0.51 \nu_{1} \lambda\right) \geq(1-0.616) \cdot\left(\frac{5}{8}-0.51\right) \nu_{1} \lambda \geq \frac{1}{16 \sqrt{2}} \nu_{1} \lambda \geq \frac{\theta_{\log }^{2}}{32} \lambda^{2} \tag{B.153}
\end{equation*}
$$

On the other hand when $\boldsymbol{\beta}_{(0)}=\beta_{\mathrm{ub}}$ :

$$
f\left(\beta_{\mathrm{ub}}\right) \geq\left(1-\beta_{\mathrm{ub}}^{2}\right)\left(\frac{\nu_{1} \lambda}{2 \boldsymbol{\beta}_{(1)}} \beta_{\mathrm{ub}}-0.51 \nu_{1} \lambda\right) \geq 0.49 \boldsymbol{\beta}_{(1)}^{2} \cdot\left(\frac{\nu_{1} \lambda}{2 \boldsymbol{\beta}_{(1)}}\left(1-0.49 \boldsymbol{\beta}_{(1)}^{2}\right)-0.51 \nu_{1} \lambda\right)
$$

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:

$$
\begin{equation*}
f\left(\beta_{\mathrm{ub}}\right) \geq 0.49\left(\frac{\lambda}{4 \log \theta^{-1}}\right)^{2} \cdot\left(\frac{1}{2}\left(1-0.49\left(\frac{\lambda}{4 \log \theta^{-1}}\right)^{2}\right)-0.51 \nu_{1} \lambda\right) \geq \frac{1}{6}\left(\frac{\lambda}{4 \log \theta^{-1}}\right)^{2} \geq \frac{\theta_{\log }^{2}}{96} \lambda^{2} \tag{B.154}
\end{equation*}
$$

Finally combine $\bar{B} .152$ with the lower bound of cubic function $\bar{B} .153-\bar{B} .154$ together with condition $c_{\mu}<\frac{c_{\lambda}^{2}}{800}$ and $\nu_{1}=\frac{\sqrt{\theta_{\text {log }}}}{2}$, obtain

$$
\begin{align*}
\left\langle\boldsymbol{\iota}^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right],-\operatorname{grad}_{\varphi_{\ell_{1}}}[\boldsymbol{a}]\right\rangle & \geq n \theta \cdot\left(\min \left\{f\left(\beta_{\mathrm{ub}}\right), f\left(\beta_{\mathrm{lb}}\right)\right\}-\frac{8 c_{\mu} \theta_{\log }^{2}}{|\boldsymbol{\tau}|}\right) \\
& \geq n \theta\left(\frac{\theta_{\log }^{2} c_{\lambda}^{2}}{96|\boldsymbol{\tau}|}-\frac{8 \theta_{\log }^{2} c_{\lambda}^{2}}{800|\boldsymbol{\tau}|}\right) \geq 6 \times 10^{-3} n \theta \theta_{\log }^{2} c_{\lambda}^{2} \tag{B.155}
\end{align*}
$$

The proof for the case where $\boldsymbol{\beta}_{(0)}$ negative can be derived in the same manner.

As a consequence, we have that

Corollary B.6.4 (Large gradient for $\varphi_{\rho}$ ). Suppose that $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ in $\mathbb{R}^{n}$, and $k, c_{\mu}$ such that $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $C, c, c^{\prime}, c^{\prime \prime}, \bar{c}>0$ such that if $\rho$ is $\delta$-smoothed $\ell^{1}$ function where $\delta \leq c^{\prime \prime} \lambda \theta^{8} / p^{2} \log ^{2} n$ with $n>C p^{5} \theta^{-2} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-c^{\prime} / n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ satisfying $\frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|>\left|\boldsymbol{\beta}_{(1)}\right|>\frac{1}{4 \log \theta^{-1} \lambda^{\prime}}$,

$$
\begin{equation*}
\left\langle\boldsymbol{\sigma}_{(0)} \iota^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right],-\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})\right\rangle \geq \operatorname{cn} \theta\left(\log ^{-2} \theta^{-1}\right) \lambda^{2} \tag{B.156}
\end{equation*}
$$

where $\boldsymbol{\sigma}_{i}=\operatorname{sign}\left(\boldsymbol{\beta}_{i}\right)$.

Proof. Choose $\iota^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right]$ as in Theorem B.6.3, and apply B.116 from Theorem B.5.6 with the constant multiplier of $\delta$ satisfies $c^{\prime \prime 4}<c / 4$, then utilize $\theta|\boldsymbol{\tau}| \log ^{2} \theta^{-1}<c_{\mu}$ from Theorem B.2.1 we have

$$
\begin{equation*}
\left\langle\boldsymbol{\sigma}_{(0)} \iota^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right],-\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})\right\rangle \geq \operatorname{cn} \theta\left(\log ^{-2} \theta^{-1}\right) \lambda-c^{\prime \prime} n \theta^{2} \geq c n \theta\left(\log ^{-2} \theta^{-1}\right) \lambda / 2 \tag{B.157}
\end{equation*}
$$

## B.6.3 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 Theorem B.6.5, and the $\varphi_{\rho}$ version in Theorem B.6.6 when $\rho$ is properly defined as in Appendix B. 5

Lemma B.6.5 (Strong convexity of $\varphi_{\ell^{1}}$ near shift). Suppose that $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ in $\mathbb{R}^{n}$, and $k, c_{\mu}$ such that $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $C, c, c^{\prime} \bar{c}>0$ such that if $n>C p^{5} \theta^{-2} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-c^{\prime} / n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ satisfying $\left|\boldsymbol{\beta}_{(1)}\right|<\frac{1}{4 \log \theta^{-1}} \lambda$ : for all $\boldsymbol{v} \in \mathbb{S}^{p-1} \cap \boldsymbol{v}^{\perp}$,

$$
\begin{equation*}
\boldsymbol{v}^{*} \widetilde{\operatorname{Hess}}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a}) \boldsymbol{v}>c n \theta \tag{B.158}
\end{equation*}
$$

furthermore, there exists $\overline{\boldsymbol{a}}$ as an local minimizer such that

$$
\begin{equation*}
\min _{\ell}\left\|\overline{\boldsymbol{a}}-s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\|_{2} \leq \frac{1}{2} \max \left\{\mu, p^{-1}\right\} \tag{B.159}
\end{equation*}
$$

Proof. 1. (Expectation of $\chi$ 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 $\boldsymbol{\beta}_{\backslash(0)}$ in an even larger shrinking ratio. Firstly we can show $\left|\left\langle\boldsymbol{\beta}_{\backslash(0)}, \boldsymbol{x}_{\backslash(0)}\right\rangle\right|$ is no larger then $\lambda / 2$ with probability at least $1-4 \theta$, since

$$
\begin{equation*}
\mathbb{P}\left[\left|\left\langle\boldsymbol{\beta}_{\backslash(0)}, \boldsymbol{x}_{\backslash(0)}\right\rangle\right|>\frac{\lambda}{2}\right] \leq \mathbb{P}\left[\left|\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau} \backslash(0)}, \boldsymbol{x}_{\boldsymbol{\tau} \backslash(0)}\right\rangle\right|>\frac{2 \lambda}{5}\right]+\mathbb{P}\left[\left|\left\langle\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}, \boldsymbol{x}_{\left.\boldsymbol{\tau}^{c}\right\rangle}\right\rangle\right|>\frac{\lambda}{10}\right] \leq 4 \theta \tag{B.160}
\end{equation*}
$$

via Theorem B.2.6 and Theorem B.2.7. Now recall from Theorem B.3.2 and the derivation of B.62-B.63, we know for every $i \neq(0)$,

$$
\begin{aligned}
\boldsymbol{\sigma}_{i} \mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} & =n \theta\left|\boldsymbol{\beta}_{i}\right| \mathbb{E}_{\boldsymbol{s}_{i}}\left[1-\operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, \boldsymbol{s}_{i}\right)\right] \\
& \leq n \theta\left|\boldsymbol{\beta}_{i}\right| \mathbb{E}_{g, \boldsymbol{x} \backslash i}\left[g^{2} \mathbf{1}_{\left\{\left|\boldsymbol{\beta}_{i} g+\boldsymbol{\beta}_{(0)} \boldsymbol{x}_{(0)}+\boldsymbol{\beta}_{\backslash\{(0), i\}}^{*} \boldsymbol{x}_{\backslash\{(0), i\}}\right|>\lambda\right\}}\right]
\end{aligned}
$$

$$
\begin{align*}
& \leq n \theta\left|\boldsymbol{\beta}_{i}\right|\left(\mathbb{E} g^{2} \mathbf{1}_{\left\{\left|\boldsymbol{\beta}_{i} g\right|>\lambda / 2\right\}}+\mathbb{P}\left[\boldsymbol{x}_{(0)} \neq 0\right]+\mathbb{P}\left[\left|\left\langle\boldsymbol{\beta}_{\backslash\{(0), i\}}, \boldsymbol{x}_{\backslash\{(0), i\}}\right\rangle\right|>\lambda / 2\right]\right) \\
& \leq n \theta\left|\boldsymbol{\beta}_{i}\right|\left(\left(\mathbb{E} g^{2}\right)^{1 / 2} \mathbb{P}\left[\left|\boldsymbol{\beta}_{(1)} g\right|>\lambda / 2\right]^{1 / 2}+\theta+4 \theta\right) \\
& \leq n \theta\left|\boldsymbol{\beta}_{i}\right|\left(\exp \left(-\log ^{2} \theta^{-1}\right)+5 \theta\right)  \tag{B.161}\\
& \leq 6 n \theta^{2}\left|\boldsymbol{\beta}_{i}\right|
\end{align*}
$$

where the third inequality is derived using union bound; the the fourth inequality is the result of (B.160), and the fifth inequality is derived from Gaussian tail bound theorem B.10.1
2. (Local strong convexity) Let $\boldsymbol{\gamma}=\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota}$, for any $\|\boldsymbol{v}\|_{2}=1$ we have $\|\boldsymbol{\gamma}\|_{2}^{2} \leq 1+\mu p$. Furthermore:

$$
\begin{align*}
\left|\gamma_{(0)}\right| & =\left|\left\langle\iota^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right], \boldsymbol{v}\right\rangle\right|=\left|\left\langle\boldsymbol{P}_{\boldsymbol{a}^{\perp}} \boldsymbol{\iota}^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right], \boldsymbol{v}\right\rangle\right|=\left|\left\langle\iota^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right]-\boldsymbol{\beta}_{(0)} \boldsymbol{a}, \boldsymbol{v}\right\rangle\right| \\
& \leq\left\|\boldsymbol{\iota}^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right]-\boldsymbol{\beta}_{(0)} \boldsymbol{a}\right\|_{2} \leq \sqrt{1-\boldsymbol{\beta}_{(0)}^{2}} . \tag{B.162}
\end{align*}
$$

Consider any such $\boldsymbol{v}$, the pseudo Hessian can be lower bounded as

$$
\begin{align*}
\boldsymbol{v}^{*} \widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a}) \boldsymbol{v} & =-\boldsymbol{\gamma}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\gamma} \\
& \geq-\boldsymbol{\gamma}_{(0)}^{2}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{e}_{(0)}\right\|_{2}^{2}-\sum_{i \neq(0)}\left\|\boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{e}_{i}\right\|_{2}^{2} \boldsymbol{\gamma}_{i}^{2}-2 \sum_{i \neq j}\left|\boldsymbol{e}_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{e}_{j}\right|\left|\boldsymbol{\gamma}_{i}\right|\left|\gamma_{j}\right| \\
& \geq-\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)\|\boldsymbol{x}\|_{2}^{2}-\max _{i \neq(0)}\left\|\boldsymbol{P}_{I(\boldsymbol{a})}^{s_{-i}}[\boldsymbol{x}]\right\|_{2}^{2}\|\boldsymbol{\gamma}\|_{2}^{2}-2 \max _{i \neq j}\left|e_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{e}_{j}\right|\|\boldsymbol{\gamma}\|_{1}^{2}, \tag{B.163}
\end{align*}
$$

where the second term is bounded by using its expectation derived in Theorem B.4.2 and utilize $\mathbb{P}\left[\left|\boldsymbol{s}_{i}\right|>\lambda / 2\right]<$ $4 \theta$ from (B.160), $\mathbb{E} \boldsymbol{\chi}$ from (B.161) and regional condition $\left|\boldsymbol{\beta}_{(1)}\right| \leq \frac{\lambda}{4 \log \theta^{-1}}$ to acquire

$$
\begin{align*}
\mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{a})}^{s_{-i}}[\boldsymbol{x}]\right\|_{2}^{2} & =n \theta\left[1-\mathbb{E}_{\boldsymbol{s}_{i}} \operatorname{erf}_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)+\mathbb{E}_{\boldsymbol{s}_{i}} f_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)\right] \\
& \leq \frac{\left|\mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}\right|}{\left|\boldsymbol{\beta}_{i}\right|}+n \theta \cdot\left(\max _{\left|s_{i}\right| \leq \frac{\lambda}{2}} f_{\boldsymbol{\beta}_{i}}\left(\lambda, s_{i}\right)+\mathbb{P}\left[\left|s_{i}\right|>\frac{\lambda}{2}\right]\right) \\
& \leq 6 n \theta^{2}+\frac{2 n \theta}{\sqrt{2 \pi}} \max _{\left|s_{i}\right| \leq \frac{\lambda}{2}}\left(\frac{\lambda+\left|s_{i}\right|}{\left|\boldsymbol{\beta}_{i}\right|} \cdot \exp \left[-\frac{\left(\lambda-\left|s_{i}\right|\right)^{2}}{2 \boldsymbol{\beta}_{i}^{2}}\right]\right)+4 n \theta^{2} \\
& \leq 10 n \theta^{2}+n \theta \cdot \log \theta^{-1} \exp \left(-2 \log ^{2} \theta^{-1}\right) \\
& \leq 11 n \theta^{2}, \tag{B.164}
\end{align*}
$$

and define the events $\mathcal{E}_{\|\boldsymbol{x}\|_{2}}, \mathcal{E}_{\text {cross }}$ and $\mathcal{E}_{\text {pcurv }}$ as follows:

$$
\left\{\begin{array}{l}
\mathcal{E}_{\text {pcurv }}:=\left\{\forall \boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right),\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}[\boldsymbol{x}]\right\|_{2}^{2} \leq 11 n \theta^{2}+\frac{c_{\mu} n \theta}{p}\right\}  \tag{B.165}\\
\mathcal{E}_{\text {cross }}:=\left\{\forall \boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right),\left|\boldsymbol{\beta}_{(1)}\right| \leq \frac{\lambda}{4 \log \theta^{-1}}, \max _{i \neq j \in[ \pm p]}\left|\boldsymbol{e}_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{P}_{I(\boldsymbol{a})} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{e}_{j}\right| \leq 8 n \theta^{3}\right\} \\
\mathcal{E}_{\|\boldsymbol{x}\|_{2}}:=\left\{\|\boldsymbol{x}\|_{2}^{2} \leq n \theta+3 \sqrt{n \theta} \log n\right\}
\end{array}\right.
$$

For the Hessian term, on the event $\mathcal{E}_{\text {pcurv }} \cap \mathcal{E}_{\text {cross }} \cap \mathcal{E}_{\|\boldsymbol{x}\|_{2}}$, 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 Theorem B.2.5. 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 (B.163) as

$$
\begin{align*}
\boldsymbol{v}^{*} \widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a}) \boldsymbol{v} & \geq-\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)(n \theta+3 \sqrt{n \theta} \log n)-(1+\mu p)\left(11 n \theta^{2}+\frac{c_{\mu} n \theta}{p}\right)-8 p(1+\mu p) \cdot 8 n \theta^{3} \\
& \geq-\frac{1}{2} n \theta \cdot\left(1-\boldsymbol{\beta}_{(0)}^{2}\right)-n \theta \cdot\left(\frac{11 c_{\mu}}{4}+c_{\mu}^{2}+\frac{64 c_{\mu}}{4}+\frac{64 c_{\mu}}{4}\right) \\
& \geq-\frac{1}{2} n \theta \cdot\left(1-\boldsymbol{\beta}_{(0)}^{2}+20 c_{\mu}\right) . \tag{B.166}
\end{align*}
$$

The bounds of $\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]$ can be derive on the event whose expectation is drawn from Theorem B.3.2 and B.161) as

$$
\mathcal{E}_{\chi}:=\left\{\begin{array}{ll}
\boldsymbol{\sigma}_{i} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \geq n \theta \mathcal{S}_{\nu_{2} \lambda}\left[\left|\boldsymbol{\beta}_{i}\right|\right]-\frac{c_{\mu} n \theta}{p}, & \forall i \in[ \pm p] \\
\boldsymbol{\sigma}_{i} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \leq 6 n \theta^{2}\left|\boldsymbol{\beta}_{i}\right|+\frac{c_{\mu} n \theta}{p^{3 / 2}}, & \forall i \neq(0)
\end{array}\right\}
$$

then use $\|\boldsymbol{\beta}\|_{1} \leq 1+\frac{\lambda p}{4 \log \theta^{-1}} \leq \frac{\lambda p}{2}$, implies:

$$
\begin{align*}
\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] & \geq n \theta\left|\boldsymbol{\beta}_{(0)}\right|\left(\left|\boldsymbol{\beta}_{(0)}\right|-\nu_{2} \lambda\right)-c_{\mu}\|\boldsymbol{\beta}\|_{1} \frac{n \theta}{p} \\
& \geq n \theta\left(\boldsymbol{\beta}_{(0)}^{2}-\sqrt{\frac{2}{\pi}} \lambda-\frac{c_{\mu}}{2} \lambda\right) \\
& \geq n \theta\left(\boldsymbol{\beta}_{(0)}^{2}-\lambda\right) . \tag{B.167}
\end{align*}
$$

Finally via the regional condition $\left|\boldsymbol{\beta}_{(1)}\right| \leq \frac{\lambda}{4 \log \theta^{-1}}$, the absolute value of leading correlation

$$
\begin{equation*}
\boldsymbol{\beta}_{(0)}^{2} \geq\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}-|\boldsymbol{\tau}| \boldsymbol{\beta}_{(1)}^{2} \geq 1-2 c_{\mu}-0.1^{2}>0.9 \tag{B.168}
\end{equation*}
$$

then we collect all above results and obtain:

$$
\begin{equation*}
\boldsymbol{v}^{*} \widetilde{\operatorname{Hess}}\left[\varphi_{\ell_{1}}\right](\boldsymbol{a}) \boldsymbol{v}=\boldsymbol{v}^{*} \widetilde{\nabla}^{2} \varphi_{\ell^{1}}(\boldsymbol{a}) \boldsymbol{v}-\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \geq\left(1.5 \boldsymbol{\beta}_{(0)}^{2}-0.5-\lambda-20 c_{\mu}\right) n \theta \geq 0.3 n \theta \tag{B.169}
\end{equation*}
$$

with probability at least

$$
\begin{equation*}
1-\underbrace{\mathbb{P}\left[\mathcal{E}_{\text {cross }}^{c}\right]}_{\text {Theorem B.4.4 }}-\underbrace{\mathbb{P}\left[\mathcal{E}_{\text {pcurv }}^{c}\right]}_{\text {Theorem B.4.3 }}-\underbrace{\mathbb{P}\left[\mathcal{E}_{\|x\|_{2}}^{c}\right]}_{\text {Theorem B.1.2 }}-\underbrace{\mathbb{P}\left[\mathcal{E}_{\chi}^{c}\right]}_{\text {Theorem B.3.4 }} \geq 1-c^{\prime} / n . \tag{B.170}
\end{equation*}
$$

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 (B.169), provides the upper bound on $\left\|\boldsymbol{a}_{*}-\boldsymbol{a}_{0}\right\|_{2}^{2}$ via

$$
\begin{align*}
& \varphi_{\ell^{1}}\left(\boldsymbol{a}_{*}\right) \geq \varphi_{\ell^{1}}\left(\boldsymbol{a}_{0}\right)+\left\langle\boldsymbol{a}_{*}-\boldsymbol{a}_{0}, \operatorname{grad}\left[\varphi_{\ell^{1}}\right]\left(\boldsymbol{a}_{0}\right)\right\rangle+\frac{0.3}{2} n \theta\left\|\boldsymbol{a}_{*}-\boldsymbol{a}_{0}\right\|_{2}^{2} \\
\Longrightarrow & \left\|\operatorname{grad}\left[\varphi_{\ell^{1}}\right]\left(\boldsymbol{a}_{0}\right)\right\|_{2} \geq 0.15 n \theta\left\|\boldsymbol{a}_{*}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.171}
\end{align*}
$$

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 $\left\|\boldsymbol{\beta}_{\backslash 0}\right\|_{\infty} \leq \mu$ hence $\left\|\boldsymbol{\beta}_{\backslash 0}\right\|_{\leq} \sqrt{2 p} \mu$. Expand the gradient term and condition on $\mathcal{E}_{\chi}$, since $\mu p^{2} \theta^{2} \leq \frac{c_{\mu}}{4}$ and $\theta<\frac{c_{\mu}}{4 \sqrt{p}}$, we can upper bound the gradient at $\boldsymbol{a}_{0}$ as

$$
\begin{align*}
\left\|\operatorname{grad}\left[\varphi_{\ell^{1}}\right]\left(\boldsymbol{a}_{0}\right)\right\|_{2} & =\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{\chi}[\boldsymbol{\beta}]-\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{e}_{0}\right)\right\|_{2} \leq\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right\|_{2}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\backslash 0}\right\|_{2} \\
& \leq \sqrt{1+\mu p}\left(6 n \theta^{2}\left\|\boldsymbol{\beta}_{\backslash 0}\right\|_{2}+n \theta \cdot \frac{c_{\mu}}{p^{3 / 2}} \cdot \sqrt{2 p}\right) \\
& \leq n \theta \sqrt{1+\mu p}\left(6 \mu \sqrt{2 p} \cdot \theta+\frac{2 c_{\mu}}{p}\right) \\
& \leq n \theta\left(3 c_{\mu} \mu+6 \mu \cdot \sqrt{2 \mu} \cdot p \theta+\frac{2 c_{\mu}}{p}+\frac{2 c_{\mu} \sqrt{\mu}}{\sqrt{p}}\right) \\
& \leq 7 \sqrt{c_{\mu}} n \theta \cdot \max \left\{\mu, \frac{1}{p}\right\} . \tag{B.172}
\end{align*}
$$

Thus we conclude that with sufficiently small $c_{\mu}$ :

$$
\begin{equation*}
\left\|\boldsymbol{a}_{*}-\boldsymbol{a}_{0}\right\|_{2} \leq 50 \sqrt{c_{\mu}} \max \left\{\mu, p^{-1}\right\} \leq \frac{1}{2} \max \left\{\mu, p^{-1}\right\} \tag{B.173}
\end{equation*}
$$

and we complete the proof by generalize this result from minima near $\boldsymbol{a}_{0}$ to any of its shifts $s_{i}\left[\boldsymbol{a}_{0}\right]$.

Similarly, for objective $\varphi_{\rho}$ we have

Corollary B.6.6 (Strong convexity of $\varphi_{\rho}$ of 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 $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $C, c, c^{\prime}, c^{\prime \prime}, \bar{c}>0$ such that if $\rho$ is $\delta$-smoothed $\ell^{1}$ function where $\delta \leq c^{\prime} \lambda \theta^{8} / p^{2} \log ^{2} n$ and $n>C p^{5} \theta^{-2} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-c^{\prime \prime} / n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ satisfying

$$
\left|\boldsymbol{\beta}_{(1)}\right|<\nu_{1} \lambda: \text { for all } \boldsymbol{v} \in \mathbb{S}^{p-1} \cap \boldsymbol{a}^{\perp}
$$

$$
\begin{equation*}
\boldsymbol{v}^{*} \widetilde{\operatorname{Hess}}\left[\varphi_{\rho}\right](\boldsymbol{a}) \boldsymbol{v}>c n \theta \tag{B.174}
\end{equation*}
$$

furthermore, there exists $\overline{\boldsymbol{a}}$ as an local minimizer such that

$$
\begin{equation*}
\min _{\ell}\left\|\overline{\boldsymbol{a}}-s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\|_{2} \leq \frac{1}{2} \max \left\{\mu, p^{-1}\right\} \tag{B.175}
\end{equation*}
$$

Proof. The strong convexity (B.174) is derived by combining B.158 and B.117 by letting constant multiplier of $\delta$ satisfies $c^{1 / 4}<10^{-3} c$. On the other hand the local minimizer near solution B.175 is derived via combining (B.171), B.115) and utilize both $\theta \sqrt{p}<c_{\mu}$ and $\mu p^{2} \theta^{2}<c_{\mu}$ such that:

$$
\begin{align*}
\left\|\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})\right\|_{2} & \leq\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right\|_{2}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]-\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}^{\delta}\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \iota \boldsymbol{a}\right]\right\|_{2}+\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right\|_{2}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\backslash 0}\right\|_{2} \\
& \leq \sqrt{1+\mu p} \cdot n \theta^{3}+7 \sqrt{c_{\mu}} n \theta \cdot \max \left\{\mu, p^{-1}\right\} \\
& \leq 8 n \theta \sqrt{c_{\mu}} \cdot \max \left\{\mu, p^{-1}\right\} \tag{B.176}
\end{align*}
$$

## B.6.4 Retraction toward subspace

As in Figure B.4, 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

$$
\begin{equation*}
\left\langle\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}(\boldsymbol{g}), \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}(\boldsymbol{a})\right\rangle>c\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\left\|\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}\right\|_{2} \tag{B.177}
\end{equation*}
$$

whenever $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \in\left[\frac{\gamma}{2}, \gamma\right]$. Apparently, the gradient will decrease $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)$, hence being addressed as retractive toward subspace $\mathcal{S}_{\tau}$. This retractive phenomenon is true for gradient of both $\varphi_{\ell^{1}}$ and $\varphi_{\rho}$.

Lemma B.6.7 (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 $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $n>C p^{5} \theta^{-2} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-c^{\prime} / n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ such that if

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \geq \gamma\left(c_{\mu}\right) / 2 \tag{B.178}
\end{equation*}
$$

then for every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{a}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}$, there exists some $\boldsymbol{\zeta}$ satisfying $\operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\zeta}$ that

$$
\begin{equation*}
\left\langle\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\rangle \geq \frac{1}{4 n \theta}\left\|\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \tag{B.179}
\end{equation*}
$$

Proof. Write $\gamma=\gamma\left(c_{\mu}\right)$ Recall the gradient can be derived as

$$
\begin{equation*}
\operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a})=-\boldsymbol{P}_{\boldsymbol{a} \perp} \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\chi}[\boldsymbol{\beta}]=\left(\boldsymbol{a} \boldsymbol{a}^{*}-\boldsymbol{I}\right) \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\chi}[\boldsymbol{\beta}]=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{\alpha}-\boldsymbol{\chi}[\boldsymbol{\beta}]\right), \tag{B.180}
\end{equation*}
$$

for every $\boldsymbol{\alpha}$ satisfies $\boldsymbol{a}=\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}$. Now via Theorem B.3.4 condition on the event:
and on this event, utilize Theorem B.2.5. bounds of $\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]$ and $\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2}$ can be derived with $c_{\mu}<\frac{1}{100}$ as:

$$
\begin{align*}
& \boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \leq n \theta\left(\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}+4 \theta|\boldsymbol{\tau}|\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}+c_{\mu}\right) \geq n \theta\left(1+c_{\mu}+4 c_{\mu}^{2}+c_{\mu}\right) \leq \frac{3}{2} n \theta  \tag{B.182}\\
& \boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \geq n \theta\left(\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{2}^{2}-\sqrt{2 / \pi} \lambda\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}}\right\|_{1}-c_{\mu}\right) \geq n \theta\left(1-4 c_{\mu}-\sqrt{2 / \pi} c_{\lambda}-c_{\mu}\right) \geq \frac{1}{2} n \theta  \tag{B.183}\\
& \left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq 4 n \theta^{2}|\boldsymbol{\tau}|\left\|\boldsymbol{\beta}_{\boldsymbol{\tau}^{c}}\right\|_{2}+\frac{c_{\mu} n \theta}{p} \sqrt{p} \leq n \theta\left(4 c_{\mu} \gamma+c_{\mu} \gamma\right) \leq \frac{1}{20} n \theta \gamma . \tag{B.184}
\end{align*}
$$

Let $\boldsymbol{\alpha}(\boldsymbol{g})=\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{\alpha}-\boldsymbol{\chi}[\boldsymbol{\beta}]$, derive

$$
\begin{align*}
& \left\langle\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\rangle-\frac{1}{4 n \theta}\left\|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \\
& \quad=\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-\left\langle\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}, \boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\rangle-\frac{1}{4 n \theta}\left\|\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}] \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}-\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \\
& \quad \geq \boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2}-\frac{1}{2 n \theta}\left|\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]\right|^{2}\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-\frac{1}{2 n \theta}\left\|\boldsymbol{\chi}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \\
& \quad \geq\left(\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]-\frac{1}{2 n \theta}\left(\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]\right)^{2}\right)\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-\frac{1}{20} n \theta \gamma\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}-\frac{1}{1000} n \theta \gamma^{2}, \tag{B.185}
\end{align*}
$$

notice that this is a quadratic function of $\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]$ with negative leading coefficient and zeros at $\{0,2 n \theta\}$, hence (B.185) is minimized when $\boldsymbol{\beta}^{*} \boldsymbol{\chi}[\boldsymbol{\beta}]=\frac{1}{2} n \theta$. Plugging in,

$$
\begin{equation*}
\text { B.185) } \geq \frac{3}{8} n \theta\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-\frac{1}{20} n \theta \gamma\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}-\frac{1}{1000} n \theta \gamma^{2} \tag{B.186}
\end{equation*}
$$

then again this is a quadratic function of $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}$ with positive leading coefficient and zeros at $\left\{0, \frac{8}{60} \gamma\right\}$, thus (B.186) is minimized at $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}=\frac{\gamma}{2}$. Plugging in again,

$$
\begin{equation*}
\text { B.186 } \geq \frac{3}{8} n \theta\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-\frac{1}{20} n \theta \gamma\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}-\frac{1}{1000} n \theta \gamma^{2} \geq\left(\frac{3}{32}-\frac{1}{80}-\frac{1}{1000}\right) n \theta \gamma^{2}>0 \tag{B.187}
\end{equation*}
$$

which concludes our proof.

As a consequence, we have that
Corollary B.6.8 (Retraction of $\varphi_{\rho}$ toward the subspace). Suppose that $\boldsymbol{x}_{0} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ in $\mathbb{R}^{n}$, and $k, c_{\mu}$ such that $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$. 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 $C, c, c^{\prime}, c^{\prime \prime}, \bar{c}>0$ such that if $\rho$ is $\delta$-smoothed $\ell^{1}$ function where $\delta \leq c^{\prime \prime} \lambda \theta^{8} / p^{2} \log ^{2} n$ and $n>C p^{5} \theta^{-2} \log p$ and $c_{\mu} \leq \bar{c}$, then with probability at least $1-c^{\prime} / n$, for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ such that if

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \geq \gamma\left(c_{\mu}\right) / 2 \tag{B.188}
\end{equation*}
$$

then for every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{a}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}$, there exists some $\boldsymbol{\zeta}$ satisfying $\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\zeta}$ that

$$
\begin{equation*}
\left\langle\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\rangle \geq \frac{1}{6 n \theta}\left\|\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \tag{B.189}
\end{equation*}
$$

Proof. Write $\gamma=\gamma\left(c_{\mu}\right)$. Define

$$
\chi_{\ell^{1}}[\boldsymbol{\beta}]=\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}], \quad \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]=\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \mathcal{S}_{\lambda}^{\delta}[\breve{\boldsymbol{a}} * \boldsymbol{y}],
$$

which, and on event (B.181) and Theorem B.5.6, we know

$$
\begin{align*}
\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}] & \leq \frac{3}{2} n \theta,  \tag{B.190}\\
\left\|\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2} & \leq \frac{1}{20} n \theta \gamma,  \tag{B.191}\\
\left\|\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]-\boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]\right\|_{2} & \leq c_{1} n \theta^{4}, \tag{B.192}
\end{align*}
$$

for some constant $c_{1}>0$. Now given any $\boldsymbol{\alpha}$ satisfies $\boldsymbol{a}=\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}$, the gradient of both objective can be derived as

$$
\begin{align*}
\operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a}) & =-\boldsymbol{P}_{\boldsymbol{a}^{ \pm}} \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \operatorname{prox}_{\lambda\|\cdot\|_{1}}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}]=\left(\boldsymbol{a} \boldsymbol{a}^{*}-\boldsymbol{I}\right) \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}] \\
& =\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}] \boldsymbol{\alpha}-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right),  \tag{B.193}\\
\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a}) & =-\boldsymbol{P}_{\boldsymbol{a}^{\perp}} \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \operatorname{prox}_{\lambda \rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]=\left(\boldsymbol{a} \boldsymbol{a}^{*}-\boldsymbol{I}\right) \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}] \\
& =\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}] \boldsymbol{\alpha}-\boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]\right) . \tag{B.194}
\end{align*}
$$

In the same spirit, define the coefficient of each gradient vector

$$
\begin{align*}
\zeta_{\ell^{1}} & =\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}] \boldsymbol{\alpha}-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}],  \tag{B.195}\\
\boldsymbol{\zeta}_{\rho} & =\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}] \boldsymbol{\alpha}-\boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}], \tag{B.196}
\end{align*}
$$

which, by norm inequality from (B.190)-(B.192) and Theorem B.6.7, we can derive

$$
\begin{align*}
&\left\|\boldsymbol{\zeta}_{\ell^{1}}-\boldsymbol{\zeta}_{\rho}\right\|_{2} \leq\left\|\left(\boldsymbol{I}-\boldsymbol{\alpha} \boldsymbol{\beta}^{*}\right)\left(\boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right)\right\|_{2} \leq c_{1} n \theta^{4},  \tag{B.197}\\
&\left\|\left(\boldsymbol{\zeta}_{\ell^{1}}\right)_{\boldsymbol{\tau}^{c}}\right\|_{2} \geq\left|\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right|\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}-\left\|\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2} \geq \frac{1}{5} n \theta \gamma,  \tag{B.198}\\
&\left\langle\left(\boldsymbol{\zeta}_{\ell^{1}}\right)_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\rangle \geq \frac{1}{4 n \theta}\left\|\left(\boldsymbol{\zeta}_{\ell^{1}}\right)_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}, \tag{B.199}
\end{align*}
$$

where the first inequality is derived by observing $\left(\boldsymbol{I}-\boldsymbol{\alpha} \boldsymbol{\beta}^{*}\right)$ is a projection operator, as such:

$$
\begin{aligned}
\boldsymbol{\beta}^{*} \boldsymbol{\alpha} & =\boldsymbol{a}^{*} \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}=\boldsymbol{a}^{*} \boldsymbol{a}=1, \\
\left(\boldsymbol{I}-\boldsymbol{\alpha} \boldsymbol{\beta}^{*}\right)^{2} & =\boldsymbol{I}-2 \boldsymbol{\alpha} \boldsymbol{\beta}^{*}+\boldsymbol{\alpha}\left(\boldsymbol{\beta}^{*} \boldsymbol{\alpha}\right) \boldsymbol{\beta}^{*}=\boldsymbol{I}-\boldsymbol{\alpha} \boldsymbol{\beta}^{*} .
\end{aligned}
$$

Now we are ready to derive B.189):

$$
\begin{align*}
\left\langle\left(\boldsymbol{\zeta}_{\rho}\right)_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\rangle \geq & \left\langle\left(\boldsymbol{\zeta}_{\ell^{1}}\right)_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\rangle-\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}\left\|\boldsymbol{\zeta}_{\rho}-\boldsymbol{\zeta}_{\ell^{1}}\right\|_{2} \\
\geq & \frac{1}{4 n \theta}\left\|\left(\boldsymbol{\zeta}_{\ell^{1}}\right)_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-c_{1} n \theta^{4} \gamma \\
\geq & \frac{1}{12 n \theta}\left\|\left(\boldsymbol{\zeta}_{\ell^{1}}\right)_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \\
& \quad \frac{1}{6 n \theta}\left(\left\|\left(\boldsymbol{\zeta}_{\rho}\right)_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-2\left\|\left(\boldsymbol{\zeta}_{\ell^{1}}\right)_{\boldsymbol{\tau}^{c}}\right\|_{2}\left\|\boldsymbol{\zeta}_{\ell^{1}}-\boldsymbol{\zeta}_{\rho}\right\|_{2}-\left\|\boldsymbol{\zeta}_{\ell^{1}}-\boldsymbol{\zeta}_{\rho}\right\|_{2}^{2}\right)-c_{1} n \theta^{4} \gamma \\
\geq & \frac{1}{6 n \theta}\left\|\left(\boldsymbol{\zeta}_{\rho}\right)_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}+\frac{1}{12 n \theta}\left(\frac{1}{5} n \theta \gamma\right)^{2}-\frac{1}{3 n \theta}\left(\frac{1}{5} n \theta \gamma\right)\left(c_{1} n \theta^{4}\right)-\frac{1}{6 n \theta}\left(c_{1} n \theta^{4}\right)^{2}-c_{1} n \theta^{4} \gamma \\
\geq & \frac{1}{6 n \theta}\left\|\left(\boldsymbol{\zeta}_{\rho}\right)_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} . \tag{B.200}
\end{align*}
$$

where the last inequality is true since $\theta^{3} \ll \gamma$.

## B.6.5 Proof of Theorem 3.4.1

By collecting result from above, we are ready to prove the acclaimed geometric result in Theorem 3.4.1 It guarantees that for every $\boldsymbol{a}$ near $\mathcal{S}_{\tau}$, either one of the following in true

$$
\begin{align*}
\lambda_{\min }\left(\operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a})\right) & \leq-c_{1} n \theta \lambda,  \tag{B.201}\\
\left\langle\boldsymbol{\sigma}_{(0)} \iota^{*} s_{(0)}\left[\boldsymbol{a}_{0}\right],-\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})\right\rangle & \geq c_{2} n \theta\left(\log ^{-2} \theta^{-1}\right) \lambda^{2},  \tag{B.202}\\
\operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a}) & \succ c_{3} n \theta \cdot \boldsymbol{P}_{\boldsymbol{a}^{\perp}}, \tag{B.203}
\end{align*}
$$

all local minimizer $\overline{\boldsymbol{a}}$ satisfies for some $\boldsymbol{a}_{*} \in\left\{ \pm \iota^{*} s_{\ell}[\boldsymbol{a}] \mid \ell \in\left[ \pm p_{0}\right]\right\}$,

$$
\begin{equation*}
\left\|\overline{\boldsymbol{a}}-\boldsymbol{a}_{*}\right\|_{2} \leq c_{4} \sqrt{c_{\mu}} \max \left\{\mu, p_{0}^{-1}\right\} \tag{B.204}
\end{equation*}
$$

and whenever $\frac{\gamma}{2} \leq d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \gamma$, coefficient of $\boldsymbol{a}$ and its gradient $\boldsymbol{g}, \boldsymbol{\alpha}$, written as $\boldsymbol{\zeta}$, satisfies

$$
\begin{equation*}
\left\langle\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\rangle \geq \frac{c_{5}}{n \theta}\left\|\boldsymbol{\zeta}_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} . \tag{B.205}
\end{equation*}
$$

To connect the geometric results introduced in Theorem B.6.1, Theorem B.6.3, Theorem B.6.5 and Theorem B.6.7, we are only required to prove the required signal condition claimed in Theorem 3.4.1 is necessary from Theorem B.2.1. In particular, when the subspace dimension $|\boldsymbol{\tau}| \leq 4 p_{0} \theta$. 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 Theorem B.6.2, Theorem B.6.4. Theorem B.6.6 and Theorem B.6.8 also holds.

Proof. Firstly we will show when largest solution subspace dimension $k=4 p_{0} \theta$, the signal condition of Theorem B.2.1 will be satisfied. Recall that the signal condition of Theorem 3.4.1 requests

$$
\begin{equation*}
\frac{2}{p_{0} \log ^{2} p_{0}} \leq \theta \leq \frac{c}{\left(p_{0} \sqrt{\mu}+\sqrt{p_{0}}\right) \log ^{2} p_{0}}, \tag{B.206}
\end{equation*}
$$

since $p=3 p_{0}-2$, this implies the lower bounds for sparsity $\theta$ as:

$$
\begin{equation*}
\theta \geq \frac{1}{2 p_{0}\left(\frac{1}{2} \log p_{0}\right)^{2}} \geq \frac{1}{p \log ^{2} \theta^{-1}} ; \tag{B.207}
\end{equation*}
$$

the upper bound of $\theta$ via $\theta \sqrt{p_{0}} \log ^{2} p_{0} \leq c$ :

$$
\begin{equation*}
\theta \leq \frac{9 c}{\sqrt{p_{0}}\left(3 \log p_{0}\right)^{2}} \leq \frac{16 c}{\sqrt{p} \log ^{2} \theta^{-1}}, \quad \theta \leq \frac{4 c^{2}}{k \log ^{4} p_{0}} \leq \frac{36 c^{2}}{k\left(3 \log p_{0}\right)^{2}} \leq \frac{36 c^{2}}{k \log ^{2} \theta^{-1}} ; \tag{B.208}
\end{equation*}
$$

and the upper bound for coherence $\mu$ as:

$$
\begin{equation*}
\mu \max \left\{k^{2},(p \theta)^{2}\right\} \log ^{2} \theta^{-1} \leq \mu \max \left\{16\left(p_{0} \theta\right)^{2}, 9\left(p_{0} \theta\right)^{2}\right\} \log ^{2} \theta^{-1} \leq 16\left(\sqrt{\mu} p_{0} \theta\right)^{2} \log ^{2} p_{0} \leq 16 c . \tag{B.209}
\end{equation*}
$$

Therefore Theorem B.2.1 holds if $\max \left\{16 c, 36 c^{2}\right\} \leq c_{\mu} / 4$ via B.207-B.209).
Furthermore, we know from lemma assumption all interested $a$ are near subspace $\mathcal{S}_{\tau}$ by

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{a} \cdot \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \frac{c}{\sqrt{p_{0}} \log ^{2} \theta^{-1}} \cdot \min \left\{\frac{1}{\sqrt{\theta}}, \frac{1}{\sqrt{\mu}} \cdot \frac{1}{\mu\left(p_{0} \theta\right)^{3 / 2}}\right\} \leq \frac{c}{\log ^{2} \theta^{-1}} \min \left\{\frac{2}{\sqrt{k}}, \frac{1}{\sqrt{p_{0} \mu}}, \frac{4}{\mu p_{0} \sqrt{\theta} k}\right\} \leq \gamma \tag{B.210}
\end{equation*}
$$

where $\gamma$ is defined in Theorem B.2.3 of widened subspace $\mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$.
Lastly, the pseudo-Huber function $\rho(x)=\sqrt{x^{2}+\delta^{2}}$ is an $\ell^{1}$ smoothed sparse surrogate defined in Theorem B.5.2, by observing that it is convex, smooth, even, whose second order derivative (according to

Table B.1 $\nabla^{2} \rho(x)=\frac{\delta^{2}}{\left(x^{2}+\delta^{2}\right)^{3 / 2}}$ is monotone decreasing in $|x|$. More importantly

$$
\begin{equation*}
\sup _{x \in \mathbb{R}}|\rho(x)-|x||=|\rho(0)-|0||=\delta . \tag{B.211}
\end{equation*}
$$

Hence, by choosing $\delta \leq \frac{c^{\prime 4} \theta^{8}}{p^{2} \log ^{2} n} \lambda$, for some sufficiently small constant $c^{\prime}$ and letting $\lambda=0.2 \sqrt{k}=$ $0.1 / \sqrt{p_{0} \theta}$ in $\varphi_{\rho}$. We obtain the geometrical results in Theorem B.6.2 when $\left|\boldsymbol{\beta}_{(1)}\right| \geq \frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right|$, Theorem B.6.4 when $\frac{4}{5}\left|\boldsymbol{\beta}_{(0)}\right| \geq\left|\boldsymbol{\beta}_{(1)}\right| \geq \frac{\lambda}{4 \log ^{2} \theta^{-1}}$ and Theorem B.6.6 when $\frac{\lambda}{4 \log ^{2} \theta^{-1}} \geq\left|\boldsymbol{\beta}_{(1)}\right|$, and the retraction result in Theorem B.6.8.

## B.7 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 $\mathcal{S}_{\tau}$. We begin by proving the initialization method guarantees that $\boldsymbol{a}^{(0)}$ is near $\mathcal{S}_{\tau}$, in the sense that

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{a}^{(0)}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \gamma, \tag{B.212}
\end{equation*}
$$

where the distance $d_{\alpha}$ is defined in 3.50 . 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 $\mathcal{S}_{\tau}$ and (ii) the geometric properties of $\varphi_{\rho}$ near $\mathcal{S}_{\tau}$.

## B.7.1 Initialization near subspace

The following lemma shows that the initialization $\boldsymbol{a}^{(0)}=\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\nabla \varphi_{\ell^{1}}\left(\boldsymbol{a}^{(-1)}\right)\right]$, where

$$
\begin{equation*}
\boldsymbol{a}^{(-1)}=\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\sum_{\ell \in \mathcal{T}} x_{0 \ell} \iota_{p_{0}}^{*} s_{\ell}\left[\boldsymbol{a}_{0}\right]\right], \tag{B.213}
\end{equation*}
$$

and is very close to the subspace $\mathcal{S}_{\tau}$ :
Lemma B.7.1 (Initialization from a piece of data). Let $\overline{\boldsymbol{x}} \in \mathbb{R}^{2 p_{0}-1}$ indexed by $\left[ \pm p_{0}\right]$, with $\overline{\boldsymbol{x}}_{i} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$. Define $\overline{\boldsymbol{y}}=\overline{\boldsymbol{x}} * \boldsymbol{a}_{0}$, and $\boldsymbol{a}^{(0)}$ as

$$
\begin{equation*}
\boldsymbol{a}^{(0)}=-\boldsymbol{P}_{\mathbb{S}^{p}-1} \nabla \varphi_{\ell^{1}}\left(\boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\mathbf{0}^{p_{0}-1} ;\left[\overline{\boldsymbol{y}}_{0} ; \cdots ; \overline{\boldsymbol{y}}_{p_{0}-1}\right] ; \boldsymbol{0}^{p_{0}-1}\right]\right), \tag{B.214}
\end{equation*}
$$

with $\lambda=0.2 / \sqrt{p \theta}$ in $\varphi_{1}$. Set $\boldsymbol{\tau}=\operatorname{supp}(\overline{\boldsymbol{x}})$. Suppose that $\left(\boldsymbol{a}_{0}, \theta, k\right)$ satisfies the sparsity-coherence condition $\operatorname{SCC}\left(c_{\mu}\right)$ and $\boldsymbol{a}_{0}$ satisfies $\max _{i \neq j}\left|\left\langle\iota_{p_{0}}^{*} s_{i}\left[\boldsymbol{a}_{0}\right], \iota_{p_{0}}^{*} s_{j}\left[\boldsymbol{a}_{0}\right]\right\rangle\right| \leq \mu$. Then there exists some constant $c, \bar{c}>0$ such that if $p_{0} \theta>1000 c$
and $c_{\mu} \leq \bar{c}$, then with probability at least $1-1 / c$, we have

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{a}^{(0)}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \frac{c_{\mu}}{4 \log ^{2} \theta^{-1}} \min \left\{\frac{1}{\sqrt{|\boldsymbol{\tau}|}}, \frac{1}{\sqrt{\mu p}}, \frac{1}{\mu p \sqrt{\theta}|\boldsymbol{\tau}|}\right\} \tag{B.215}
\end{equation*}
$$

Proof. 1. (Distance to $\mathcal{S}_{\boldsymbol{\tau}}$ from $\left.\boldsymbol{a}^{(0)}\right)$ Let $\eta=\left\|\boldsymbol{\iota}_{p_{0}}^{*}\left(\boldsymbol{a}_{0} * \boldsymbol{x}\right)\right\|_{2}=\left\|\boldsymbol{\iota}_{p_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}\right\|_{2}$ and $\gamma=\gamma\left(c_{\mu}\right)$, as in B.215). Expand the expression of $\boldsymbol{a}^{(0)}$ from B.214 we have

$$
\begin{equation*}
\boldsymbol{a}^{(0)}=\boldsymbol{P}_{\mathbb{S}^{p}-1} \iota^{*} \breve{\boldsymbol{C}}_{\boldsymbol{y}} \mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{y}} \boldsymbol{\iota}_{p_{0}} \boldsymbol{P}_{\mathbb{S}^{p_{0}-1}} \iota_{p_{0}}^{*}\left(\boldsymbol{a}_{0} * \boldsymbol{x}\right)\right]=\boldsymbol{P}_{\mathbb{S}^{p}-1} \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\chi}\left[\frac{1}{\eta} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \iota_{p_{0}} \iota_{p_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}\right] \tag{B.216}
\end{equation*}
$$

To relate $\boldsymbol{a}^{(0)}$ to its coefficient, introduce the truncated autocorrelation matrix $\widetilde{\boldsymbol{M}}=\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota}_{p_{0}} \iota_{p_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}$, define $\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}}$ as

$$
\begin{equation*}
\widetilde{\boldsymbol{\beta}}=\frac{1}{\eta} \widetilde{\boldsymbol{M}} \boldsymbol{x}, \quad \widetilde{\boldsymbol{\alpha}}=\chi\left[\frac{1}{\eta} \widetilde{\boldsymbol{M}} \boldsymbol{x}\right]=\boldsymbol{\chi}[\widetilde{\boldsymbol{\beta}}] \tag{B.217}
\end{equation*}
$$

and note that $\widetilde{M}$ is bounded entrywise as

$$
\left|\widetilde{\boldsymbol{M}}_{i j}\right| \leq \begin{cases}1 & i=j \in\left[-p_{0}+1, p_{0}-1\right]  \tag{B.218}\\ \mu & i \neq j \in\left[-p_{0}+1, p_{0}-1\right],|i-j|<p_{0} \\ 0 & \text { otherwise }\end{cases}
$$

 coefficient vector for $\boldsymbol{a}^{(0)}$. Let $\boldsymbol{\tau}^{c}=\left[ \pm 2 p_{0}\right] \backslash \boldsymbol{\tau}$. The distance $d_{\boldsymbol{\alpha}}$ to subspace $\mathcal{S}_{\boldsymbol{\tau}}$ (3.50) is upper bounded as

$$
\begin{aligned}
d_{\alpha}\left(\boldsymbol{a}^{(0)}, \mathcal{S}_{\boldsymbol{\tau}}\right) & \leq \frac{\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2}}{\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \widetilde{\boldsymbol{\alpha}}\right\|_{2}} \leq \frac{\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2}}{\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\right\|_{2}-\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2}} \\
& \leq \frac{\| \widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c} \|_{2}}}{\sqrt{1-\mu|\boldsymbol{\tau}|}\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\right\|_{2}-\sqrt{1+\mu p}\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2}}
\end{aligned}
$$

where the last inequality is derived with Theorem B.2.4. Therefore, it is sufficient to show

$$
\begin{equation*}
(1+\gamma \sqrt{1+\mu p})\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \gamma \sqrt{1-\mu|\boldsymbol{\tau}|}\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\right\|_{2} \tag{B.219}
\end{equation*}
$$

to complete the proof that $d_{\alpha}\left(\boldsymbol{a}^{(0)}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \gamma$.
2. (Bound $\eta$ ) Condition on the following two events

$$
\begin{equation*}
\mathcal{E}_{\tau}:=\left\{|\boldsymbol{\tau}|<4 p_{0} \theta\right\}, \quad \mathcal{E}_{\|\boldsymbol{x}\|_{2}}:=\left\{\sqrt{p_{0} \theta} \leq\|\boldsymbol{x}\|_{2} \leq \sqrt{3 p_{0} \theta}\right\} \tag{B.220}
\end{equation*}
$$

and utilize $\mu$ bound from Theorem B.2.5 such that $\mu|\boldsymbol{\tau}|<0.1$. An upper bound on $\eta$ can be obtained using
properties of $\widetilde{M}$ of (B.218):

$$
\begin{equation*}
\eta=\left\|\iota_{p_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}\right\|_{2} \leq\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}\right\|_{2} \leq \sqrt{1+\mu|\boldsymbol{\tau}|}\|\boldsymbol{x}\|_{2} \leq 2 \sqrt{p_{0} \theta} \tag{B.221}
\end{equation*}
$$

To lower bound $\eta$, use $\eta^{2}=\boldsymbol{g}^{*} \boldsymbol{P}_{\tau} \widetilde{\boldsymbol{M}} \boldsymbol{P}_{\tau} \boldsymbol{g}$ where $\boldsymbol{g}$ is the standard Gaussian vector. Observe the submatrix of $\widetilde{M}$ is diagonal dominant:

$$
\left\{\begin{array}{l}
\widetilde{\boldsymbol{M}}_{i i}=\left\|\iota_{p_{0}}^{*} s_{i}\left[\boldsymbol{a}_{0}\right]\right\|_{2}^{2} \in[0,1]  \tag{B.222}\\
\operatorname{tr}(\widetilde{\boldsymbol{M}})=\sum_{i \in\left[ \pm p_{0}\right]}\left\|\iota_{p_{0}}^{*} s_{i}\left[\boldsymbol{a}_{0}\right]\right\|_{2}^{2}=\left\|\boldsymbol{a}_{0}\right\|_{2}^{2}+\sum_{i=1}^{p_{0}-1}\left(\left\|\iota_{p_{0}}^{*} s_{i}\left[\boldsymbol{a}_{0}\right]\right\|_{2}^{2}+\left\|\boldsymbol{\iota}_{p_{0}}^{*} s_{i-p_{0}}\left[\boldsymbol{a}_{0}\right]\right\|_{2}^{2}\right)=p_{0}
\end{array} .\right.
$$

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 $\boldsymbol{P}_{\tau} \widetilde{\boldsymbol{M}} \boldsymbol{P}_{\tau}$ can be written as sum of independent r.v.s as:

$$
\operatorname{tr}\left(\boldsymbol{P}_{\tau} \widetilde{\boldsymbol{M}} \boldsymbol{P}_{\tau}\right)=\sum_{i \in\left[ \pm p_{0}\right]} w_{i}\left\|\iota_{p_{0}}^{*} s_{i}\left[\boldsymbol{a}_{0}\right]\right\|_{2}^{2},
$$

Bernstein inequality Theorem B.10.4 and (B.222) gives

$$
\begin{equation*}
\mathbb{P}\left[\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}} \widetilde{\boldsymbol{M}} \boldsymbol{P}_{\boldsymbol{\tau}}\right)<\frac{3 p_{0} \theta}{4}\right] \leq \mathbb{P}\left[\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}} \widetilde{\boldsymbol{M}} \boldsymbol{P}_{\boldsymbol{\tau}}\right)-p_{0} \theta \leq-\frac{p_{0} \theta}{4}\right] \leq 2 \exp \left(\frac{-\left(p_{0} \theta / 4\right)^{2}}{2 p_{0} \theta+p_{0} \theta / 2}\right) \leq 2 \exp \left(\frac{-p_{0} \theta}{40}\right) \tag{B.223}
\end{equation*}
$$

thus condition on $\boldsymbol{\omega}$ satisfies $\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}} \widetilde{M} \boldsymbol{P}_{\tau}\right) \geq 3 p_{0} \theta / 4$ and $\mathcal{E}_{\boldsymbol{\tau}}$, expectation $\eta^{2}$ has lower bound

$$
\mathbb{E}_{\boldsymbol{g} \mid \boldsymbol{w}} \eta^{2}=\mathbb{E}_{\boldsymbol{g} \mid \boldsymbol{w}}\left[\boldsymbol{g}^{*} \boldsymbol{P}_{\boldsymbol{\tau}} \widetilde{\boldsymbol{M}} \boldsymbol{P}_{\boldsymbol{\tau}} \boldsymbol{g}\right]=\operatorname{tr}\left(\boldsymbol{P}_{\boldsymbol{\tau}} \widetilde{M} \boldsymbol{P}_{\boldsymbol{\tau}}\right) \geq \frac{3 p_{0} \theta}{4}
$$

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 $|\boldsymbol{\tau}|<4 p_{0} \theta$ and square orthobasis $\boldsymbol{U}$. Let $\boldsymbol{g}^{\prime}=\boldsymbol{U}^{*} \boldsymbol{g}$, then $\boldsymbol{g}^{\prime}$ is standard i.i.d. Gaussian vector, provides alternative expression $\eta^{2}<\sum_{i=1}^{4 p_{0} \theta} g_{i}^{\prime 2} \sigma_{i}$ where $\sigma_{i} \leq 1+\mu|\boldsymbol{\tau}| \leq 1.1$. We obtain probability of $\eta^{2}$ to be small as

$$
\begin{equation*}
\mathbb{P}_{\boldsymbol{g} \mid \boldsymbol{w}}\left[\eta^{2}<\frac{p_{0} \theta}{2}\right] \leq \mathbb{P}_{\boldsymbol{g} \mid \boldsymbol{w}}\left[\eta^{2}-\mathbb{E}_{\boldsymbol{g} \mid \boldsymbol{w}} \eta^{2}<-\frac{p_{0} \theta}{4}\right] \leq 2 \exp \left(\frac{-\left(p_{0} \theta / 4\right)^{2}}{2(1+\mu|\boldsymbol{\tau}|)\left(12 p_{0} \theta+p_{0} \theta / 2\right)}\right) \leq 2 \exp \left(\frac{-p_{0} \theta}{440}\right) \tag{B.224}
\end{equation*}
$$

by applying moment bounds $\left(\sigma^{2}, R\right)=\left(12 p_{0} \theta(1+\mu|\boldsymbol{\tau}|), 2(1+\mu|\boldsymbol{\tau}|)\right)$. We thereby define event

$$
\begin{equation*}
\mathcal{E}_{\eta}=\left\{\sqrt{p_{0} \theta / 2} \leq \eta \leq 2 \sqrt{p_{0} \theta}\right\}, \tag{B.225}
\end{equation*}
$$

which holds w.h.p. based on (B.220), (B.223) and (B.224).
3. (Bound $\widetilde{\boldsymbol{\alpha}}$ ) Condition on $\mathcal{E}_{\eta} \cap \mathcal{E}_{\|x\|_{2}} \cap \mathcal{E}_{\tau}$. Use definition $\widetilde{\boldsymbol{\beta}}=\frac{1}{\eta} \widetilde{\boldsymbol{M}} \boldsymbol{x}$ from (B.217), and properties of $\widetilde{\boldsymbol{M}}$ from B.218 we can obtain:

$$
\left\{\begin{array}{l}
\left\|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \frac{1}{\eta}\left\|\iota_{\boldsymbol{\tau}^{c}}^{*} \widetilde{\boldsymbol{M}} \iota_{\boldsymbol{\tau}}\right\|_{2}\|\boldsymbol{x}\|_{2} \leq \frac{\mu \sqrt{p_{0}|\boldsymbol{\tau}|}}{\sqrt{p_{0} \theta / 2}} \cdot \sqrt{3 p_{0} \theta} \leq 3 \mu \sqrt{p_{0}|\boldsymbol{\tau}|}  \tag{B.226}\\
\left\|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}}\right\|_{2} \geq \frac{1}{\eta}\left\|\iota_{\boldsymbol{\tau}}^{*} \widetilde{\boldsymbol{M}} \iota_{\boldsymbol{\tau}}\right\|_{2}\|\boldsymbol{x}\|_{2} \geq \frac{\sqrt{1-\mu|\boldsymbol{\tau}|}}{2 \sqrt{p_{0} \theta}} \cdot \sqrt{p_{0} \theta} \geq 0.45
\end{array}\right.
$$

Use definition $\|\widetilde{\boldsymbol{\alpha}}\|_{2}=\|\boldsymbol{\chi}[\widetilde{\boldsymbol{\beta}}]\|_{2}$, condition on event

$$
\mathcal{E}_{\chi}:=\left\{\left\{\begin{array}{ll}
\sigma_{i} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \geq n \theta \mathcal{S}_{\nu_{2} \lambda}\left[\left|\boldsymbol{\beta}_{i}\right|\right]-\frac{c_{\mu}^{2} n \theta}{p}, & \forall i \in \boldsymbol{\tau} \\
\sigma_{i} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i} \leq 4 n \theta^{2}|\boldsymbol{\tau}|\left|\boldsymbol{\beta}_{i}\right|+\frac{c_{\mu} n \theta}{p}, & \forall i \in \boldsymbol{\tau}^{c}
\end{array}\right\}\right.
$$

also from Theorem B.2.1 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:

$$
\left\{\begin{array}{l}
\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq 4 n \theta^{2}|\boldsymbol{\tau}| \cdot\left\|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}^{c}}\right\|_{2}+\frac{c_{\mu} n \theta}{p} \cdot \sqrt{2 p_{0}} \leq 3 c_{\mu} n \theta\left(\frac{\sqrt{\theta}}{\log ^{2} \theta^{-1}}+\frac{c_{\mu}}{p}\right)  \tag{B.227}\\
\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\right\|_{2} \geq n \theta\left(\left\|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\tau}}\right\|_{2}-\nu_{2} \lambda \sqrt{|\boldsymbol{\tau}|}-\frac{c_{\mu}}{p} \sqrt{|\boldsymbol{\tau}|}\right) \geq n \theta\left(0.45-\sqrt{\frac{2}{\pi}} \cdot \frac{1}{5}-c_{\mu}\right) \geq 0.2 n \theta
\end{array}\right.
$$

since both $\theta|\boldsymbol{\tau}|, \mu p \theta|\boldsymbol{\tau}|<c_{\mu}$, we have

$$
\left\{\begin{array}{l}
\sqrt{1+\mu p}\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq 3 c_{\mu} n \theta \sqrt{1+\mu p}\left(\sqrt{\theta}+p^{-1}\right) \leq 6 c_{\mu} n \theta \\
\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \frac{6 c_{\mu}^{3 / 2} n \theta}{\log ^{2} \theta^{-1}} \min \left\{\frac{1}{\sqrt{|\boldsymbol{\tau}|}}, \frac{1}{\sqrt{\mu p}}, \frac{1}{\mu p \sqrt{\theta}|\boldsymbol{\tau}|}\right\} \leq 24 \sqrt{c_{\mu}} n \theta \gamma
\end{array}\right.
$$

which satisfies $\overline{\mathrm{B} .219}$, since $\mu|\tau|<c_{\mu}<\frac{1}{1000}$,

$$
\begin{equation*}
(1+\gamma \sqrt{1+\mu p})\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq\left(24 \sqrt{c_{\mu}}+6 c_{\mu}\right) n \theta \gamma \leq 0.1 n \theta \gamma \leq \gamma \sqrt{1-\mu|\boldsymbol{\tau}|}\left\|\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\tau}}\right\|_{2} \tag{B.228}
\end{equation*}
$$

Finally, given $p_{0} \theta>1000 c$, this result holds with probability at least

$$
\begin{equation*}
1-\underbrace{\mathbb{P}\left[\mathcal{E}_{\tau}^{c}\right]}_{\text {Theorem B.1.1 }}-\underbrace{\mathbb{P}\left[\mathcal{E}_{\|x\|_{2}}^{c}\right]}_{\text {Theorem B.1.2] }}-\underbrace{\mathbb{P}\left[\mathcal{E}_{\eta}^{c}\right]}_{\text {B.225] }}-\underbrace{\mathbb{P}\left[\mathcal{E}_{\chi}^{c}\right]}_{\text {Theorem B.3.4 }} \geq 1-\frac{2}{p_{0} \theta}-\frac{1}{n}-4 \exp \left(\frac{-p_{0} \theta}{440}\right) \geq 1-\frac{1}{c} \tag{B.229}
\end{equation*}
$$

## B.7.2 Minimization near subspace (Proof of Theorem 3.5.1)

Before we start the proof of theorem, writing $\boldsymbol{g}=\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})$ and $\boldsymbol{H}=\operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a})$, we will first restate the results of Theorem 3.4.1 in simplified terms. The theorem shows that for any $\boldsymbol{a} \in \mathbb{S}^{p-1}$ whose distance to
subspace $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \gamma$, then at least one of the the following statement hold:

$$
\begin{align*}
& \|\boldsymbol{g}\|_{2} \geq \eta_{g}  \tag{B.230}\\
& \lambda_{\min }(\boldsymbol{H}) \leq-\eta_{v}  \tag{B.231}\\
& \boldsymbol{H} \succ \eta_{c} \cdot \boldsymbol{P}_{\boldsymbol{a}^{\perp}} \tag{B.232}
\end{align*}
$$

Furthermore, $\varphi_{\rho}$ is retractive near $\mathcal{S}_{\boldsymbol{\tau}}$ : wherever $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right) \geq \frac{\gamma}{2}$, writing $\boldsymbol{\alpha}(\boldsymbol{a}), \boldsymbol{\alpha}(\boldsymbol{g})$ to be the coefficient of $\boldsymbol{a}$, $\boldsymbol{g}$, we have

$$
\begin{equation*}
\left\langle\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\rangle \geq \eta_{r}\left\|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\|_{2} . \tag{B.233}
\end{equation*}
$$

Also, the the gradient, Hessian and the third order derivative are all bounded as follows:

Remark B.7.2. With high probability, for every $\boldsymbol{a}$ whose $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)<\gamma$, its $\max \left\{\|\boldsymbol{g}\|_{2},\|\boldsymbol{H}\|_{2},\|\nabla \boldsymbol{H}\|_{2}\right\} \leq \bar{\eta}=$ $\operatorname{poly}(n, p)$.

We state Theorem B.7.2 without explicit proof since its derivation is similar to the proof in Theorem 3.4.1,
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 $\boldsymbol{v}=\mathbf{0}$ ), then the iterates

$$
\begin{equation*}
\boldsymbol{a}^{(k+1)}=\boldsymbol{P}_{\mathbb{S} p-1}\left[\boldsymbol{a}^{(k)}-t \boldsymbol{g}^{(k)}-t^{2} \boldsymbol{v}^{(k)}\right] \tag{B.234}
\end{equation*}
$$

converges toward the minimizer $\overline{\boldsymbol{a}}$ 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 $\overline{\boldsymbol{a}}$ 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 $\left\|\boldsymbol{a}^{(k)}-\overline{\boldsymbol{a}}\right\|_{2} \leq\left(1-t \eta_{c}\right)^{-(k-\bar{k})}\left\|\boldsymbol{a}^{(\bar{k})}-\overline{\boldsymbol{a}}\right\|_{2}$ [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 $\overline{\boldsymbol{a}}$.

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 (3.68)
is satisfied. Note that since $\varphi_{\rho}$ is a smooth function, $\boldsymbol{a} \rightarrow \varphi_{\rho} \circ \boldsymbol{P}_{\mathbb{S}^{p}-1}(\boldsymbol{a})$ admits a version of Taylor's theorem (see also AMS09](Section 7.1.3)): for any $\boldsymbol{\xi} \perp \boldsymbol{a}$, writing $\boldsymbol{a}^{+}=\boldsymbol{P}_{\mathbb{S}^{p-1}}[\boldsymbol{a}+\boldsymbol{\xi}]$,

$$
\begin{equation*}
\left|\varphi_{\rho}\left(\boldsymbol{a}^{+}\right)-\left(\varphi_{\rho}(\boldsymbol{a})+\left\langle\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a}), \boldsymbol{\xi}\right\rangle+\frac{1}{2} \boldsymbol{\xi}^{*} \operatorname{Hess}\left[\varphi_{\rho}\right](\boldsymbol{a}) \boldsymbol{\xi}\right)\right| \leq \bar{\eta}\|\boldsymbol{\xi}\|_{2}^{3} \tag{B.235}
\end{equation*}
$$

using $\|\nabla \boldsymbol{H}\|_{2} \leq \bar{\eta}$. Now, let $\boldsymbol{\xi}=-t \boldsymbol{g}-t^{2} \boldsymbol{v}$ as in the iterates (3.67). Suppose the Armijo step condition (3.68) does not hold, so

$$
\begin{equation*}
\varphi_{\rho}\left(\boldsymbol{a}^{+}\right)>\varphi_{\rho}(\boldsymbol{a})-\frac{1}{2}\left(t\|\boldsymbol{g}\|_{2}^{2}+\frac{1}{2} t^{4} \eta_{v}\|\boldsymbol{v}\|_{2}^{2}\right) \tag{B.236}
\end{equation*}
$$

Since $\boldsymbol{g}^{*} \boldsymbol{v} \geq 0$ and $\boldsymbol{v}^{*} \boldsymbol{H} \boldsymbol{v} \leq-\eta_{v}\|\boldsymbol{v}\|_{2}^{2}$ or $\boldsymbol{v}=\mathbf{0}$, using $\|\boldsymbol{a}+\boldsymbol{b}\|_{2}^{3} \leq 4\|\boldsymbol{a}\|_{2}^{3}+4\|\boldsymbol{b}\|_{2}^{3}$ (Hölder's inequality) and $\|\boldsymbol{H}\|_{2}<\bar{\eta}$, we can derive

$$
\begin{align*}
& \left\langle\boldsymbol{g},-t \boldsymbol{g}-t^{2} \boldsymbol{v}\right\rangle+\frac{1}{2}\left(t \boldsymbol{g}+t^{2} \boldsymbol{v}\right)^{*} \boldsymbol{H}\left(t \boldsymbol{g}+t^{2} \boldsymbol{v}\right)+c\left\|t \boldsymbol{g}+t^{2} \boldsymbol{v}\right\|_{2}^{3}>-\frac{1}{2}\left(t\|\boldsymbol{g}\|_{2}^{2}+\frac{1}{2} t^{4} \eta_{v}\|\boldsymbol{v}\|_{2}^{2}\right) \\
\Longrightarrow & -\frac{1}{2} t\|\boldsymbol{g}\|_{2}^{2}+\frac{1}{2} t^{2} \boldsymbol{g}^{*} \boldsymbol{H} \boldsymbol{g}+t^{3} \boldsymbol{v}^{*} \boldsymbol{H} \boldsymbol{g}-\frac{1}{4} t^{4} \eta_{v}\|\boldsymbol{v}\|_{2}^{2}+4 \bar{\eta} t^{3}\|\boldsymbol{g}\|_{2}^{3}+4 \bar{\eta} t^{6}\|\boldsymbol{v}\|_{2}^{3}>0 \\
\Longrightarrow & -\frac{1}{2} t\|\boldsymbol{g}\|_{2}^{2}+t^{2}\left(\frac{1}{2} \bar{\eta}\|\boldsymbol{g}\|_{2}^{2}+t \bar{\eta}\|\boldsymbol{v}\|_{2}\|\boldsymbol{g}\|_{2}+4 \bar{\eta} t\|\boldsymbol{g}\|_{2}^{3}\right)-\frac{1}{4} t^{4} \eta_{v}\|\boldsymbol{v}\|_{2}^{2}+4 \bar{\eta} t^{6}\|\boldsymbol{v}\|_{2}^{3}>0 . \tag{B.237}
\end{align*}
$$

If

$$
\begin{equation*}
t<T=\min \left\{\frac{\|\boldsymbol{g}\|_{2}}{\bar{\eta}\|\boldsymbol{g}\|_{2}+2 \bar{\eta} t\|\boldsymbol{v}\|_{2}+8 \bar{\eta} t\|\boldsymbol{g}\|_{2}^{2}}, \sqrt{\frac{\eta_{v}}{16 \bar{\eta}\|\boldsymbol{v}\|_{2}}}\right\} \tag{B.238}
\end{equation*}
$$

then B.237) $<0$ contradicting B.236. Using our bounds on $\|\boldsymbol{g}\|_{2}, \bar{\eta}, \eta_{v}$ and $\|\boldsymbol{v}\|$, it follows that $T$ is lower bounded by a polynomial poly $\left(n^{-1}, p^{-1}\right)$.
2.(Bounds on $\left.d_{\alpha}\left(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}}\right), d_{\alpha}\left(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}\right)\right)$ We will show there are numerical constants $c_{g}, c_{v}$ such that

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq c_{g} n \theta \gamma \quad \text { and } \quad d_{\alpha}\left(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq c_{v} n \theta p \tag{B.239}
\end{equation*}
$$

Define

$$
\chi_{\ell^{1}}[\boldsymbol{\beta}]=\breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \operatorname{prox}_{\lambda \ell^{1}}[\check{\boldsymbol{a}} * \boldsymbol{y}], \quad \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]=\widetilde{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \operatorname{prox}_{\lambda \rho}[\check{\boldsymbol{a}} * \boldsymbol{y}]
$$

then the gradient can be written as B.193

$$
\begin{align*}
\operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a}) & =\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}] \boldsymbol{\alpha}-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right)  \tag{B.240}\\
\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a}) & =\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}] \boldsymbol{\alpha}-\boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]\right) \tag{B.241}
\end{align*}
$$

Use the following inequalities:

$$
\begin{aligned}
\frac{1}{2} n \theta \leq\left|\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right| & \leq \frac{3}{2} n \theta, \\
\left\|\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2} & \leq \frac{1}{20} n \theta \gamma, \\
\left\|\boldsymbol{I}-\boldsymbol{\alpha} \boldsymbol{\beta}^{*}\right\|_{2} & \leq 4 \sqrt{p}, \\
\left\|\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]-\boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]\right\|_{2} & \leq n \theta^{4},
\end{aligned}
$$

where the first and second bounds of $\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]$ based on event B.181); the third by observing $\|\boldsymbol{\alpha}\|_{2} \leq 2$ and $\|\boldsymbol{\beta}\|_{2} \leq 2+c_{\mu} \sqrt{p}$; the last from (B.115) of Theorem B.5.6 when $\delta$ is sufficiently small. Hence, by definition of $d_{\alpha}\left(\cdot, \mathcal{S}_{\boldsymbol{\tau}}\right)$ 3.50 and knowing $\boldsymbol{a}$ is close to subspace $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2} \leq \gamma$, via triangle inequality, we get

$$
\begin{align*}
d_{\alpha}\left(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}}\right) & \leq d_{\alpha}\left(\operatorname{grad}\left[\varphi_{\ell^{1}}\right](\boldsymbol{a}), \mathcal{S}_{\boldsymbol{\tau}}\right)+d_{\alpha}\left(\operatorname{grad}\left[\varphi_{\rho}\right](\boldsymbol{a})-\operatorname{grad}\left[\boldsymbol{\varphi}_{\ell^{1}}\right](\boldsymbol{a}), \mathcal{S}_{\boldsymbol{\tau}}\right) \\
& \left.\leq \| \boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}] \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right]_{\boldsymbol{\tau}^{c}}\left\|_{2}+\right\|\left(\boldsymbol{I}-\boldsymbol{\alpha} \boldsymbol{\beta}^{*}\right)\left(\boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right) \|_{2} \\
& \leq \frac{3}{2} n \theta \gamma+\frac{1}{20} n \theta \gamma+4 \sqrt{p} n \theta^{4} \\
& \leq 3 n \theta \gamma . \tag{B.242}
\end{align*}
$$

To bound the $d_{\alpha}$ norm of least eigenvector $\boldsymbol{v}$, note that $\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]>0$, we can conclude

$$
\boldsymbol{v}^{*} \nabla^{2} \varphi_{\rho}(\boldsymbol{a}) \boldsymbol{v} \leq \boldsymbol{v}^{*} \boldsymbol{P}_{\boldsymbol{a}^{\perp}} \nabla^{2} \varphi_{\rho}(\boldsymbol{a}) \boldsymbol{P}_{\boldsymbol{a}^{\perp}} \boldsymbol{v}+\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]=\boldsymbol{v}^{*} \boldsymbol{H} \boldsymbol{v}<-\eta_{v},
$$

expand $\nabla^{2} \varphi_{\rho}(\boldsymbol{a})$ as in (B.102), and since $\boldsymbol{v}$ is the eigenvector of smallest eigenvalue $\lambda_{\min }<-\eta_{v}$,

$$
\begin{equation*}
\boldsymbol{P}_{\boldsymbol{a} \perp} \nabla^{2} \varphi_{\rho}(\boldsymbol{a}) \boldsymbol{P}_{\boldsymbol{a} \perp} \boldsymbol{v}=\left(\boldsymbol{I}-\boldsymbol{a} \boldsymbol{a}^{*}\right) \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \nabla \operatorname{prox}_{\lambda \rho}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}] \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{v} \boldsymbol{v}=\lambda_{\min } \boldsymbol{v} \tag{B.243}
\end{equation*}
$$

hence there exists $\alpha(\boldsymbol{v})$ satisfies $\boldsymbol{v}=\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\alpha}(\boldsymbol{v})$ and

$$
\boldsymbol{\alpha}(\boldsymbol{v})=\lambda_{\min }^{-1}\left[\check{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \nabla \operatorname{prox}_{\lambda \rho}[\breve{\boldsymbol{a}} * \boldsymbol{y}] \check{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota} \boldsymbol{v}-\left(\boldsymbol{\beta}^{*} \check{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \nabla \operatorname{prox}_{\lambda \rho}[\breve{\boldsymbol{a}} * \boldsymbol{y}] \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota} \boldsymbol{v}\right) \boldsymbol{\alpha}\right] .
$$

Now since $\nabla \operatorname{prox}_{\lambda \rho}[\widetilde{\boldsymbol{a}} * \boldsymbol{y}]$ is a diagonal matrix with entries in $[0,1]$,

$$
\begin{equation*}
d_{\alpha}\left(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq\|\boldsymbol{\alpha}(\boldsymbol{v})\|_{2} \leq\left|\lambda_{\min }\right|^{-1}\left\|\iota \boldsymbol{C}_{\boldsymbol{a}_{0}}\right\|_{2}\left\|\boldsymbol{x}_{0}\right\|_{1}^{2}\|\boldsymbol{v}\|_{2}\left(1+\|\boldsymbol{\alpha}\|_{2}\|\boldsymbol{\beta}\|_{2}\right)<c_{v} n \theta p \tag{B.244}
\end{equation*}
$$

where we use upper bound of $\left\|\boldsymbol{x}_{0}\right\|_{1}<c n \theta$ from Theorem B.1.2 and $\left|\lambda_{\min }\right|>\eta_{v}>c n \theta \lambda$ from Theorem B.6.2
3. (Iterates stay within widened subspace). Suppose (B.233) holds. We will show that whenever

$$
\begin{equation*}
t \leq T^{\prime}=\frac{1}{10 n \theta} \tag{B.245}
\end{equation*}
$$

then setting $\boldsymbol{a}^{+}=\boldsymbol{P}_{\mathbb{S}^{p}-1}\left[\boldsymbol{a}-t \boldsymbol{g}-t^{2} \boldsymbol{v}\right]$, we have

$$
\begin{equation*}
\left|d_{\alpha}\left(\boldsymbol{a}^{+}, \mathcal{S}_{\tau}\right)-d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\tau}\right)\right| \leq \frac{\gamma}{2}, \tag{B.246}
\end{equation*}
$$

and whenever $d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\tau}\right) \in\left[\frac{\gamma}{2}, \gamma\right]$

$$
\begin{equation*}
d_{\alpha}^{2}\left(\boldsymbol{a}^{+}, \mathcal{S}_{\tau}\right) \leq d_{\alpha}^{2}\left(\boldsymbol{a}, \mathcal{S}_{\tau}\right)-t \cdot c^{\prime} n \theta \gamma^{2} . \tag{B.247}
\end{equation*}
$$

If both B.246) and B.247) hold, then all iterates $\boldsymbol{a}^{(k)}$ will stay near the subspace: $d_{\alpha}\left(\boldsymbol{a}^{(k)}, \mathcal{S}_{\tau}\right)<\gamma$.
To derive (B.246), since both $\boldsymbol{g} \perp \boldsymbol{a}$ and $\boldsymbol{v} \perp \boldsymbol{a}$ we have $\left\|\boldsymbol{a}-t \boldsymbol{g}-t^{2} \boldsymbol{v}\right\|_{2}^{2}=\|\boldsymbol{a}\|_{2}^{2}+\left\|\boldsymbol{g}+t^{2} \boldsymbol{v}\right\|_{2}^{2}>1$, and since $d_{\alpha}\left(\cdot, \mathcal{S}_{\boldsymbol{\tau}}\right)$ is a seminorm Theorem B.2.2

$$
\begin{align*}
d_{\alpha}\left(\boldsymbol{a}^{+}, \mathcal{S}_{\tau}\right) & =d_{\alpha}\left(\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a}-t \boldsymbol{g}-t^{2} \boldsymbol{v}\right], \mathcal{S}_{\boldsymbol{\tau}}\right) \leq d_{\alpha}\left(\boldsymbol{a}-t \boldsymbol{g}-t^{2} \boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}\right) \\
& \leq d_{\alpha}\left(\boldsymbol{a}, \mathcal{S}_{\tau}\right)+t d_{\alpha}\left(\boldsymbol{g}, \mathcal{S}_{\boldsymbol{\tau}}\right)+t^{2} d_{\alpha}\left(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}\right) \tag{B.248}
\end{align*}
$$

suggests (B.246) holds via B.239) and let $n>C p^{5} \theta^{-2}$, we have

$$
\begin{equation*}
t d_{\alpha}\left(\boldsymbol{g}, \mathcal{S}_{\tau}\right)+t^{2} d_{\alpha}\left(\boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}\right) \leq \frac{c_{g} n \theta \gamma}{10 n \theta}+\frac{c_{v} n \theta_{p}}{(10 n \theta)^{2}}<\frac{\gamma}{2} \tag{B.249}
\end{equation*}
$$

with sufficiently large $C$.
 and (B.244), define

$$
\begin{align*}
& \boldsymbol{\alpha}(\boldsymbol{g})=\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}] \boldsymbol{\alpha}(\boldsymbol{a})-\chi_{\rho}[\boldsymbol{\beta}]  \tag{B.250}\\
& \boldsymbol{\alpha}(\boldsymbol{v})=\lambda_{\min }^{-1} \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \nabla \operatorname{prox}_{\lambda_{\rho}}[\check{\boldsymbol{a}} * \boldsymbol{y}] \breve{\boldsymbol{C}}_{\boldsymbol{x}_{0}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \iota \boldsymbol{v} . \tag{B.251}
\end{align*}
$$

By the retraction property and norm bounds,

$$
\begin{align*}
\left\langle\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\rangle & \geq \frac{1}{6 n \theta}\left\|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}  \tag{B.252}\\
\left\|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^{c}}\right\|_{2} & \leq \gamma  \tag{B.253}\\
\|\boldsymbol{\alpha}(\boldsymbol{v})\|_{2} & \leq c_{v} n \theta p . \tag{B.254}
\end{align*}
$$

Since $\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}>\frac{\gamma}{2}$,

$$
\begin{aligned}
\left\|\boldsymbol{a}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\|_{2} & \geq\left\|\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}] \boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2}-\left\|\left(\boldsymbol{I}-\boldsymbol{\alpha} \boldsymbol{\beta}^{*}\right)\left(\boldsymbol{\chi}_{\boldsymbol{\rho}}[\boldsymbol{\beta}]-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right)\right\|_{2} \\
& \geq\left|\boldsymbol{\beta}^{*} \boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right|\left\|\boldsymbol{\alpha}_{\boldsymbol{\tau}^{c}}\right\|_{2}-\left\|\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]_{\boldsymbol{\tau}^{c}}\right\|_{2}-\left\|\left(\boldsymbol{I}-\boldsymbol{\alpha} \boldsymbol{\beta}^{*}\right)\right\|_{2}\left\|\left(\boldsymbol{\chi}_{\rho}[\boldsymbol{\beta}]-\boldsymbol{\chi}_{\ell^{1}}[\boldsymbol{\beta}]\right)\right\|_{2}
\end{aligned}
$$

$$
\begin{align*}
& \geq \frac{1}{2} n \theta \times \frac{\gamma}{2}-\frac{1}{20} n \theta \gamma+2 n \theta^{4} \\
& \geq \frac{1}{10} n \theta \gamma \tag{B.255}
\end{align*}
$$

Finally, we can bound $d_{\alpha}\left(\boldsymbol{a}^{+}, \mathcal{S}_{\boldsymbol{\tau}}\right)$ as

$$
\begin{align*}
d_{\alpha}^{2}\left(\boldsymbol{a}^{+}, \mathcal{S}_{\boldsymbol{\tau}}\right) & \leq d_{\alpha}^{2}\left(\boldsymbol{a}-t \boldsymbol{g}-t^{2} \boldsymbol{v}, \mathcal{S}_{\boldsymbol{\tau}}\right) \\
& \leq\left\|\left[\boldsymbol{\alpha}(\boldsymbol{a})-t \boldsymbol{\alpha}(\boldsymbol{g})-t^{2} \boldsymbol{\alpha}(\boldsymbol{v})\right]_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \\
& =\left\|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-2 t\left\langle\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^{c}},[\boldsymbol{\alpha}(\boldsymbol{g})+t \boldsymbol{\alpha}(\boldsymbol{v})]_{\boldsymbol{\tau}^{c}}\right\rangle+t^{2}\left\|[\boldsymbol{\alpha}(\boldsymbol{g})+t \boldsymbol{\alpha}(\boldsymbol{v})]_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2} \\
& \leq\left\|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-2 t\left\langle\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^{c}}, \boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\rangle+2 t^{2}\left\|\boldsymbol{\alpha}(\boldsymbol{a})_{\boldsymbol{\tau}^{c}}\right\|_{2}\|\boldsymbol{\alpha}(\boldsymbol{v})\|_{2}+2 t^{2}\left\|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}+2 t^{4}\|\boldsymbol{\alpha}(\boldsymbol{v})\|_{2}^{2} \\
& \leq d^{2}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)-2 t\left[\left(\frac{1}{3 n \theta}-t\right)\left\|\boldsymbol{\alpha}(\boldsymbol{g})_{\boldsymbol{\tau}^{c}}\right\|_{2}^{2}-t n \theta p \gamma-t^{3}\left(c_{v} n \theta p\right)^{2}\right] \\
& \leq d^{2}\left(\boldsymbol{a}, \mathcal{S}_{\boldsymbol{\tau}}\right)-t \cdot c^{\prime} n \theta \gamma^{2} \tag{B.256}
\end{align*}
$$

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 $p, \operatorname{poly}(n, p)$, 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 $\left\|\boldsymbol{a}^{(k)}-\overline{\boldsymbol{a}}\right\|_{2} \leq \varepsilon$ is bounded by $\operatorname{poly}(n, p) \log \varepsilon^{-1}$. In particular, the number of iterations required to achieve $\left\|\boldsymbol{a}^{(k)}-\overline{\boldsymbol{a}}\right\|_{2} \leq \mu+1 / p$ is bounded by a polynomial in $(n, p)$, as claimed.

## B. 8 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 $\left(\boldsymbol{a}_{0}, \boldsymbol{x}_{0}\right)$. Set

$$
\begin{equation*}
\lambda^{(0)} \leftarrow 5 \kappa_{I} \widetilde{\mu} \quad \text { and } \quad I^{(0)} \leftarrow \operatorname{supp}\left(\mathcal{S}_{\lambda}\left[\boldsymbol{C}_{\boldsymbol{a}^{(0)}}^{*} \boldsymbol{y}\right]\right) \tag{B.257}
\end{equation*}
$$

where as each iteration of the algorithm consists of the following key steps:

- Sparse Estimation using Reweighted Lasso: Set

$$
\begin{equation*}
\boldsymbol{x}^{(k+1)} \leftarrow \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{a}^{(k)} * \boldsymbol{x}-\boldsymbol{y}\right\|_{2}^{2}+\sum_{i \notin I^{(k)}} \lambda^{(k)}\left|\boldsymbol{x}_{i}\right| ; \tag{B.258}
\end{equation*}
$$

- Kernel Estimation using Least Squares: Set

$$
\begin{equation*}
\boldsymbol{a}^{(k+1)} \leftarrow \boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\underset{\boldsymbol{a}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{x}^{(k+1)}-\boldsymbol{y}\right\|_{2}^{2}\right] ; \tag{B.259}
\end{equation*}
$$

- Continuation and reweighting by decreasing sparsity regularizer: Set

$$
\begin{equation*}
\lambda^{(k+1)} \leftarrow \frac{1}{2} \lambda^{(k)} \quad \text { and } \quad I^{(k+1)} \leftarrow \operatorname{supp}\left(\boldsymbol{x}^{(k+1)}\right) . \tag{B.260}
\end{equation*}
$$

Our analysis will show that $\boldsymbol{a}^{(k)}$ converges to $a_{0}$ 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

$$
\begin{equation*}
\widetilde{\mu}=\max \left\{\mu, p^{-1}\right\} . \tag{B.261}
\end{equation*}
$$

Our refinement algorithm will demand an initialization satisfying

$$
\begin{equation*}
\left\|\boldsymbol{a}^{(0)}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu} . \tag{B.262}
\end{equation*}
$$

Support density of $x_{0}$ 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

$$
\begin{equation*}
I=\operatorname{supp}\left(\boldsymbol{x}_{0}\right), \quad I^{(k)}=\operatorname{supp}\left(\boldsymbol{x}^{(k)}\right) . \tag{B.263}
\end{equation*}
$$

It should be clear that exact recovery is unlikely if $x_{0}$ 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 $x_{0}$ in a length- $p$ window:

$$
\begin{equation*}
\kappa_{I}=6 \max \{\theta p, \log n\}, \tag{B.264}
\end{equation*}
$$

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

$$
\begin{equation*}
\max _{\ell}|I \cap([p]+\ell)| \leq \kappa_{I} . \tag{B.265}
\end{equation*}
$$

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:

$$
\begin{equation*}
\left\|\boldsymbol{x}_{0}\right\|_{\square}:=\max _{\ell}\left\|\boldsymbol{P}_{([p]+\ell)} \boldsymbol{x}_{0}\right\|_{2} . \tag{B.266}
\end{equation*}
$$

Below, we will sometimes work with the $\square$-induced operator norm:

$$
\begin{equation*}
\|\boldsymbol{M}\|_{\square \rightarrow \square}=\sup _{\|\boldsymbol{x}\|_{\square} \leq 1}\|\boldsymbol{M} \boldsymbol{x}\|_{\square} \tag{B.267}
\end{equation*}
$$

For now, we note that in the Bernoulli-Gaussian model, $\left\|x_{0}\right\|_{\square}$ is typically not large

$$
\begin{equation*}
\left\|x_{0}\right\|_{\square} \leq \sqrt{\kappa_{I}} . \tag{B.268}
\end{equation*}
$$

## B.8.1 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

$$
\begin{equation*}
\sum_{i \in T^{c}}\left|x_{i}\right| \tag{B.269}
\end{equation*}
$$

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 B.8.1 (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 $\left\|\boldsymbol{x}_{0}\right\|_{\square} \leq \sqrt{\kappa_{I}}$ with $\kappa_{I} \geq 1$. If $\widetilde{\mu} \kappa_{I}^{2} \leq c_{\mu}$, then for every $T \subseteq I$ and $\boldsymbol{a}$ satisfying $\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu}$, the solution $\boldsymbol{x}^{+}$to the optimization problem

$$
\begin{equation*}
\min _{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}-\boldsymbol{y}\|_{2}^{2}+\lambda \sum_{i \in T^{c}}\left|\boldsymbol{x}_{i}\right|\right\} \tag{B.270}
\end{equation*}
$$

with

$$
\begin{equation*}
\lambda>5 \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.271}
\end{equation*}
$$

is unique with the form

$$
\begin{equation*}
\boldsymbol{x}^{+}=\boldsymbol{\iota}_{J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{\iota}_{J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{y}-\lambda \boldsymbol{P}_{J \backslash T} \boldsymbol{\sigma}\right) \tag{B.272}
\end{equation*}
$$

where $\boldsymbol{\sigma}=\operatorname{sign}\left(\boldsymbol{x}^{+}\right)$. Its support set $J$ satisfies

$$
\begin{equation*}
\left(T \cup I_{\geq 3 \lambda}\right) \subseteq J \subseteq I \tag{B.273}
\end{equation*}
$$

and its entrywise error is bounded as

$$
\begin{equation*}
\left\|\boldsymbol{x}^{+}-\boldsymbol{x}_{0}\right\|_{\infty} \leq 3 \lambda \tag{B.274}
\end{equation*}
$$

Above, $c_{\mu}>0$ is a positive numerical constant.

We prove Theorem B.8.1 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 B.8.4.

Proof. 1. (Restricted support Lasso problem). We first consider the restricted problem

$$
\begin{equation*}
\min _{\boldsymbol{w} \in \mathbb{R}^{|I|}}\left\{\frac{1}{2}\left\|\boldsymbol{a} * \boldsymbol{\iota}_{I} \boldsymbol{w}-\boldsymbol{y}\right\|_{2}^{2}+\lambda \sum_{i \in T^{c}}\left|\left(\boldsymbol{\iota}_{I} \boldsymbol{w}\right)_{i}\right|\right\} \tag{B.275}
\end{equation*}
$$

Under our assumptions, provided $c<\frac{1}{9}$, Theorem B.8.6 implies that

$$
\begin{equation*}
\iota_{I}^{*} C_{a}^{*} C_{a} \iota_{I}=\left[C_{a}^{*} C_{a}\right]_{I, I} \succ \mathbf{0} \tag{B.276}
\end{equation*}
$$

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

$$
\begin{equation*}
\boldsymbol{\iota}_{I}^{*} \boldsymbol{C}_{a}^{*} \boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{\iota}_{I} \boldsymbol{w}_{\star} \in \boldsymbol{\iota}_{I}^{*} \boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{y}-\lambda \partial\left\|\boldsymbol{P}_{T^{c}}[\cdot]\right\|_{1}\left(\boldsymbol{w}_{\star}\right) \tag{B.277}
\end{equation*}
$$

Writing $J=\operatorname{supp}\left(\boldsymbol{\iota}_{I} \boldsymbol{w}_{\star}\right) \subseteq I, \boldsymbol{C}_{\boldsymbol{a} J}=\boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{\iota}_{J}, \boldsymbol{w}_{J}=\boldsymbol{\iota}_{J}^{*} \boldsymbol{\iota}_{I} \boldsymbol{w}_{\star}$ the corresponding sub-vector containing the nonzero entries of $\boldsymbol{w}_{\star}$ and $\boldsymbol{\sigma}_{J \backslash T}=\boldsymbol{\iota}_{J}^{*} \boldsymbol{P}_{T^{c}}\left[\operatorname{sign}\left(\boldsymbol{\iota}_{I} \boldsymbol{w}_{*}\right)\right]$, the condition B.277 is satisfied if and only if

$$
\begin{align*}
& \boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J} \boldsymbol{w}_{J}=\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{y}-\lambda \boldsymbol{\sigma}_{J \backslash T},  \tag{B.278}\\
& \left\|\boldsymbol{C}_{\boldsymbol{a}}{ }_{I \backslash J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a} J} \boldsymbol{w}_{J}-\boldsymbol{y}\right)\right\|_{\infty} \leq \lambda . \tag{B.279}
\end{align*}
$$

We will argue that under our assumptions, $J$ necessarily contains all of the large entries of $\boldsymbol{x}_{0}$ :

$$
\begin{equation*}
I_{>3 \lambda}=\left\{\ell \in I| | \boldsymbol{x}_{0 \ell} \mid>3 \lambda\right\} \subseteq J \tag{B.280}
\end{equation*}
$$

We show this by contradiction - namely, if some large entry of $\boldsymbol{x}_{0}$ is not in $J$, then the dual condition B.279) is violated, contradicting the optimality of $\boldsymbol{w}_{\star}$. To this end, note that by Theorem B.8.7, $C_{a}{ }_{J}^{*} C_{a, J}$ has full rank. From B.278,

$$
\begin{equation*}
\boldsymbol{w}_{J}=\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right]^{-1}\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{y}-\lambda \boldsymbol{\sigma}_{J \backslash T}\right] . \tag{B.281}
\end{equation*}
$$

Write $\boldsymbol{x}_{0 J}=\boldsymbol{\iota}_{J}^{*} \boldsymbol{x}_{0}$ and $\left(\boldsymbol{x}_{0}\right)_{I \backslash J}=\boldsymbol{P}_{I \backslash J} \boldsymbol{x}_{0}$. We can further notice that

$$
\begin{align*}
& \boldsymbol{C}_{\boldsymbol{a} J} \boldsymbol{w}_{J}-\boldsymbol{y}=\left(\boldsymbol{C}_{\boldsymbol{a} J}\left[\boldsymbol{C}_{a_{J}}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*}-\boldsymbol{I}\right) \boldsymbol{y}-\lambda \boldsymbol{C}_{\boldsymbol{a} J}\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{\sigma}_{J \backslash T} \\
& =\left(\boldsymbol{C}_{\boldsymbol{a} J}\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*}-\boldsymbol{I}\right) \boldsymbol{C}_{\boldsymbol{a}_{0} J} \boldsymbol{x}_{0 J}+\left(\boldsymbol{C}_{\boldsymbol{a} J}\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*}-\boldsymbol{I}\right) \boldsymbol{C}_{\boldsymbol{a}_{0} I \backslash J}\left(\boldsymbol{x}_{0}\right)_{I \backslash J} \\
& -\lambda \boldsymbol{C}_{\boldsymbol{a} J}\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{\sigma}_{J \backslash T} \\
& =\left(\boldsymbol{C}_{\boldsymbol{a} J}\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{C}_{a_{J}}^{*}-\boldsymbol{I}\right) \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a} J} \boldsymbol{x}_{0 J}+\left(\boldsymbol{C}_{\boldsymbol{a} J}\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*}-\boldsymbol{I}\right) \boldsymbol{C}_{\boldsymbol{a}_{0} I \backslash J}\left(\boldsymbol{x}_{0}\right)_{I \backslash J} \\
& -\lambda \boldsymbol{C}_{\boldsymbol{a} J}\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{\sigma}_{J \backslash T}, \tag{B.282}
\end{align*}
$$

where in the final line we have used that

$$
\begin{equation*}
\left(C_{a, J}\left[C_{a}{ }_{J}^{*} C_{a J}\right]^{-1} C_{a}^{*}-I\right) C_{a J}=0 \tag{B.283}
\end{equation*}
$$

Suppose that $J$ is a strict subset of $I$ (otherwise, if $J=I$, we are done). Take any $i \in I \backslash J$ such that $\left|\boldsymbol{x}_{0 i}\right|=\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty^{\prime}}$ and let $\xi=\operatorname{sign}\left(\boldsymbol{x}_{0 i}\right)$. Using B.282, Theorem B.8.7 and Theorem B.8.8, we have

$$
\begin{align*}
& -\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}\left[\boldsymbol{a}_{0}\right] \boldsymbol{x}_{0 i} \\
& +\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}}\left(\boldsymbol{x}_{0}\right)_{I \backslash(J \cup\{i\})} \\
& +\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} \\
& +\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 \backslash T}  \tag{B.284}\\
& \geq\left(\left\langle s_{i}[\boldsymbol{a}], s_{i}\left[\boldsymbol{a}_{0}\right]\right\rangle-\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 \rightarrow \infty}\left\|\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} s_{i}\left[\boldsymbol{a}_{0}\right]\right\|_{\infty}\right)\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \\
& -\left(\left\|s_{i}[\boldsymbol{a}]^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} I \backslash\{i\}}\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 \rightarrow \infty}\left\|\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} I \backslash J}\right\|_{\infty \rightarrow \infty}\right)\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \\
& -\left(\left\|s_{i}[\boldsymbol{a}]^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a} \boldsymbol{a}_{J}}\right\|_{2}+\left\|s_{i}[\boldsymbol{a}]^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right\|_{2}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{J} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1}\right\|_{\square \rightarrow \square}\left\|\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}_{J}}\right\|_{\square \rightarrow \square}\right) \sqrt{2}\left\|\boldsymbol{x}_{0}\right\|_{\square} \\
& -\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 \rightarrow \infty}\left\|\boldsymbol{\sigma}_{J \backslash T}\right\|_{\infty}  \tag{B.285}\\
& \geq\left(\left(1-\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}\right)-C_{1} \kappa_{I} \widetilde{\mu} \times 1 \times \widetilde{\mu}\right)\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty}
\end{align*}
$$

$$
\begin{align*}
&-C_{2}\left(\kappa_{I} \widetilde{\mu}+\kappa_{I} \widetilde{\mu} \times 1 \times \kappa_{I} \widetilde{\mu}\right)\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \\
&-\left(2 \sqrt{\kappa_{I}}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}+C_{3} \sqrt{\kappa_{I}} \widetilde{\mu} \times 1 \times \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}\right)\left\|\boldsymbol{x}_{0}\right\|_{\square} \\
&-\lambda C_{4} \kappa_{I} \widetilde{\mu}  \tag{B.286}\\
& \geq \quad\left(1-C_{1}^{\prime} \kappa_{I} \widetilde{\mu}-C_{2}\left(\kappa_{I} \widetilde{\mu}\right)^{2}\right)\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \\
&-2 \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}-\left(C_{3} \kappa_{I}^{3 / 2} \widetilde{\mu}\right) \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}-\left(C_{4} \kappa_{I} \widetilde{\mu}\right) \lambda  \tag{B.287}\\
& \geq \quad \frac{1}{2}\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty}-\lambda / 2 \tag{B.288}
\end{align*}
$$

where the last line holds provided $\widetilde{\mu} \kappa_{I}^{2} \leq c_{\mu}$ to be a sufficiently small numerical constants. If $\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty}>3 \lambda$, 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

$$
\begin{equation*}
\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \leq 3 \lambda \tag{B.289}
\end{equation*}
$$

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

$$
\begin{equation*}
\min _{\boldsymbol{x}} \varphi_{\text {lasso }}(\boldsymbol{x}) \equiv \frac{1}{2}\|\boldsymbol{a} * \boldsymbol{x}-\boldsymbol{y}\|_{2}^{2}+\lambda \sum_{i \in T^{c}}\left|\boldsymbol{x}_{i}\right| \tag{B.290}
\end{equation*}
$$

To prove that $\boldsymbol{x}^{+}$is the unique optimal solution, it suffices to show that for every $i \in I^{c}$,

$$
\begin{equation*}
\left|s_{i}[\boldsymbol{a}]^{*}\left(\boldsymbol{a} * \boldsymbol{x}^{+}-\boldsymbol{y}\right)\right|<\lambda . \tag{B.291}
\end{equation*}
$$

Indeed, suppose that this inequality is in force. Write $\varepsilon=\lambda-\max _{i \in I^{c}}\left|s_{i}[\boldsymbol{a}]^{*}\left(\boldsymbol{a} * \boldsymbol{x}^{+}-\boldsymbol{y}\right)\right|$, and notice that from the KKT conditions for the restricted problem,

$$
\begin{equation*}
\mathbf{0} \in \boldsymbol{P}_{I} \partial_{\boldsymbol{x}} \varphi_{\mathrm{lasso}}(\boldsymbol{x}) \tag{B.292}
\end{equation*}
$$

Combining with B.291, we have that for every vector $\boldsymbol{\zeta}$ with $\operatorname{supp}(\boldsymbol{\zeta}) \subseteq I^{c}$ and $\|\boldsymbol{\zeta}\|_{\infty} \leq 1$, then $\varepsilon \boldsymbol{\zeta} \in$ $\partial \varphi_{\text {lasso }}\left(\boldsymbol{x}^{+}\right)$. Let $\boldsymbol{x}^{\prime}$ be any vector with $\boldsymbol{x}_{I^{c}}^{\prime} \neq \mathbf{0}$ and set $\boldsymbol{\zeta}=\mathcal{P}_{I^{c}} \operatorname{sign}\left(\boldsymbol{x}^{\prime}\right)$, then from the subgradient inequality,

$$
\begin{align*}
\varphi_{\mathrm{lasso}}\left(\boldsymbol{x}^{\prime}\right) & \geq \varphi_{\mathrm{lasso}}\left(\boldsymbol{x}^{+}\right)+\left\langle\varepsilon \boldsymbol{\zeta}, \boldsymbol{x}^{\prime}-\boldsymbol{x}^{+}\right\rangle \\
& \geq \varphi_{\mathrm{lasso}}\left(\boldsymbol{x}^{+}\right)+\varepsilon\left\|\boldsymbol{x}_{I^{c}}^{\prime}\right\|_{1} \tag{B.293}
\end{align*}
$$

which is strictly larger than $\varphi_{\text {lasso }}\left(\boldsymbol{x}^{+}\right)$. Hence, when B.291 holds, any optimal solution $\overline{\boldsymbol{x}}$ to the full Lasso problem must satisfy $\operatorname{supp}(\overline{\boldsymbol{x}}) \subseteq I$. By strong convexity of the restricted problem, the solution to (B.290) is
unique and equal to $\boldsymbol{x}^{+}$.
We finish by showing (B.291). Using the same expansion as above, we obtain

$$
\begin{align*}
& \left|s_{i}[\boldsymbol{a}]^{*}\left(\boldsymbol{C}_{\boldsymbol{a} J} \boldsymbol{w}_{J}-\boldsymbol{y}\right)\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 \backslash J}\left(\boldsymbol{x}_{0}\right)_{I \backslash 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 \backslash T}\right|  \tag{B.294}\\
& \leq\left(\left\|s_{i}[\boldsymbol{a}]^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} I \backslash 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 J}\right]^{-1}\right\|_{\infty \rightarrow \infty}\left\|\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} I \backslash J}\right\|_{\infty \rightarrow \infty}\right)\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \\
& +\left(\left\|s_{i}[\boldsymbol{a}]^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}_{J}}\right\|_{2}+\left\|s_{i}[\boldsymbol{a}]^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right\|_{2}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1}\right\|_{\square \rightarrow \square}\left\|\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}_{J}}\right\|_{\square \rightarrow \square}\right) \sqrt{2}\left\|\boldsymbol{x}_{0}\right\|_{\square} \\
& +\lambda\left\|s_{i}[\boldsymbol{a}]^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right\|_{1}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{J}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1}\right\|_{\infty \rightarrow \infty}\left\|\boldsymbol{\sigma}_{J \backslash T}\right\|_{\infty}  \tag{B.295}\\
& \leq C_{1}\left(\widetilde{\mu} \kappa_{I}+\widetilde{\mu} \kappa_{I} \times 1 \times \widetilde{\mu} \kappa_{I}\right) \times 2 \lambda \\
& +\left(2 \sqrt{\kappa_{I}}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}+C_{2} \sqrt{\kappa_{I}} \tilde{\mu} \times 1 \times \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}\right) \times \sqrt{\kappa_{I}} \\
& +\lambda C_{3} \times \widetilde{\mu} \kappa_{I}  \tag{B.296}\\
& \leq\left(\left(C_{1}+C_{3}\right) \widetilde{\mu} \kappa_{I}+C_{1}\left(\widetilde{\mu} \kappa_{I}\right)^{2}\right) \lambda+\left(2+C_{2} \widetilde{\mu} \kappa_{I}\right) \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}  \tag{B.297}\\
& <\lambda, \tag{B.298}
\end{align*}
$$

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 $x_{0}$ ) Finally we will be controlling $\left\|x_{J}^{+}-\left(x_{0}\right)_{J}\right\|_{\infty}$. Indeed, from Theorem B.8.8,

$$
\begin{align*}
\left\|\boldsymbol{x}_{J}^{+}-\left(\boldsymbol{x}_{0}\right)_{J}\right\|_{\infty}= & \left\|\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}_{0}-\lambda\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{\sigma}_{J \backslash T}-\left(\boldsymbol{x}_{0}\right)_{J}\right\|_{\infty} \\
\leq & \left\|\left[\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}_{J}}\left(\boldsymbol{x}_{0}\right)_{J}\right\|_{\infty}+\lambda\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{\sigma}_{J \backslash T}\right\|_{\infty} \\
& \quad+\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right]^{-1} \boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} I \backslash J}\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \\
\leq & 2\left\|\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}_{J}}\right\|_{\square \rightarrow \infty}\left\|\left(\boldsymbol{x}_{0}\right)_{J}\right\|_{\square}+2 \lambda+2\left\|\boldsymbol{C}_{\boldsymbol{a}}{ }_{J}^{*} \boldsymbol{C}_{\boldsymbol{a} I \backslash J}\right\|_{\infty \rightarrow \infty}\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \\
\leq & 2 \sqrt{2 \kappa_{I}}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}\left\|\boldsymbol{x}_{0}\right\|_{\square}+2 \lambda+2 \times 3 \widetilde{\mu} \times 2 \kappa_{I \backslash J} \times 3 \lambda \\
\leq & 3 \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}+2 \lambda+36 \lambda \widetilde{\mu} \kappa_{I} \\
\leq & 3 \lambda, \tag{B.299}
\end{align*}
$$

establishing the claim.

## B.8.2 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 $a$, we will show the solution of the least squares problem

$$
\begin{equation*}
\boldsymbol{a}^{+} \leftarrow \underset{\boldsymbol{a}^{\prime} \in \mathbb{R}^{p}}{\operatorname{argmin}} \frac{1}{2}\left\|\boldsymbol{a}^{\prime} * \boldsymbol{x}-\boldsymbol{y}\right\|_{2}^{2} \tag{B.300}
\end{equation*}
$$

is closer to $\boldsymbol{a}_{0}$ compared to $\boldsymbol{a}$. Observe that in Theorem B.8.1, the solution of B.272

$$
\begin{equation*}
\boldsymbol{x}=\boldsymbol{\iota}_{J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{\iota}_{J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}_{0}-\lambda \boldsymbol{P}_{J \backslash T} \boldsymbol{\sigma}\right), \tag{B.301}
\end{equation*}
$$

by assuming $\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J} \approx \boldsymbol{I}, \boldsymbol{a} \approx \boldsymbol{a}_{0}$ and $J \backslash T \approx \emptyset$, is a good approximation to the true sparse map $\boldsymbol{x}_{0}$

$$
\begin{equation*}
\boldsymbol{x} \approx \boldsymbol{I}\left(\boldsymbol{x}_{0}-\mathbf{0}\right)=\boldsymbol{x}_{0} \tag{B.302}
\end{equation*}
$$

furthermore, its difference to the true sparse map $\left\|\boldsymbol{x}_{0}-\boldsymbol{x}\right\|_{2}$ is proportional to $\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2}$ as

$$
\begin{equation*}
\boldsymbol{x}-\boldsymbol{x}_{0} \approx \boldsymbol{P}_{I}\left(\boldsymbol{C}_{a}^{*} \boldsymbol{C}_{a_{0}} \boldsymbol{x}_{0}-\boldsymbol{C}_{a}^{*} \boldsymbol{C}_{a} \boldsymbol{x}_{0}\right) \approx \boldsymbol{P}_{I}\left[\boldsymbol{C}_{a_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right)\right] \tag{B.303}
\end{equation*}
$$

To this end, since we know the solution of least square problem $\boldsymbol{a}^{+}$is simply

$$
\begin{equation*}
\boldsymbol{a}^{+}=\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}} \iota\right)^{-1}\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota \boldsymbol{a}_{0}\right) \tag{B.304}
\end{equation*}
$$

this implies the difference between the new $\boldsymbol{a}^{+}$and $\boldsymbol{a}_{0}$, has the relationship with $\boldsymbol{a}-\boldsymbol{a}_{0}$ roughly

$$
\begin{align*}
\boldsymbol{a}^{+}-\boldsymbol{a}_{0} & =\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{\iota}\right)^{-1}\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota \boldsymbol{a}_{0}-\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}} \iota \boldsymbol{a}_{0}\right) \approx(n \theta)^{-1} \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{x}_{0}-\boldsymbol{x}\right) \\
& \approx(n \theta)^{-1} \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}-\boldsymbol{a}_{0}\right) . \tag{B.305}
\end{align*}
$$

To make this point precise, we introduce the following lemma:

Lemma B.8.2 (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 $C, C^{\prime}, c, c^{\prime}, c_{\mu}$ such that if $\lambda<c^{\prime} \widetilde{\mu} \kappa_{I}, \widetilde{\mu} \kappa_{I}^{2} \leq c_{\mu}$ and $n>C p^{2} \log p$, then with probability at least $1-c / n$, for every $\boldsymbol{a}$ satisfying $\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu}$ and $\boldsymbol{x}$ of the form

$$
\begin{equation*}
\boldsymbol{x}=\boldsymbol{\iota}_{J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \iota_{J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{y}-\lambda \boldsymbol{P}_{J \backslash T} \boldsymbol{\sigma}\right) \tag{B.306}
\end{equation*}
$$

where the set $J, T$ satisfies $I_{>6 \lambda} \subseteq T \subseteq J \subseteq I$, we have

$$
\begin{equation*}
\frac{1}{n \theta}\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}} \boldsymbol{\iota} \boldsymbol{a}_{0}-\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right)\right\|_{2} \leq C^{\prime} \lambda\left(\widetilde{\lambda}+\widetilde{\mu} \kappa_{I}\right)+\frac{1}{32}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.307}
\end{equation*}
$$

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 Theorem B.8.1 and our assumptions on the set $T$, then $\boldsymbol{x}$ approximates $\boldsymbol{x}_{0}$ in the sense that

$$
\begin{align*}
\left\|\boldsymbol{x}-\boldsymbol{x}_{0}\right\|_{\infty} & \leq 3 \lambda  \tag{B.308}\\
\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} & \leq 3 \lambda  \tag{B.309}\\
\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash T}\right\|_{\infty} & \leq 6 \lambda \tag{B.310}
\end{align*}
$$

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 B.309) we know $|I \backslash J|=\left|\left\{i| | \boldsymbol{g}_{i} \mid \leq 3 \lambda, \boldsymbol{\omega}_{i} \neq 0\right\}\right|$. Since $\mathbb{P}\left[\boldsymbol{\omega}_{i} \neq 0\right]=\theta$ and $\mathbb{P}\left[\left|\boldsymbol{g}_{i}\right| \leq 3 \lambda\right] \leq 3 \lambda$, Theorem B.1.1 1 implies that with probability at least $1-2 / n$ :

$$
\begin{align*}
& |I \backslash J| \leq 3 \lambda n \theta+6 \sqrt{\lambda n \theta} \log n \leq 3 \widetilde{\lambda} n \theta  \tag{B.311}\\
& |I \backslash T| \leq 6 \lambda n \theta+12 \sqrt{\lambda n \theta} \log n \leq 6 \widetilde{\lambda} n \theta \tag{B.312}
\end{align*}
$$

and

$$
\begin{equation*}
\left|(I \backslash J) \cap s_{\ell}[I]\right| \leq 3 \lambda n \theta^{2}+6 \sqrt{\lambda n \theta^{2}} \log n \leq 3 \tilde{\lambda} n \theta^{2} \tag{B.313}
\end{equation*}
$$

together with base on properties of Bernoulli-Gaussian vector $\boldsymbol{x}_{0}$ from Appendix B. 1 and we conclude with probability at least $1-c / n$, all the following events hold:

$$
\begin{align*}
\frac{1}{2} n \theta \leq|I| & \leq 2 n \theta,  \tag{B.314}\\
\max _{\ell \neq 0}\left|I \cap s_{\ell}[I]\right| & \leq 2 n \theta^{2}  \tag{B.315}\\
\max _{\ell \neq 0}\left|(I \backslash J) \cap s_{\ell}[I]\right| & \leq 6 \widetilde{\lambda} n \theta^{2},  \tag{B.316}\\
\left\|\boldsymbol{x}_{0}\right\|_{\square}^{2} & \leq \kappa_{I},  \tag{B.317}\\
\left\|\widetilde{\boldsymbol{a}}_{0} * \boldsymbol{x}_{0}\right\|_{\square}^{2} & \leq \kappa_{I},  \tag{B.318}\\
\left\|\boldsymbol{x}_{0}\right\|_{2}^{2} & \leq 2 n \theta,  \tag{B.319}\\
\left\|\boldsymbol{x}_{0}\right\|_{1} & \leq 2 n \theta  \tag{B.320}\\
\max _{\ell \neq 0}\left\|\boldsymbol{P}_{I \cap s_{\ell}[I]} \boldsymbol{x}_{0}\right\|_{2}^{2} & \leq 2 n \theta^{2},  \tag{B.321}\\
\max _{\ell \neq 0}\left\|\boldsymbol{P}_{I \cap s_{\ell}[I \backslash J]} \boldsymbol{x}_{0}\right\|_{1} & \leq 12 \widetilde{\lambda} n \theta^{2} \tag{B.322}
\end{align*}
$$

$$
\begin{equation*}
\left\|\boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\right\|_{2}^{2} \leq 3 n \theta \tag{B.323}
\end{equation*}
$$

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

$$
\begin{equation*}
\iota^{*} C_{x}^{*} C_{x-x_{0}} \iota a_{0}=\iota^{*} C_{x_{0}}^{*} C_{x-x_{0}} \iota a_{0}+\iota^{*} C_{x-x_{0}}^{*} C_{x-x_{0}} \iota a_{0} \tag{B.324}
\end{equation*}
$$

where

$$
\begin{align*}
\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}} \boldsymbol{\iota} \boldsymbol{a}_{0}\right\|_{2} & \leq\left\|\boldsymbol{a}_{0}\right\|_{2}\left\|\boldsymbol{x}-\boldsymbol{x}_{0}\right\|_{2}^{2}+\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}\right\|_{2} \sqrt{2 p} \max _{\ell \neq 0}\left|\left\langle s_{\ell}\left[\boldsymbol{x}-\boldsymbol{x}_{0}\right], \boldsymbol{x}-\boldsymbol{x}_{0}\right\rangle\right| \\
& \leq\left\|\boldsymbol{x}-\boldsymbol{x}_{0}\right\|_{\infty}^{2} \times|I|+\sqrt{2 \widetilde{\mu} p^{2}}\left(\left\|\boldsymbol{x}-\boldsymbol{x}_{0}\right\|_{\infty}^{2} \times \max _{\ell \neq 0}\left|I \cap s_{\ell}[I]\right|\right) \\
& \leq C_{1}\left(\lambda^{2} n \theta+\sqrt{2 \widetilde{\mu} p^{2}}\left(\lambda^{2} n \theta^{2}\right)\right) \\
& \leq 2 C_{1} \lambda^{2} n \theta \tag{B.325}
\end{align*}
$$

we have that

$$
\begin{equation*}
\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}} \iota \boldsymbol{a}_{0}-\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}} \iota \boldsymbol{a}_{0}\right\|_{2} \leq 2 C_{1} \lambda^{2} n \theta \tag{B.326}
\end{equation*}
$$

2. (Extract the $\boldsymbol{a}_{0}-\boldsymbol{a}$ term) Observe that

$$
\begin{align*}
\iota^{*} & \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}} \boldsymbol{\iota} \boldsymbol{a}_{0} \\
= & \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{x}-\boldsymbol{x}_{0}\right) \\
= & \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\iota_{J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \iota_{J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}_{0}-\lambda \boldsymbol{P}_{J \backslash 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)\left(\boldsymbol{x}_{0}\right)_{J}-\boldsymbol{P}_{I \backslash J} \boldsymbol{x}_{0}\right) \\
= & \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} J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}} \boldsymbol{x}_{0}\right) \\
& +\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} J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{x}_{0}-\boldsymbol{C}_{\boldsymbol{a} J}\left(\boldsymbol{x}_{0}\right)_{J}\right) \\
& -\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I \backslash J} \boldsymbol{x}_{0} \\
& -\lambda \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} \iota_{J}^{*} \boldsymbol{P}_{J \backslash T} \boldsymbol{\sigma}, \tag{B.327}
\end{align*}
$$

where, the second term in B.327 is bounded as

$$
\begin{aligned}
& \left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{C}_{\boldsymbol{a} J}^{*}\left(\boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{x}_{0}-\boldsymbol{C}_{\boldsymbol{a} J}\left(\boldsymbol{x}_{0}\right)_{J}\right)\right\|_{2} \\
& \quad \leq\left\|\boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\right\|_{2} \times\left\|\boldsymbol{C}_{\boldsymbol{a}_{0} J}\right\|_{2}\left\|\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1}\right\|_{2} \times\left\|\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} I \backslash J}\right\|_{2} \times\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{2}
\end{aligned}
$$

$$
\begin{align*}
& \leq C_{2}\left(\sqrt{n \theta} \times 3 \times \widetilde{\mu} \kappa_{I} \times \lambda \sqrt{\widetilde{\lambda} n \theta}\right) \\
& \leq 3 C_{2} \widetilde{\mu} \kappa_{I} \lambda n \theta \tag{B.328}
\end{align*}
$$

the third term in B.327) is bounded as

$$
\begin{align*}
\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I \backslash J} \boldsymbol{x}_{0}\right\|_{2} & =\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{P}_{[ \pm p] \backslash 0}+\boldsymbol{e}_{0} \boldsymbol{e}_{0}^{*}\right) \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I \backslash J} \boldsymbol{x}_{0}\right\|_{2} \\
& \leq\left\|\boldsymbol{a}_{0}\right\|_{2}\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{2}^{2}+\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}\right\|_{2} \times \sqrt{2 p} \times \max _{\ell \neq 0}\left\|\boldsymbol{P}_{I \cap s_{\ell}[I \backslash J]} \boldsymbol{x}_{0}\right\|_{1} \times\left\|\left(\boldsymbol{x}_{0}\right)_{I \backslash J}\right\|_{\infty} \\
& \leq C_{3}\left(\lambda^{2} \times \widetilde{\lambda} n \theta+\sqrt{\widetilde{\mu} p^{2}} \times \widetilde{\lambda} n \theta^{2} \times \lambda\right) \\
& \leq 2 C_{3} \widetilde{\lambda} \lambda n \theta \tag{B.329}
\end{align*}
$$

and finally, write $\boldsymbol{\Delta}=\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1}-\boldsymbol{I}$, then the forth term in B.327 is bounded as

$$
\begin{align*}
\lambda \| \iota^{*} & \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}_{J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{\iota}_{J}^{*} \boldsymbol{P}_{J \backslash T} \boldsymbol{\sigma} \|_{2} \\
= & \lambda\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\left(\boldsymbol{P}_{[ \pm p] \backslash 0}+\boldsymbol{e}_{0} \boldsymbol{e}_{0}^{*}\right) \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{\iota}_{J}(\boldsymbol{I}+\boldsymbol{\Delta}) \boldsymbol{\iota}_{J}^{*} \boldsymbol{P}_{J \backslash T} \boldsymbol{\sigma}\right\|_{2} \\
\leq & \lambda\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota}\right\|_{2} \sqrt{2 p} \max _{\ell \neq 0}\left\|\boldsymbol{P}_{I \cap s_{\ell}[I \backslash T]} \boldsymbol{x}_{0}\right\|_{1}+\lambda\left\|\boldsymbol{a}_{0}\right\|_{2}\left\|\boldsymbol{P}_{I \backslash T} \boldsymbol{x}_{0}\right\|_{1} \\
& +\lambda\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota}\right\|_{2} \sqrt{2 p}\left\|\boldsymbol{P}_{I \cap s_{\ell}[I]} \boldsymbol{x}_{0}\right\|_{1}\|\boldsymbol{\Delta}\|_{\infty \rightarrow \infty}+\lambda\left\|\boldsymbol{a}_{0}\right\|_{2}\left\|\boldsymbol{x}_{0}\right\|_{2}\|\boldsymbol{\Delta}\|_{2} \sqrt{|J \backslash T|} \\
\leq & C_{4} \lambda\left(\sqrt{\widetilde{\mu} p^{2}} \times \widetilde{\lambda} n \theta^{2}+\lambda \widetilde{\lambda} n \theta+\sqrt{\widetilde{\mu} p^{2}} \times n \theta^{2} \times \widetilde{\mu} \kappa_{I}+\sqrt{n \theta} \times \widetilde{\mu} \kappa_{I} \sqrt{\widetilde{\lambda} n \theta}\right) \\
\leq & 2 C_{4}\left(\widetilde{\lambda}+\widetilde{\mu} \kappa_{I}\right) \lambda n \theta . \tag{B.330}
\end{align*}
$$

Therefore, combining B.328-B.330 we obtain

$$
\begin{equation*}
\left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}} \boldsymbol{\iota} \boldsymbol{a}_{0}-\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}} \boldsymbol{x}_{0}\right\|_{2} \leq C_{5}\left(\widetilde{\lambda}+\widetilde{\mu} \kappa_{I}\right) \lambda n \theta \tag{B.331}
\end{equation*}
$$

3. (Extract the set $J$ ) Lastly, we will further simplify the term with $\boldsymbol{a}-\boldsymbol{a}_{0}$ in B.331 by extracting the set $J$ :

$$
\begin{align*}
& \iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}} \boldsymbol{x}_{0} \\
& =\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J}(\boldsymbol{I}+\boldsymbol{\Delta}) \boldsymbol{C}_{\boldsymbol{a}_{0}+\left(\boldsymbol{a}-\boldsymbol{a}_{0}\right) J}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right) \\
& =\iota^{*} C_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{a_{0}} \boldsymbol{P}_{I} C_{a_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right) \\
& +\iota^{*} C_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J} \Delta C_{a_{0} J}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}_{0}-a\right)+\iota^{*} C_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J}\left(C_{a J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} C_{a-a_{0} J}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}_{0}-a\right) \\
& -\iota^{*} C_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I \backslash J} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right), \tag{B.332}
\end{align*}
$$

where, the latter terms in B.332 are bounded as

$$
\begin{align*}
& \left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J} \boldsymbol{\Delta} \boldsymbol{C}_{\boldsymbol{a}_{0} J}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\right\|_{2} \leq\left\|\boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\right\|_{2}^{2}\left\|\boldsymbol{C}_{\boldsymbol{a}_{0} J}\right\|_{2}^{2}\|\boldsymbol{\Delta}\|_{2} \leq C_{6} \widetilde{\mu} \kappa_{I} n \theta \\
& \left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{C}_{\boldsymbol{a}-\boldsymbol{a}_{0} J}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\right\|_{2} \leq\left\|\boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\right\|_{2}^{2}\left\|\boldsymbol{C}_{\boldsymbol{a}_{0} J}\right\|_{2}\left\|\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1}\right\|_{2}\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a} \boldsymbol{\iota}_{J}}\right\|_{2} \leq C_{7} \widetilde{\mu} \sqrt{\kappa_{I}} n \theta \\
& \left\|\boldsymbol{P}_{I \backslash J} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\right\|_{2}^{2} \leq|I \backslash J|\left\|\check{\boldsymbol{a}}_{0} * \boldsymbol{x}_{0}\right\|_{\square}^{2} \leq C_{8} \widetilde{\lambda} n \theta \times \kappa_{I} \leq C_{8}\left(\lambda \kappa_{I}+\frac{\kappa_{I} \log n}{\sqrt{n \theta^{2}}}\right) n \theta, \tag{B.333}
\end{align*}
$$

whence we conclude, that since $c_{\mu} \kappa_{I}^{2} \leq c_{\mu}$ and $\lambda \kappa_{I} \leq 5 c_{\mu}$, as long as $c_{\mu}<\frac{1}{100}\left(\frac{1}{C_{6}}+\frac{1}{C_{7}}+\frac{1}{5 C_{8}}\right)$ and $n>10^{6} C_{8}^{2} \theta^{-2} \kappa_{I}^{2} \log ^{2} n$, we gain:

$$
\begin{align*}
& \left\|\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0} J}\left(\boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a} J}\right)^{-1} \boldsymbol{C}_{\boldsymbol{a} J}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}} \boldsymbol{x}_{0}-\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right)\right\|_{2} \\
& \quad \leq\left(\frac{3}{100}+\frac{1}{1000}\right) n \theta\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2} \\
& \quad \leq \frac{1}{32} n \theta\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2} . \tag{B.334}
\end{align*}
$$

The claimed result therefore is followed by combining $\overline{B .326},(\overline{B .331}$ and $(\bar{B} .334)$.

Contraction of least square estimate of $\boldsymbol{a}$ toward $a_{0}$. The next thing is to show the operator

$$
\begin{equation*}
(n \theta)^{-1}\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\right) \tag{B.335}
\end{equation*}
$$

contracts $\boldsymbol{a}$ toward $\boldsymbol{a}_{0}$. We first will show that

$$
\begin{equation*}
(n \theta)^{-1}\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\right) \approx \boldsymbol{a}_{0} \boldsymbol{a}_{0}^{*} \tag{B.336}
\end{equation*}
$$

by seeing $\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota \approx(n \theta) \boldsymbol{e}_{0} \boldsymbol{e}_{0}^{*}$ 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

$$
\begin{equation*}
\left|\left\langle\boldsymbol{a}_{0}, \boldsymbol{a}-\boldsymbol{a}_{0}\right\rangle\right| \leq \frac{1}{2}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}^{2} \tag{B.337}
\end{equation*}
$$

Again, we introduce the following lemma to solidify our claim:

Lemma B.8.3 (Contraction of $\boldsymbol{a}$ to $\boldsymbol{a}_{0}$ ). Given $\boldsymbol{a}_{0} \in \mathbb{R}^{p_{0}}$ to be $\tilde{\mu}$-shift coherent and $\boldsymbol{x}_{0} \sim \operatorname{BG}(\theta) \in \mathbb{R}^{n}$. There exists some constants $C, C^{\prime}, c, c^{\prime}, c_{\mu}$ such that if $\lambda<c^{\prime} \widetilde{\mu} \kappa_{I}, \widetilde{\mu} \kappa_{I}^{2} \leq c_{\mu}$ and $n>C \theta^{-2} p^{2} \log p$, then with probability at least $1-c / n$, for every $\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu}$,

$$
\begin{equation*}
\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right)\right\|_{2} \leq \frac{1}{32}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} n \theta \tag{B.338}
\end{equation*}
$$

Proof. Since $\mathbb{E}\left\langle\boldsymbol{P}_{I} s_{i}\left[\boldsymbol{x}_{0}\right], s_{j}\left[\boldsymbol{x}_{0}\right]\right\rangle=0$ for all $i \neq j$ and set $I$, we calculate

$$
\begin{align*}
\mathbb{E}\left[\boldsymbol{\iota}_{[ \pm p]}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}_{[ \pm p]}\right] & =\sum_{i \in[ \pm p]} \mathbb{E}\left[\boldsymbol{e}_{i}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{e}_{i}\right] \boldsymbol{e}_{i} \boldsymbol{e}_{i}^{*}=\mathbb{E}\left\|\boldsymbol{x}_{0}\right\|_{2}^{2} \boldsymbol{e}_{0} \boldsymbol{e}_{0}^{*}+\sum_{i \in[ \pm p] \backslash 0} \mathbb{E}\left\|\boldsymbol{P}_{I} s_{i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2} \boldsymbol{e}_{i} \boldsymbol{e}_{i}^{*} \\
& =n \theta \boldsymbol{e}_{0} \boldsymbol{e}_{0}^{*}+n \theta^{2} \boldsymbol{P}_{[ \pm p] \backslash 0}=n \theta^{2} \boldsymbol{I}+n \theta(1-\theta) \boldsymbol{e}_{0} \boldsymbol{e}_{0}^{*} \tag{B.339}
\end{align*}
$$

whence

$$
\begin{equation*}
\mathbb{E}\left[\iota^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\right]=\iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \mathbb{E}\left[\boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}}\right] \boldsymbol{C}_{\boldsymbol{a}_{0}} \iota=n \theta^{2} \iota^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \iota+n \theta(1-\theta) \boldsymbol{a}_{0} \boldsymbol{a}_{0}^{*} \tag{B.340}
\end{equation*}
$$

implying the expectation is a contraction mapping for $\boldsymbol{a}_{0}-\boldsymbol{a}$ when $c_{\mu}<\frac{1}{200}$ :

$$
\begin{align*}
\left\|\mathbb{E}\left[\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\right]\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right)\right\|_{2} & \leq n \theta^{2}\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}\right\|_{2}\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2}+n \theta\left\|\boldsymbol{a}_{0}\right\|_{2}\left|\left\langle\boldsymbol{a}_{0}, \boldsymbol{a}_{0}-\boldsymbol{a}\right\rangle\right| \\
& \leq n \theta^{2} \times 2 \widetilde{\mu} p \times\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2}+\frac{1}{2} n \theta\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2}^{2} \\
& \leq\left(2 c_{\mu}+\frac{1}{2} c_{\mu}\right)\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2} n \theta \\
& \leq \frac{1}{64}\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2} n \theta \tag{B.341}
\end{align*}
$$

For each entry of $\boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}}$, again from Appendix B.1 we know with probability at least $1-c / n$ :

$$
\left|\boldsymbol{e}_{i}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{e}_{j}-\mathbb{E}\left[\boldsymbol{e}_{i}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{e}_{j}\right]\right| \leq \begin{cases}C^{\prime} \sqrt{n \theta \log n} & i=j=0 \\ C^{\prime} \sqrt{n \theta^{2} \log n} & \text { otherwise }\end{cases}
$$

Thus via Gershgorin disc theorem, when $n>10^{3} C^{2} \theta^{-2} p^{2} \log n$ :

$$
\begin{equation*}
\lambda_{\max }\left(\boldsymbol{\iota}_{[ \pm p]}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}_{[ \pm p]}-\mathbb{E}\left[\boldsymbol{\iota}_{[ \pm p]}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}_{[ \pm p]}\right]\right) \leq C^{\prime} p \sqrt{n \theta^{2} \log n} \leq \frac{1}{64} n \theta^{2} \tag{B.342}
\end{equation*}
$$

Finally we combine $\overline{B .341}, \overline{B .342}$ and get

$$
\begin{equation*}
\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{P}_{I} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right)\right\|_{2} \leq\left(\frac{1}{64} n \theta+\frac{1}{64} n \theta^{2}\left\|\boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{\iota}_{ \pm p}\right\|_{2}^{2}\right)\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2} \leq \frac{1}{32}\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2} n \theta \tag{B.343}
\end{equation*}
$$

Theorem B.8.1 B.8.3 together implies the single iterate of alternating minimization contracts $\boldsymbol{a}$ toward $\boldsymbol{a}_{0}$. We show it with the following lemma:

Lemma B.8.4 (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 $C, C^{\prime}, c, c_{\mu}$ 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

$$
\begin{equation*}
5 \widetilde{\mu} \kappa_{I} \geq \lambda \geq 5 \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.344}
\end{equation*}
$$

and suppose $\boldsymbol{x}^{+}$has the form of (B.272), then the solution $\boldsymbol{a}^{+}$to

$$
\begin{equation*}
\min _{\boldsymbol{a}^{\prime} \in \mathbb{R}^{p}}\left\{\left\|\boldsymbol{a}^{\prime} * \boldsymbol{x}^{+}-\boldsymbol{y}\right\|_{2}^{2}\right\} \tag{B.345}
\end{equation*}
$$

is unique and satisfies

$$
\begin{equation*}
\left\|\boldsymbol{P}_{\mathbb{S}^{p-1}}\left[\boldsymbol{a}^{+}\right]-\boldsymbol{a}_{0}\right\|_{2} \leq \frac{1}{2}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.346}
\end{equation*}
$$

Proof. Write $\boldsymbol{x}$ as $\boldsymbol{x}^{+}$, then

$$
\begin{align*}
\lambda_{p}\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{\iota}\right) & =\sigma_{\min }^{2}\left(\boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}+\boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}} \boldsymbol{\iota}\right) \\
& \geq\left[\sigma_{\min }\left(\boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\right)-\left\|\boldsymbol{C}_{\boldsymbol{x}-\boldsymbol{x}_{0}} \boldsymbol{\iota}\right\|\right]_{+}^{2} \\
& \geq\left[\sigma_{\min }\left(\boldsymbol{C}_{\boldsymbol{x}_{0}} \boldsymbol{\iota}\right)-2 \sqrt{\kappa_{I}}\left\|\boldsymbol{x}-\boldsymbol{x}_{0}\right\|_{2}\right]_{+}^{2} \\
& \geq\left[\frac{2}{3} \sqrt{\theta n}-8 \lambda \sqrt{\kappa_{I}} \sqrt{\theta n}\right]_{+}^{2} \\
& \geq \frac{1}{2} \theta n, \tag{B.347}
\end{align*}
$$

where the fourth inequality is derived from using the upper bound of sparse convolution matrix from Theorem B.1.6, and the last line holds by knowing $\lambda<5 c_{\mu} \kappa_{I}^{-1}$. From B.347) we know the least square problem of B.345 has unique solution $\boldsymbol{a}^{+}$, written as

$$
\begin{equation*}
\boldsymbol{a}^{+}=\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}} \iota\right)^{-1} \iota \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{y} \tag{B.348}
\end{equation*}
$$

whence

$$
\begin{equation*}
\boldsymbol{a}^{+}-\boldsymbol{a}_{0}=\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}} \iota\right)^{-1}\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}} \iota\right) \boldsymbol{a}_{0}-\boldsymbol{a}_{0}=\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}} \iota\right)^{-1}\left(\iota^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}-\boldsymbol{x}} \iota\right) \boldsymbol{a}_{0} \tag{B.349}
\end{equation*}
$$

Combine Theorem B.8.2 and Theorem B.8.3, we know

$$
\begin{equation*}
\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}-\boldsymbol{x}} \boldsymbol{\iota}\right\|_{2} \leq\left(C_{1} \lambda\left(\widetilde{\lambda}+\widetilde{\mu} \kappa_{I}\right)+\frac{1}{16}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}\right) n \theta \tag{B.350}
\end{equation*}
$$

for some constant $C_{1}$. Combine B.347, (B.349, B.350) and since $\lambda<\tilde{\mu} \kappa_{I}$, by letting $c_{\mu}<\frac{1}{4 C_{1}}$, we gain

$$
\begin{equation*}
\left\|\boldsymbol{a}^{+}-\boldsymbol{a}_{0}\right\|_{2} \leq \frac{\left\|\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}_{0}-\boldsymbol{x}} \boldsymbol{\iota}\right\|_{2}}{\lambda_{p}\left(\boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{x}}^{*} \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{\iota}\right)} \leq 2 C_{1} \lambda\left(\widetilde{\lambda}+\widetilde{\mu} \kappa_{I}\right)+\frac{1}{8}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \frac{1}{4} \tag{B.351}
\end{equation*}
$$

For the final bound,

$$
\begin{align*}
\left\|\frac{\boldsymbol{a}^{+}}{\left\|\boldsymbol{a}^{+}\right\|_{2}}-\boldsymbol{a}_{0}\right\|_{2} & \leq \frac{\left\|\boldsymbol{a}^{+}-\boldsymbol{a}_{0}\right\|_{2}+\left|\left\|\boldsymbol{a}^{+}\right\|_{2}-1\right|}{\left\|\boldsymbol{a}^{+}\right\|_{2}} \leq \frac{2\left\|\boldsymbol{a}^{+}-\boldsymbol{a}_{0}\right\|_{2}}{1-\left\|\boldsymbol{a}^{+}-\boldsymbol{a}_{0}\right\|_{2}} \leq \frac{8}{3}\left\|\boldsymbol{a}^{+}-\boldsymbol{a}_{0}\right\|_{2} \\
& \leq C_{2} \lambda\left(\widetilde{\lambda}+\widetilde{\mu} \kappa_{I}\right)+\frac{1}{3}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.352}
\end{align*}
$$

and since $\lambda>\kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}$, finally we gain

$$
\begin{align*}
\text { B.352 } & \leq C_{2}\left(\lambda \kappa_{I}+\frac{p \kappa_{I} \log n}{n \theta}+\widetilde{\mu} \kappa_{I}^{2}\right)\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}+\frac{1}{3}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \\
& \leq \frac{1}{2}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.353}
\end{align*}
$$

as long as $n>20 C_{2} \theta^{-1} p \kappa_{I} \log n$ and $c_{\mu}<\frac{1}{20 C_{2}}$.

## B.8.3 Linear convergence of alternating minimization (Proof of Theorem 3.5.2)

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 $I^{(0)}=\left\{i:\left|s_{i}\left[\boldsymbol{a}^{(0)}\right]^{*} \boldsymbol{\iota}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}_{0}\right|>\lambda^{(0)}\right\}$. Then for all $\left|\boldsymbol{x}_{j}\right|>6 \lambda^{(0)}$, we have

$$
\begin{align*}
\left|s_{j}\left[\boldsymbol{a}^{(0)}\right]^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}} \boldsymbol{x}_{0}\right| & \geq\left(1-\left|\left\langle\boldsymbol{a}^{(0)} \boldsymbol{a}_{0}\right\rangle\right|\right)\left|\boldsymbol{x}_{j}\right|-\left\|\boldsymbol{P}_{[ \pm p] \backslash\{j\}} \boldsymbol{C}_{\boldsymbol{a}_{0}}^{*} \boldsymbol{\iota} s_{j}\left[\boldsymbol{a}^{(0)}\right]\right\|_{2} \times \sqrt{2}\left\|\boldsymbol{x}_{0}\right\|_{\square} \\
& \geq(1-2 \widetilde{\mu}) 6 \lambda^{(0)}-2 \widetilde{\mu} \sqrt{\kappa_{I}} \times \sqrt{2 \kappa_{I}} \\
& \geq 5 \lambda^{(0)}-4 \lambda^{(0)} \\
& =\lambda^{(0)} . \tag{B.354}
\end{align*}
$$

hence $I_{>6 \lambda^{(0)}} \subseteq I^{(0)}$, therefore the condition of Theorem B.8.4 is satisfied, implies 3.89 holds for $k=0$.
Suppose it is true for $1,2, \ldots, k-1$, such that

$$
\begin{equation*}
\kappa_{I}\left\|\boldsymbol{a}^{(k)}-\boldsymbol{a}_{0}\right\|_{2} \leq \frac{1}{2} \lambda^{(k-1)}=\lambda^{(k)}, \quad \text { and } \quad I_{>3 \lambda^{(k-1)}} \subseteq I^{(k)} \tag{B.355}
\end{equation*}
$$

and since $I_{>6 \lambda^{(k)}}=I_{>3 \lambda^{(k-1)}} \subseteq I^{(k)}$, we can again apply Theorem B.8.4 resulting

$$
\begin{equation*}
\kappa_{I}\left\|\boldsymbol{a}^{(k+1)}-\boldsymbol{a}\right\|_{2} \leq \frac{1}{2} \kappa_{I}\left\|\boldsymbol{a}^{(k)}-\boldsymbol{a}_{0}\right\|_{2} \leq \frac{1}{2} \lambda^{(k)} \tag{B.356}
\end{equation*}
$$

as claimed.

## B.8.4 Supporting lemmas for refinement

The following lemma controls the shift coherence of $\boldsymbol{a}$ :
Lemma B.8.5 (Coherence of $\boldsymbol{a}$ near $\boldsymbol{a}_{0}$ ). Suppose that $\boldsymbol{a}_{0}$ is $\widetilde{\mu}$-shift coherent, and $\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu}$. Then

$$
\begin{align*}
\| \text { off }\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{a_{0}}\right] \|_{\infty} & \leq 2 \widetilde{\mu}  \tag{B.357}\\
\| \text { off }\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right] \|_{\infty} & \leq 3 \widetilde{\mu} \tag{B.358}
\end{align*}
$$

Proof. Notice that for any $\ell \neq 0,\left|\left\langle\boldsymbol{a}, s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\rangle\right| \leq\left|\left\langle\boldsymbol{a}_{0}, s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\rangle\right|+\left|\left\langle\boldsymbol{a}-\boldsymbol{a}_{0}, s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\rangle\right| \leq \widetilde{\mu}+\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2} \leq 2 \widetilde{\mu}$. Similarly, $\left|\left\langle\boldsymbol{a}, s_{\ell}[\boldsymbol{a}]\right\rangle\right| \leq\left|\left\langle\boldsymbol{a}-\boldsymbol{a}_{0}, s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\rangle\right|+\left|\left\langle\boldsymbol{a}, s_{\ell}\left[\boldsymbol{a}_{0}\right]\right\rangle\right| \leq\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}+2 \widetilde{\mu} \leq 3 \widetilde{\mu}$, as claimed.

From this we obtain the following spectral control on $C_{a}^{*} C_{a}$, to simply the notations, we will write

$$
\begin{equation*}
C_{a I}^{*} C_{a I}=\iota_{I}^{*} C_{a}^{*} C_{a} \iota_{I}=\left[C_{a}^{*} C_{a}\right]_{I, I} \tag{B.359}
\end{equation*}
$$

in the latter part of this section.
Lemma B.8.6 (Off-diagonals of $\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right]_{I, I}$ ). Suppose that $\boldsymbol{a}_{0}$ is $\widetilde{\mu}$-shift coherent and $\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu}$. Then

$$
\begin{equation*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C} \boldsymbol{a}-\boldsymbol{I}\right]_{I, I}\right\|_{2} \leq 9 \kappa_{I} \widetilde{\mu} \tag{B.360}
\end{equation*}
$$

We prove this lemma by noting that $C_{a}^{*} C_{a}=C_{r_{a, a}}$ is the convolution matrix associated with the autocorrelation $\boldsymbol{r}_{\boldsymbol{a}, \boldsymbol{a}}$ of $\boldsymbol{a}$. Since $\operatorname{supp}\left(\boldsymbol{r}_{\boldsymbol{a}, \boldsymbol{a}}\right) \subseteq\{-p+1, \ldots, p-1\}$ is confined to a (cyclic) stripe of width $2 p-1$, 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 $\ell$ :

$$
\operatorname{supp}\left(\left[C_{a}^{*} C_{a}\right]_{I_{\ell}, I}\right) \subseteq I_{\ell} \times\left(I_{\ell-1} \uplus I_{\ell} \uplus I_{\ell+1}\right),
$$

where $\ell+1$ and $\ell-1$ are interpreted cyclically modulo $r$.

For an arbitrary $\boldsymbol{v} \in \mathbb{R}^{|I|}$, we calculate

$$
\begin{align*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{I, I} \boldsymbol{v}\right\|_{2}^{2} & =\sum_{\ell=0}^{r-1}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{I_{\ell}, I} \boldsymbol{v}\right\|_{2}^{2}  \tag{B.361}\\
& =\sum_{\ell=0}^{r-1}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{I_{\ell}, I_{\ell-1} \uplus I_{\ell} \uplus I_{\ell+1}} \boldsymbol{v}_{I_{\ell-1} \uplus I_{\ell} \uplus I_{\ell+1}}\right\|_{2}^{2}  \tag{B.362}\\
& \leq \sum_{\ell=0}^{r-1}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{I_{\ell}, I_{\ell-1} \uplus I_{\ell} \uplus I_{\ell+1}}\right\|_{F}^{2}\left\|\boldsymbol{v}_{I_{\ell-1} \uplus I_{\ell} \uplus I_{\ell+1}}\right\|_{2}^{2}  \tag{B.363}\\
& \leq 3 \kappa_{I}^{2} \times(3 \widetilde{\mu})^{2} \times \sum_{\ell=0}^{r-1}\left\|\boldsymbol{v}_{I_{\ell-1} \uplus I_{\ell} \uplus I_{\ell+1}}\right\|_{2}^{2}  \tag{B.364}\\
& \leq 3 \kappa_{I}^{2} \times 9 \widetilde{\mu}^{2} \times 3\|\boldsymbol{v}\|_{2}^{2} \tag{B.365}
\end{align*}
$$

giving the claimed result.

As a consequence, we have that
Corollary B.8.7 (Inverse of $\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right]_{J, J}$ ). Suppose that $\boldsymbol{a}_{0}$ is $\mu$-shift coherent, that $\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu}$ and that $\kappa_{I} \widetilde{\mu}<\frac{1}{18}$. Then for every $J \subseteq I$ and any norm $\|\cdot\|_{\diamond} \in\left\{\|\cdot\|_{\square \rightarrow \square},\|\cdot\|_{\infty \rightarrow \infty},\|\cdot\|_{2}\right\}$, we have

$$
\begin{align*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{J, J}\right\|_{\diamond} & \leq 9 \kappa_{I} \widetilde{\mu}  \tag{B.366}\\
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right]_{J, J}^{-1}-\boldsymbol{I}\right\|_{\diamond} & \leq 18 \kappa_{I} \widetilde{\mu}  \tag{B.367}\\
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right]_{J, J}^{-1}\right\|_{\diamond} & \leq 2 \tag{B.368}
\end{align*}
$$

Proof. First we prove

$$
\begin{equation*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{J, J}\right\|_{2} \leq 9 \kappa_{I} \widetilde{\mu}, \quad\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{J, J}\right\|_{\infty \rightarrow \infty} \leq 6 \kappa_{I} \widetilde{\mu}, \quad\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{J, J}\right\|_{\square \rightarrow \square} \leq 6 \kappa_{I} \tilde{\mu} \tag{B.369}
\end{equation*}
$$

Where the first claim follows from Theorem B.8.6. 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

$$
\begin{align*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{J, J}\right\|_{\square \rightarrow \square} & \leq \max _{\ell, \ell^{\prime}}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}-\boldsymbol{I}\right]_{J \cap([p]+\ell), J \cap\left([2 p]+\ell^{\prime}\right)}\right\|_{F} \\
& \leq 6 \kappa_{I} \widetilde{\mu} \tag{B.370}
\end{align*}
$$

Then we prove

$$
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right]_{J, J}^{-1}-\boldsymbol{I}\right\|_{2} \leq 18 \kappa_{I} \widetilde{\mu}, \quad\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right]_{J, J}^{-1}-\boldsymbol{I}\right\|_{\infty \rightarrow \infty} \leq 12 \kappa_{I} \widetilde{\mu}, \quad\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}}\right]_{J, J}^{-1}-\boldsymbol{I}\right\|_{\square \rightarrow \square} \leq 12 \kappa_{I} \widetilde{\mu}, \quad \text { (B.371) }
$$

which are followed from the fact that if $\|\cdot\|_{\diamond}$ is a matrix norm and $\|\Delta\|_{\diamond}<1$, then

$$
\left\|(\boldsymbol{I}+\boldsymbol{\Delta})^{-1}-\boldsymbol{I}\right\|_{\diamond} \leq \frac{\|\boldsymbol{\Delta}\|_{\diamond}}{1-\|\boldsymbol{\Delta}\|_{\diamond}}
$$

Finally, B.368 follows from the triangle inequality.

Also, we need to bound the convolution of $\boldsymbol{a}_{0}-\boldsymbol{a}$ with $\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2}$ requiring for bounds of the lasso solution:

Lemma B.8.8 (Convolution of $\boldsymbol{a}_{0}-\boldsymbol{a}$ ). Suppose that $\boldsymbol{a}_{0}$ is $\mu$-shift coherent and $\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu}$, then for every $J \subseteq I$,

$$
\begin{align*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}}\right]_{J, J}\right\|_{\square \rightarrow \infty} & \leq \sqrt{2 \kappa_{I}}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}  \tag{B.372}\\
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}}\right]_{J, J}\right\|_{\square \rightarrow \square} & \leq \sqrt{2} \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.373}
\end{align*}
$$

Proof. For the first inequality, we have

$$
\begin{align*}
\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\left.\boldsymbol{a}_{0}-\boldsymbol{a}\right]_{J, J} \boldsymbol{v} \|_{\square \rightarrow \infty}}\right. & =\max _{j \in J,\|\boldsymbol{v}\|_{\square}=1}\left|\left\langle s_{j}[\boldsymbol{a}],\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right) * \boldsymbol{v}\right\rangle\right| \\
& \leq \max _{j \in[n],\|\boldsymbol{v}\|_{\square}=1}\left\|\boldsymbol{P}_{[p]+j}\left[\left(\boldsymbol{a}_{0}-\boldsymbol{a}\right) * \boldsymbol{v}\right]\right\|_{2} \\
& \leq\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \times \max _{j \in[n],\|\boldsymbol{v}\|_{\square}=1}\left\|\boldsymbol{P}_{[ \pm p]+j} \boldsymbol{v}\right\|_{1} \\
& \leq \sqrt{2 \kappa_{I}}\left\|\boldsymbol{a}_{0}-\boldsymbol{a}\right\|_{2} \tag{B.374}
\end{align*}
$$

The second inequality is derived by

$$
\begin{align*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}}\right]_{J, J}\right\|_{\square \rightarrow \square} & \leq \max _{\ell, \ell^{\prime}}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}-\boldsymbol{a}}\right]_{J \cap([p]+\ell), J \cap\left([2 p]+\ell^{\prime}\right)}\right\|_{F} \\
& \leq \sqrt{2 \kappa_{I}^{2} \max _{i, j}\left|\left\langle s_{i}[\boldsymbol{a}], s_{j}\left[\boldsymbol{a}_{0}-\boldsymbol{a}\right]\right\rangle\right|^{2}} \\
& \leq \sqrt{2} \kappa_{I}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \tag{B.375}
\end{align*}
$$

finishing the proof.

Again, using a variant of the argument for Theorem B.8.6, we have the following:

Lemma B.8.9 (Off-diagonal of submatrix of $\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}$ ). Suppose that $\boldsymbol{a}_{0}$ is $\mu$-shift coherent and $\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2} \leq \widetilde{\mu}$. For any $J \subset I$, if

$$
\begin{align*}
\kappa_{J} & =\max _{\ell}|J \cap\{\ell, \ell+1, \ldots, \ell+p-1\}|  \tag{B.376}\\
\kappa_{I \backslash J} & =\max _{\ell}|(I \backslash J) \cap\{\ell, \ell+1, \ldots, \ell+p-1\}| \tag{B.377}
\end{align*}
$$

Then

$$
\begin{equation*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right]_{J, I \backslash J}\right\|_{2} \leq 6 \sqrt{\kappa_{J} \kappa_{I \backslash J}} \widetilde{\mu} . \tag{B.378}
\end{equation*}
$$

Proof. Take $r=\lceil n / p\rceil$ and for $\ell=0, \ldots, r-1$, write

$$
J_{\ell}=J \cap([p]+p \ell), \quad L_{\ell}=(I \backslash J) \cap([p]+p \ell),
$$

Take $\boldsymbol{v} \in \mathbb{R}^{|I \backslash J|}$ arbitrary and notice that

$$
\begin{align*}
\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right]_{J, I \backslash J} \boldsymbol{v}\right\|_{2}^{2} & =\sum_{\ell=0}^{r-1}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right]_{J_{\ell}, I \backslash J} \boldsymbol{v}\right\|_{2}^{2} \\
& =\sum_{\ell=0}^{r-1}\left\|\left[\boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}_{0}}\right]_{J_{\ell}, L_{\ell-1} \cup L_{\ell} \cup L_{\ell+1}} \boldsymbol{v}_{L_{\ell-1} \cup L_{\ell} \cup L_{\ell+1}}\right\|_{2}^{2} \\
& \leq 4 \widetilde{\mu}^{2} \times \kappa_{J} \times 3 \kappa_{I \backslash J} \times \sum_{\ell=0}^{r-1}\left\|\boldsymbol{v}_{L_{\ell-1} \cup L_{\ell} \cup L_{\ell+1}}\right\|_{2}^{2} \\
& \leq 4 \widetilde{\mu}^{2} \times \kappa_{J} \times 3 \kappa_{I \backslash J} \times 3\|\boldsymbol{v}\|_{2}^{2} \tag{B.379}
\end{align*}
$$

giving the result.

Lemma B.8.10 (Perturbation of vector over sphere). If both $\boldsymbol{a}, \boldsymbol{a}_{0}$ are unit vectors in inner product space, then

$$
\begin{equation*}
\left|\left\langle\boldsymbol{a}, \boldsymbol{a}-\boldsymbol{a}_{0}\right\rangle\right| \leq \frac{1}{2}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}^{2} \tag{B.380}
\end{equation*}
$$

Proof. Via simple norm inequalities:

$$
\begin{equation*}
\frac{1}{2}\left\|\boldsymbol{a}-\boldsymbol{a}_{0}\right\|_{2}^{2}=1-\left\langle\boldsymbol{a}, \boldsymbol{a}_{0}\right\rangle=1-\left\langle\boldsymbol{a}, \boldsymbol{a}_{0}-\boldsymbol{a}+\boldsymbol{a}\right\rangle=\left\langle\boldsymbol{a}, \boldsymbol{a}-\boldsymbol{a}_{0}\right\rangle>0 \tag{B.381}
\end{equation*}
$$

Lemma B.8.11 (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

$$
\begin{equation*}
\max _{\ell \in[n]}|I \cap([p]+\ell)| \leq \kappa \tag{B.382}
\end{equation*}
$$

then

$$
\begin{equation*}
\|\boldsymbol{\delta} * \boldsymbol{v}\|_{2} \leq \sqrt{2 \kappa}\|\boldsymbol{\delta}\|_{2}\|\boldsymbol{v}\|_{2} \tag{B.383}
\end{equation*}
$$

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:

$$
\begin{align*}
I_{1} & =\left\{i_{1}, i_{\kappa+1}, i_{2 \kappa+1}, \ldots\right\} \cap\{0, \ldots, n-p-1\} \\
I_{2} & =\left\{i_{2}, i_{\kappa+2}, i_{2 \kappa+2}, \ldots\right\} \cap\{0, \ldots, n-p-1\} \\
& \vdots \\
I_{\kappa} & =\left\{i_{\kappa}, i_{2 \kappa}, i_{3 \kappa}, \ldots\right\} \cap\{0, \ldots, n-p-1\}  \tag{B.384}\\
R & =I \cap\{n-p, \ldots, n-1\} . \tag{B.385}
\end{align*}
$$

Then the p-separating property gives $\left\|\boldsymbol{\delta} * \boldsymbol{P}_{I_{i}} \boldsymbol{v}\right\|_{2}=\|\boldsymbol{\delta}\|_{2}\left\|\boldsymbol{P}_{I_{i}} \boldsymbol{v}\right\|_{2}$. Hence:

$$
\begin{align*}
\left\|\boldsymbol{\delta} * \boldsymbol{P}_{I} \boldsymbol{v}\right\|_{2} & =\left\|\sum_{i \in \kappa} \boldsymbol{\delta} * \boldsymbol{P}_{I_{i}} \boldsymbol{v}+\boldsymbol{\delta} * \boldsymbol{P}_{R} \boldsymbol{v}\right\|_{2} \leq \sum_{i \in \kappa}\left\|\boldsymbol{\delta} * \boldsymbol{P}_{I_{i}} \boldsymbol{v}\right\|_{2}+\left\|\boldsymbol{\delta} * \boldsymbol{P}_{R} \boldsymbol{v}\right\| \\
& =\|\boldsymbol{\delta}\|_{2} \sum_{i \in \kappa}\left\|\boldsymbol{v}_{I_{i}}\right\|_{2}+\|\boldsymbol{\delta}\|_{2}\left\|\boldsymbol{P}_{R} \boldsymbol{v}\right\|_{1} \\
& \leq \sqrt{\kappa}\left\|\boldsymbol{v}_{I_{1}, \uplus, \ldots, \uplus I_{\kappa}}\right\|_{2}\|\boldsymbol{\delta}\|_{2}+\sqrt{\kappa}\left\|\boldsymbol{v}_{R}\right\|_{2}\|\boldsymbol{\delta}\|_{2} \\
& \leq \sqrt{2 \kappa}\|\boldsymbol{v}\|_{2}\|\boldsymbol{\delta}\|_{2} \tag{B.386}
\end{align*}
$$

where the last two inequalities were coming from Cauchy-Schwartz.

## B. 9 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 $\left\|\boldsymbol{P}_{I(\boldsymbol{a})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}$, and the set of gradients discontinuity entries $\left|J_{B}(\boldsymbol{a})\right|$.

## B.9.1 Proof of Theorem B.3.4

Proof. 1. ( $\varepsilon$-net) Write $\boldsymbol{x}$ as $\boldsymbol{x}_{0}$ and $\|\boldsymbol{\beta}\|_{2}=\eta$ through out this proof, firstly from Theorem B.2.1 for every $\boldsymbol{a} \in \cup_{|\boldsymbol{\tau}| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$, we know $\eta \leq 1+c_{\mu}+\frac{c_{\mu}}{\sqrt{\theta} k \log \theta^{-1}} \leq \sqrt{p}$. Define $\varepsilon=\frac{c_{2}}{2 n^{3 / 2} p^{3 / 2}}$ and consider the $\varepsilon$-net
$\mathcal{N}_{\varepsilon}$ for sphere of radius $\eta$. From Theorem B.10.5 we know for any $c_{2}<1$ :

$$
\begin{equation*}
\left|\mathcal{N}_{\varepsilon}\right| \leq\left(\frac{3 \eta}{\varepsilon}\right)^{2 p} \leq\left(\frac{3 n^{3 / 2} p^{2}}{c_{2}}\right)^{2 p} \leq\left(\frac{3 n p^{2}}{c_{2}}\right)^{3 p} \tag{B.387}
\end{equation*}
$$

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:

$$
\begin{equation*}
\mathcal{E}_{\mathrm{Net}}:=\left\{\forall i \in[n], \quad \boldsymbol{\sigma}_{i} n^{-1} \boldsymbol{\chi}\left[\boldsymbol{\beta}_{\varepsilon}\right]_{i}-\boldsymbol{\sigma}_{i} n^{-1} \overline{\mathbb{E} \boldsymbol{\chi}\left[\boldsymbol{\beta}_{\varepsilon}\right]_{i}}<\frac{c_{1} \theta}{p^{3 / 2}} \quad \forall \boldsymbol{\beta}_{\varepsilon} \in \mathcal{N}_{\varepsilon, i}\right\} \tag{B.388}
\end{equation*}
$$

2. (Lipschitz constant) The Lipschitz constant $L$ of $\boldsymbol{\chi}[]_{i}$ w.r.t $\boldsymbol{\beta}$ is bounded in terms of $\boldsymbol{x}$ regardless of entry $i$ :

$$
\begin{align*}
\left|\boldsymbol{\chi}[\boldsymbol{\beta}]_{i}-\boldsymbol{\chi}\left[\boldsymbol{\beta}^{\prime}\right]_{i}\right| & \leq\left|\boldsymbol{e}_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\beta}\right]-\boldsymbol{e}_{i}^{*} \breve{\boldsymbol{C}}_{\boldsymbol{x}} \mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\beta}^{\prime}\right]\right| \leq\|\boldsymbol{x}\|_{2}\left\|\mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\beta}\right]-\mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\beta}^{\prime}\right]\right\|_{2} \\
& \leq\|\boldsymbol{x}\|_{2} \sqrt{\sum_{j \in[n]}\left|\mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\beta}\right]_{j}-\mathcal{S}_{\lambda}\left[\breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\beta}^{\prime}\right]_{j}\right|^{2}} \leq\|\boldsymbol{x}\|_{2}\left\|\breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\beta}-\breve{\boldsymbol{C}}_{\boldsymbol{x}} \boldsymbol{\beta}^{\prime}\right\|_{2} \\
& \leq\|\boldsymbol{x}\|_{2} \cdot\|\boldsymbol{x}\|_{1} \cdot\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}\right\|_{2}=: L\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}\right\|_{2} \tag{B.389}
\end{align*}
$$

Define the event that $\boldsymbol{\chi}[\boldsymbol{\beta}]_{i}$ that has small Lipschitz constant as

$$
\begin{equation*}
\mathcal{E}_{\text {Lip }}:=\left\{L<2 n^{3 / 2} \theta\right\} \tag{B.390}
\end{equation*}
$$

on the event $\mathcal{E}_{\text {Lip }}$, for every points in $\mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ and $i \in[n]$, there exists some $\boldsymbol{\beta}_{\varepsilon} \in \mathcal{N}_{\varepsilon, i}$ such that

$$
\begin{equation*}
\left|\left(\boldsymbol{\sigma}_{i} n^{-1} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}-\boldsymbol{\sigma}_{i} n^{-1} \overline{\mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}}\right)-\left(\boldsymbol{\sigma}_{i} n^{-1} \boldsymbol{\chi}\left[\boldsymbol{\beta}_{\varepsilon}\right]_{i}-\boldsymbol{\sigma}_{i} n^{-1} \overline{\mathbb{E} \boldsymbol{\chi}\left[\boldsymbol{\beta}_{\varepsilon}\right]_{i}}\right)\right| \leq 2 L \varepsilon \leq \frac{c_{2} \theta}{p^{3 / 2}} \tag{B.391}
\end{equation*}
$$

On event $\mathcal{E}_{\text {Lip }} \cap \mathcal{E}_{\text {Net }},(\overline{B .388}), \overline{B .391}$ implies $\boldsymbol{\chi}[\boldsymbol{\beta}]$ is well concentrated entrywise and anywhere in $\cup_{|\tau| \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ :

$$
\begin{equation*}
\left|\boldsymbol{\sigma}_{i} n^{-1} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}-\boldsymbol{\sigma}_{i} n^{-1} \overline{\mathbb{E} \boldsymbol{\chi}[\boldsymbol{\beta}]_{i}}\right| \leq \frac{\left(c_{1}+c_{2}\right) \theta}{p^{3 / 2}}, \quad \forall \boldsymbol{a} \in \cup_{k \leq k} \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right), \forall i \in[n] \tag{B.392}
\end{equation*}
$$

as desired, where, using Theorem B.1.2

$$
\begin{equation*}
\mathbb{P}\left[\mathcal{E}_{\text {Lip }}^{c}\right] \leq \mathbb{P}\left[\|x\|_{2}^{2}>2 n \theta\right] \leq 1 / n \tag{B.393}
\end{equation*}
$$

and using union bound,

$$
\begin{align*}
\mathbb{P}\left[\mathcal{E}_{\mathrm{Net}}^{c}\right] & \leq \mathbb{P}\left[\max _{\substack{\boldsymbol{a}_{\varepsilon} \in \mathcal{N}_{\varepsilon}, i \\
i \in[n]}} \boldsymbol{\sigma}_{i} n^{-1} \boldsymbol{\chi}\left[\boldsymbol{\beta}_{\varepsilon}\right]_{i}-\boldsymbol{\sigma}_{i} n^{-1} \overline{\mathbb{E} \boldsymbol{\chi}\left[\boldsymbol{\beta}_{\varepsilon}\right]_{i}}>\frac{c_{1} \theta}{p^{3 / 2}}\right] \\
& \leq n\left|\mathcal{N}_{\varepsilon}\right| \mathbb{P}\left[\boldsymbol{\sigma}_{0} n^{-1} \boldsymbol{\chi}\left[\boldsymbol{\beta}_{\varepsilon}\right]_{0}-\boldsymbol{\sigma}_{0} n^{-1} \mathbb{E} \boldsymbol{\chi}\left[\boldsymbol{\beta}_{\varepsilon}\right]_{0}>\frac{c_{1} \theta}{p^{3 / 2}}\right] . \tag{B.394}
\end{align*}
$$

3. (Bound $\left.\mathbb{P}\left[\mathcal{E}_{\text {Net }}^{c}\right]\right)$ Wlog write $n=t \cdot(2 p)$ for some integer $t$ and $2 p \geq 4 p_{0}-3$ and replace $\boldsymbol{x}_{0}$ with $\boldsymbol{x}$. Observe that $\boldsymbol{Z}_{j}(\boldsymbol{\beta})$ from $(\overline{\mathrm{B} .61})$ is independent of $\boldsymbol{Z}_{j+2 p}(\boldsymbol{\beta})$ 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

$$
\boldsymbol{\chi}[\boldsymbol{\beta}]_{0}=\sum_{j \in[n]} \boldsymbol{Z}_{j}(\boldsymbol{\beta})=\sum_{k \in[2 p]}\left(\sum_{t=0}^{n / 2 p-1} \boldsymbol{Z}_{k+2 t p}(\boldsymbol{\beta})\right)
$$

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

$$
\begin{equation*}
\mathbb{P}\left[n^{-1} \boldsymbol{\chi}[\boldsymbol{\beta}]_{0}>n^{-1}{\overline{\mathbb{E}} \boldsymbol{\chi}(\boldsymbol{\beta})_{0}}^{1}+\frac{c_{1} \theta}{p^{3 / 2}}\right] \leq 2 p \cdot \mathbb{P}\left[\sum_{t=0}^{n / 2 p-1} \boldsymbol{Z}_{2 t p}(\boldsymbol{\beta})>\frac{n}{2 p} \mathbb{E} \boldsymbol{Z}(\boldsymbol{\beta})+\frac{c_{1} n \theta}{2 p^{5 / 2}}\right] \tag{B.395}
\end{equation*}
$$

The moments of $\boldsymbol{Z}_{0}$ can be bounded by using $\left|\boldsymbol{Z}_{0}(\boldsymbol{\beta})\right| \leq\left|\boldsymbol{x}_{0}\right|\left|\boldsymbol{\beta}_{0} \boldsymbol{x}_{0}+\boldsymbol{s}_{0}\right| \leq \boldsymbol{\beta}_{0} \boldsymbol{x}_{0}^{2}+\left|\boldsymbol{x}_{0}\right|\left|\boldsymbol{s}_{0}\right|$ 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

$$
\begin{equation*}
\mathbb{E}\left|s_{0}\right|^{2}=\mathbb{E}\left|\sum_{\ell} \boldsymbol{x}_{\ell} \boldsymbol{\beta}_{\ell}\right|^{2} \leq \theta\|\boldsymbol{\beta}\|_{2}^{2} \leq \theta\left(1+c_{\mu}+\frac{c_{\mu}}{\theta k^{2}}\right) \leq \frac{1}{2} \tag{B.396}
\end{equation*}
$$

As for the $q$-norm, use the moment generating function bound, such that for all $t \geq 0$ :

$$
\begin{align*}
\mathbb{E}\left|\boldsymbol{s}_{0}\right|^{q} & \leq q!t^{-q} \mathbb{E} \exp \left[t\left|s_{0}\right|\right] \leq q!t^{-q} \prod_{\ell} \mathbb{E}_{\boldsymbol{\omega}_{\ell}, \boldsymbol{g}_{\ell}} \exp \left[t \boldsymbol{\omega}_{\ell}\left|\boldsymbol{g}_{\ell}\right|\left|\boldsymbol{\beta}_{\ell}\right|\right] \leq 2 q!t^{-q} \prod_{\ell} \mathbb{E}_{\boldsymbol{\omega}_{\ell}} \exp \left[\boldsymbol{\omega}_{\ell} t^{2} \boldsymbol{\beta}_{\ell}^{2} / 2\right] \\
& \leq 2 q!t^{-q} \prod_{\ell}\left(1-\theta+\theta \exp \left[t^{2} \boldsymbol{\beta}_{\ell}^{2} / 2\right]\right) \tag{B.397}
\end{align*}
$$

notice that the entrywise twice derivative of B.397) w.r.t. $\boldsymbol{\beta}_{\ell}^{2 \prime}$ 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 B.397) w.r.t. $\boldsymbol{\beta}_{\ell}^{2}$ 's occurs and a vertex point where $\boldsymbol{\beta}_{0}^{2}=\|\boldsymbol{\beta}\|_{2}^{2}$. Thus

$$
\text { B.397) } \leq 2 q!t^{-q}\left(1-\theta+\theta \exp \left[t^{2}\|\boldsymbol{\beta}\|_{2}^{2} / 2\right]\right) \prod_{\ell \neq 0}\left(1-\theta+\theta e^{0}\right) \leq 2 q!t^{-q}\left(1+\theta \exp \left[\|\boldsymbol{\beta}\|_{2}^{2} t^{2} / 2\right]\right)
$$

Choose $t=\sqrt{q} /\|\boldsymbol{\beta}\|_{2}$, use $q!!>(q!/ 2) \cdot(e / q)^{q / 2}$, we have

$$
\begin{equation*}
\mathbb{E}\left|\boldsymbol{s}_{0}\right|^{q} \leq 2 q!q^{-q / 2}\|\boldsymbol{\beta}\|_{2}^{q}(1+\theta \exp [q / 2]) \leq 8\|\boldsymbol{\beta}\|_{2}^{q} \max \left\{e^{-q / 2}, \theta\right\} q!! \tag{B.398}
\end{equation*}
$$

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 Theorem B.10.2, B.396 and (B.398), obtain for $q \geq 3$,

$$
\begin{aligned}
& \mathbb{E} Z(\boldsymbol{\beta})^{2} \leq \mathbb{E}\left(\boldsymbol{\beta}_{0} \boldsymbol{x}_{0}^{2}+\left|\boldsymbol{x}_{0}\right|\left|\boldsymbol{s}_{0}\right|\right)^{2} \leq 2 \mathbb{E}\left[\boldsymbol{\beta}_{0}^{2} \boldsymbol{x}_{0}^{4}+\boldsymbol{x}_{0}^{2} \boldsymbol{s}_{0}^{2}\right] \leq 6 \theta+2 \theta^{2}\|\boldsymbol{\beta}\|_{2}^{2} \leq 7 \theta \\
& \mathbb{E} \boldsymbol{Z}(\boldsymbol{\beta})^{q} \leq \mathbb{E}\left(\boldsymbol{\beta}_{0} \boldsymbol{x}_{0}^{2}+\left|\boldsymbol{x}_{0}\right|\left|\boldsymbol{s}_{0}\right|\right)^{q} \leq 2^{q-1}\left(\mathbb{E} \boldsymbol{x}_{0}^{2 q}+\mathbb{E}\left|\boldsymbol{x}_{0}\right|^{q} \mathbb{E}\left|\boldsymbol{s}_{0}\right|^{q}\right)
\end{aligned}
$$

$$
\begin{aligned}
& \leq \theta 2^{q-1}(2 q-1)!!+\theta 2^{q-1}(q-1)!!\left(8\|\boldsymbol{\beta}\|_{2}^{q} \max \left\{e^{-q / 2}, \theta\right\} q!!\right) \\
& \leq \theta 4^{q} q!+\theta 2^{q}\|\boldsymbol{\beta}\|_{2}^{q} q!
\end{aligned}
$$

Thus, recall that $\|\boldsymbol{\beta}\|_{2}=\eta$, use $\left(\sigma^{2}, R\right)=\left(8 \theta \eta^{2}, 4 \eta\right)$, from B.394 -B.395, apply Bernstein inequality Theorem B.10.4 with $n \geq C p^{5} \theta^{-2} \log p$, and $c_{1}, c_{2} \in[0,1]$ we have

$$
\begin{align*}
\mathbb{P}\left[\mathcal{E}_{\mathrm{Net}}^{c}\right] & \leq 2 n p\left|\mathcal{N}_{\varepsilon}\right| \cdot \mathbb{P}\left[\sum_{t=0}^{n / 2 p-1} \boldsymbol{Z}_{2 t p}(\boldsymbol{\beta})>\frac{n}{2 p} \mathbb{E} \boldsymbol{Z}(\boldsymbol{\beta})+\frac{c_{1} n \theta}{2 p^{5 / 2}}\right] \leq 2 n p\left(\frac{3 n p^{2}}{c_{2}}\right)^{3 p} \exp \left(\frac{-\left(c_{1} n \theta / 2 p^{5 / 2}\right)^{2}}{16 n \theta \eta^{2} / 2 p+8 \eta c_{1} n \theta / 2 p^{5 / 2}}\right) \\
& \leq \exp \left(4 p \log \left(\frac{3 n p^{2}}{c_{2}}\right)-\frac{\left(c_{1} n \theta / 2 p^{5 / 2}\right)^{2}}{16 n \theta \eta^{2} / p}\right) \leq \exp \left(4 p \log \left(\frac{3 n p^{2}}{c_{2}}\right)-\frac{c_{1}^{2} n \theta^{2}}{64 p^{4}}\right) \\
& \leq \exp \left(\frac{-c_{1}^{2} n \theta^{2}}{100 p^{4}}\right) \leq \frac{1}{n} \tag{B.399}
\end{align*}
$$

when $\frac{C}{\log C}>\frac{10^{5}}{c_{1}^{2} c_{2}}$. The proof of lower bound and negative $\boldsymbol{\beta}_{0}$ is derived in the same manner.

## B.9.2 Proof of Theorem B.4.3

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 $\boldsymbol{Z}_{i j}(\boldsymbol{\beta})$ as

$$
\begin{equation*}
\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}[\boldsymbol{x}]\right\|_{2}^{2}=\sum_{j \in[n]} \boldsymbol{x}_{j}^{2} \mathbf{1}_{\left\{\left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[ \pm p]-i+j}\right\rangle\right|>\lambda\right\}}=: \sum_{j \in[n]} \boldsymbol{Z}_{i j}(\boldsymbol{\beta}) \tag{B.400}
\end{equation*}
$$

and define $\left\{\overline{\boldsymbol{Z}}_{i j}(\boldsymbol{\beta})\right\}_{j \in[n]}$ that are independent r.v.s. and as a upper bounding function of $\boldsymbol{Z}_{i j}(\boldsymbol{\beta})$ as

$$
\overline{\boldsymbol{Z}}_{i j}(\boldsymbol{\beta}):= \begin{cases}\boldsymbol{x}_{j}^{2}, & \left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[ \pm p]-i+j}\right\rangle\right|>\lambda  \tag{B.401}\\ 0, & \left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[ \pm p]-i+j}\right\rangle\right|<\lambda / 2 \\ \frac{\boldsymbol{x}_{j}^{2}}{\lambda / 2}\left(\left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[ \pm p]-i+j}\right\rangle\right|-\lambda / 2\right), & \text { otherwise }\end{cases}
$$

Similar to proof of Theorem B.3.4 Let $\|\boldsymbol{\beta}\|_{2} \leq \eta \leq \sqrt{p}$. Define $\varepsilon=\frac{c_{2}^{\prime} \lambda}{24 n p \sqrt{p \theta \log n \log \theta^{-1}}}$ for some $c_{2}^{\prime}>0$ and consider the $\varepsilon$-net $\mathcal{N}_{\varepsilon}$ for sphere of radius $\eta$. FromTheorem B.10.5 we know

$$
\begin{equation*}
\left|\mathcal{N}_{\varepsilon}\right| \leq\left(\frac{3 \eta}{\varepsilon}\right)^{2 p} \leq\left(\frac{72}{c_{2}^{\prime} c_{\lambda}} n p^{2} \sqrt{\theta|\boldsymbol{\tau}| \log n \log \theta^{-1}}\right)^{2 p} \leq\left(\frac{72}{c_{2}^{\prime} c_{\lambda}} n p^{2} \log n\right)^{2 p} \tag{B.402}
\end{equation*}
$$

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:

$$
\begin{equation*}
\mathcal{E}_{\text {Net }}:=\left\{\forall i \in[n], \quad\left|n^{-1} \sum_{j \in[n]} \overline{\boldsymbol{Z}}_{i j}\left(\boldsymbol{\beta}_{\varepsilon}\right)-\mathbb{E} \overline{\boldsymbol{Z}}_{i}\left(\boldsymbol{\beta}_{\varepsilon}\right)\right| \leq \frac{c_{1}^{\prime} \theta}{p} \quad \forall \boldsymbol{\beta}_{\varepsilon} \in \mathcal{N}_{\varepsilon, i}\right\}, \tag{B.403}
\end{equation*}
$$

Also, $\sum_{j} \overline{\boldsymbol{Z}}_{i j}(\boldsymbol{\beta})$ is a Lipchitz function over $\boldsymbol{\beta}$ for every $i \in[n]$ as

$$
\begin{align*}
\left|\sum_{j \in[n]} \overline{\boldsymbol{Z}}_{i j}(\boldsymbol{\beta})-\sum_{j \in[n]} \overline{\boldsymbol{Z}}_{i j}\left(\boldsymbol{\beta}^{\prime}\right)\right| & \leq \sum_{j \in[n]} \frac{\boldsymbol{x}_{j}^{2}}{\lambda / 2}\left|\left\langle\boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}, \boldsymbol{x}_{[ \pm p]-i+j}\right\rangle\right| \leq \sum_{j \in[n]} \frac{\boldsymbol{x}_{j}^{2}\left\|\boldsymbol{x}_{[ \pm p]-i+j}\right\|_{2}}{\lambda / 2}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}\right\|_{2} \\
& \leq \frac{1}{\lambda / 2}\|\boldsymbol{x}\|_{2}^{2} \cdot \max _{j \in[n]}\left\|\boldsymbol{x}_{[ \pm p]+j}\right\|_{2} \cdot\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}\right\|_{2}:=L\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}\right\|_{2} \tag{B.404}
\end{align*}
$$

and define event $\mathcal{E}_{\text {Lip }}$ such that the Lipchitz constant is bounded as

$$
\begin{equation*}
\mathcal{E}_{\text {Lip }}:=\left\{L \leq 12 n \theta \sqrt{p \theta \log n \log \theta^{-1}} \lambda^{-1}\right\} \tag{B.405}
\end{equation*}
$$

then on event $\mathcal{E}_{\text {Lip }}$, for any points $\boldsymbol{\beta}$ in $\mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ and $i \in[n]$, there exists some $\boldsymbol{\beta}_{\varepsilon}$ in $\mathcal{N}_{\varepsilon, i}$ with $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_{\varepsilon}\right\|_{2} \leq \varepsilon$, and thus

$$
\begin{equation*}
\left|\left(n^{-1} \sum_{j \in[n]} \overline{\boldsymbol{Z}}_{i j}(\boldsymbol{\beta})-\mathbb{E} \overline{\boldsymbol{Z}}_{i}(\boldsymbol{\beta})\right)-\left(n^{-1} \sum_{j \in[n]} \overline{\boldsymbol{Z}}_{i j}\left(\boldsymbol{\beta}_{\varepsilon}\right)-\mathbb{E} \overline{\boldsymbol{Z}}_{i}\left(\boldsymbol{\beta}_{\varepsilon}\right)\right)\right| \leq 2 L \varepsilon \leq \frac{c_{2}^{\prime} \theta}{p} . \tag{B.406}
\end{equation*}
$$

On event $\mathcal{E}_{\text {Lip }} \cap \mathcal{E}_{\text {Net }}$, from (B.403, , B.406, we can conclude that for all $\boldsymbol{\beta} \in \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ and $i \in[n]$ that:

$$
\begin{equation*}
n^{-1}\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2}-n^{-1} \mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}\left[\boldsymbol{x}_{0}\right]\right\|_{2}^{2} \leq n^{-1} \sum_{j \in[n]} \overline{\boldsymbol{Z}}_{i j}(\boldsymbol{\beta})-\mathbb{E} \overline{\boldsymbol{Z}}_{i}(\boldsymbol{\beta}) \leq \frac{\left(c_{1}^{\prime}+c_{2}^{\prime}\right) \theta}{p} \tag{B.407}
\end{equation*}
$$

as desired, where the error probability of $\mathcal{E}_{\text {Lip }}^{c}$ is bounded using Theorem B.1.2 and Theorem B.1.3. which give

$$
\begin{equation*}
\mathbb{P}\left[\mathcal{E}_{\text {Lip }}^{c}\right] \leq \mathbb{P}\left[\|\boldsymbol{x}\|_{2}^{2}>2 n \theta\right]+\mathbb{P}\left[\max _{j \in[n]}\left\|\boldsymbol{x}_{[ \pm p]+j}\right\|_{2}>3 \sqrt{p \theta \log n \log \theta^{-1}}\right] \leq 3 / n \tag{B.408}
\end{equation*}
$$

when $n>10^{3} \theta^{-1}$. As for $\mathcal{E}_{\text {Net }}^{c}$ use union bound and split the r.v.s since $\boldsymbol{Z}_{j}, \boldsymbol{Z}_{j+2 p}$ are independent for all $j$ :

$$
\mathbb{P}\left[\mathcal{E}_{\mathrm{Net}}^{c}\right] \leq 2 n p \cdot\left|\mathcal{N}_{\varepsilon}\right| \cdot \mathbb{P}\left[\left|\sum_{k}^{n / 2 p} \overline{\boldsymbol{Z}}_{i, 2 k j}(\boldsymbol{\beta})-\frac{n}{2 p} \mathbb{E} \overline{\boldsymbol{Z}}_{i}(\boldsymbol{\beta})\right| \geq \frac{c_{1}^{\prime} n \theta}{2 p^{2}}\right]
$$

Now we calculate the variance and $L^{q}$-norm of $\sum_{k} \bar{Z}_{i, 2 k j}$ for $q \geq 3$ :

$$
\left\{\begin{array}{l}
\mathbb{E} \overline{\boldsymbol{Z}}_{i, j}^{2} \leq \mathbb{E} \boldsymbol{x}_{j}^{4} \leq 3 \theta  \tag{B.409}\\
\mathbb{E} \overline{\boldsymbol{Z}}_{i, j}^{q} \leq \mathbb{E} \boldsymbol{x}_{j}^{2 q} \leq \theta(2 q-1)!!\leq \frac{1}{2} \cdot(3 \theta) \cdot 2^{q-2} q!
\end{array}\right.
$$

and apply Bernstein inequality with $\left(\sigma^{2}, R\right)=(3 \theta, 2)$, then use $n \geq C p^{4} \theta^{-1} \log p$ and $c_{1}^{\prime}, c_{2}^{\prime}<1$ to obtain

$$
\begin{align*}
2 n p\left|\mathcal{N}_{\varepsilon}\right| \mathbb{P}\left[\left|\sum_{k}^{n / 2 p} \overline{\boldsymbol{Z}}_{i, 2 k j}(\boldsymbol{\beta})-\frac{n}{2 p^{2}} \mathbb{E} \overline{\boldsymbol{Z}}_{i}\right| \geq \frac{c_{1}^{\prime} n \theta}{2 p^{2}}\right] & \leq \exp \left[\log (2 n p)+2 p \log \left(\frac{72}{c_{2}^{\prime} c_{\lambda}} n p^{2} \log n\right)-\frac{\left(c_{1}^{\prime} n \theta / 2 p^{2}\right)^{2}}{6 n \theta / 2 p+4 c_{1}^{\prime} n \theta / 2 p^{2}}\right] \\
& \leq \exp \left[3 p \log \left(\frac{72}{c_{2}^{\prime} c_{\lambda}} n p^{2} \log n\right)-\frac{c_{1}^{\prime 2} n \theta}{24 p^{3}}\right] \\
& \leq \exp \left[-c_{1}^{\prime 2} n \theta /\left(50 p^{3}\right)\right] \leq 1 / n, \tag{B.410}
\end{align*}
$$

where the last two inequalities holds when $\frac{C}{\log C} \geq \frac{10^{5}}{c_{1}^{1} c_{2}^{\prime} c_{\lambda}}$. The other side of inequality of (B.91) can be derived by defining $\underline{Z}_{i j}$ as

$$
\underline{\boldsymbol{Z}}_{i j}(\boldsymbol{\beta}):= \begin{cases}\boldsymbol{x}_{j}^{2}, & \left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[ \pm p]-i+j}\right\rangle\right|>3 \lambda / 2  \tag{B.411}\\ 0, & \left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[ \pm p]-i+j}\right\rangle\right|<\lambda \\ \frac{\boldsymbol{x}_{j}^{2}}{\lambda / 2}\left(\left|\left\langle\boldsymbol{\beta}, \boldsymbol{x}_{[ \pm p]-i+j}\right\rangle\right|-\lambda\right), & \text { otherwise }\end{cases}
$$

and define $\mathcal{E}_{\text {Net }}, \mathcal{E}_{\text {Lip }}$ similarly, such that on intersection of these events,

$$
\begin{equation*}
n^{-1}\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}[\boldsymbol{x}]\right\|_{2}^{2}-n^{-1} \mathbb{E}\left\|\boldsymbol{P}_{I(\boldsymbol{\beta})} s_{-i}[\boldsymbol{x}]\right\|_{2}^{2} \geq n^{-1} \sum_{j \in[n]} \underline{\boldsymbol{Z}}_{i j}(\boldsymbol{\beta})-\mathbb{E} \underline{\boldsymbol{Z}}_{i}(\boldsymbol{\beta}) \geq \frac{\left(c_{1}^{\prime}+c_{2}^{\prime}\right) \theta}{p} \tag{B.412}
\end{equation*}
$$

as desired.

## B.9.3 Proof of Theorem B.5.5

Proof. 1. (Expectation upper bound) We will write $\boldsymbol{x}$ as $\boldsymbol{x}_{0}$. Similar to proof of Theorem B.3.4 let $\|\boldsymbol{\beta}\|_{2} \leq \eta \leq$ $\sqrt{p}$. For each $i \in[n]$, define the random variable

$$
\begin{equation*}
\boldsymbol{X}_{i}(\boldsymbol{\beta})=\mathbf{1}_{\left\{\left|\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}\right\rangle-\lambda\right| \leq B\right\}}+\mathbf{1}_{\left\{\left|\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}\right\rangle+\lambda\right| \leq B\right\}}, \tag{B.413}
\end{equation*}
$$

then number of indices for vector $\boldsymbol{x} * \breve{\boldsymbol{\beta}}$ 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

$$
\overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})= \begin{cases}\frac{1}{M}\left(\left|\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}\right\rangle\right|-(\lambda-B-M)\right) & \left|\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}\right\rangle\right| \in[\lambda-B-M, \lambda-B]  \tag{B.414}\\ 1 & \left|\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}\right\rangle\right| \in[\lambda-B, \lambda+B] \\ \frac{1}{M}\left((\lambda+B+M)-\left|\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}\right\rangle\right|\right) & \left|\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}\right\rangle\right| \in[\lambda+B, \lambda+B+M] \\ 0 & \text { else }\end{cases}
$$

where $B<M=c \lambda \theta^{2} /(p \log n) \leq \lambda / 4$ for some constant $0<c<1$.
Notice that $\boldsymbol{x} \sim_{\text {i.i.d. }} \operatorname{BG}(\theta)$ 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 $I(\boldsymbol{a}),\langle\boldsymbol{x}, \boldsymbol{\beta}\rangle=\left\langle\boldsymbol{g}, \boldsymbol{P}_{I(\boldsymbol{a})} \boldsymbol{\beta}\right\rangle \sim \mathcal{N}\left(0,\left\|\boldsymbol{P}_{I(\boldsymbol{a})} \boldsymbol{\beta}\right\|_{2}^{2}\right)$. 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 $f_{\sigma}(t)=(\sqrt{2 \pi} \sigma)^{-1} \exp \left(-t^{2} / 2 \sigma^{2}\right)$ denote the pdf of an $\mathcal{N}\left(0, \sigma^{2}\right)$ random variable, the expectation of $\sum_{i} \overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})$ can be upper bounded as

$$
\begin{align*}
\sum_{i \in[n]} \mathbb{E}\left[\overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})\right] & \leq(2 n) \cdot \mathbb{P}[\langle\boldsymbol{x}, \boldsymbol{\beta}\rangle \in[\lambda-B-M, \lambda+B+M]] \\
& \leq(2 n) \cdot 2(B+M) \sup _{\sigma^{2} \in(0, p]} t \in[\lambda-B-M, \lambda+B+M] \\
& \max _{\sigma}(t) \\
& \leq 4 n(B+M) \sup _{\sigma^{2} \in(0, p]} f_{\sigma}(\lambda-B-M)  \tag{B.415}\\
& \leq 4 n(B+M) \sup _{\sigma^{2} \in(0, p]} f_{\sigma}(\lambda / 2) .
\end{align*}
$$

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

$$
\begin{equation*}
\sup _{\sigma^{2} \in(0, p]} f_{\sigma}(\lambda / 2) \leq f_{\lambda / 2}(\lambda / 2) \leq \frac{1}{\sqrt{2 \pi}(\lambda / 2)} \exp \left(-\frac{1}{2}\right) \leq \frac{1}{2 \lambda}, \tag{B.416}
\end{equation*}
$$

whence, by letting $B \leq c \lambda \theta^{2} /(p \log n)$, the upper bound of expectation become:

$$
\begin{equation*}
\sum_{i \in[n]} \mathbb{E}\left[\overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})\right] \leq \frac{4 n}{2 \lambda}(B+M) \leq \frac{4 c n \theta^{2}}{p \log n}=: n \mathbb{E} \overline{\boldsymbol{X}(\boldsymbol{\beta})} . \tag{B.417}
\end{equation*}
$$

2. ( $\underline{\varepsilon \text {-net }})$ Define $\varepsilon=\frac{c^{2} \lambda \theta^{3.5}}{3 p^{2.5} \log ^{2.5} n \log ^{0.5} \theta^{-1}}$. Write $\lambda=c_{\lambda} / \sqrt{|\boldsymbol{\tau}|}$ and consider the $\varepsilon$-net $\mathcal{N}_{\varepsilon}$ for sphere of radius $\eta \leq \sqrt{p}$. From Theorem B.10.5 we know

$$
\begin{equation*}
\left|\mathcal{N}_{\varepsilon}\right| \leq\left(\frac{3 \eta}{\varepsilon}\right)^{2 p} \leq\left(\frac{81|\boldsymbol{\tau}| p^{6} \log ^{5} n \log \theta^{-1}}{c^{4} c_{\lambda}^{2} \theta^{7}}\right)^{p} \leq\left(\frac{2 p \log n}{c \cdot c_{\lambda}}\right)^{13 p} \tag{B.418}
\end{equation*}
$$

and define an event such that all center of subsets in $\mathcal{N}_{\varepsilon}$ are being well-behaved:

$$
\begin{equation*}
\mathcal{E}_{\mathrm{Net}}:=\left\{\sum_{i \in[n]} \overline{\boldsymbol{X}}_{i}\left(\boldsymbol{\beta}_{\varepsilon}\right)-n \mathbb{E} \overline{\boldsymbol{X}}\left(\boldsymbol{\beta}_{\varepsilon}\right)<\frac{18 c n \theta^{2}}{p \log n} \quad \forall \boldsymbol{\beta}_{\varepsilon} \in \mathcal{N}_{\varepsilon},\right\} \tag{B.419}
\end{equation*}
$$

3. (Lipschitz constant) Furthermore, the function $\sum_{i}^{n} \overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})$ is Lipchitz over $\boldsymbol{\beta}$ such that

$$
\left|\sum_{i \in[n]} \overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})-\sum_{i \in[n]} \overline{\boldsymbol{X}}_{i}\left(\boldsymbol{\beta}^{\prime}\right)\right| \leq \sum_{i \in[n]}^{n} \frac{1}{M}\left|\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}\right\rangle\right| \leq \frac{n}{M} \max _{i \in[n]}\left\|\boldsymbol{P}_{[ \pm p]+i} \boldsymbol{x}\right\|_{2}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}\right\|_{2}=: L\left\|\boldsymbol{\beta}-\boldsymbol{\beta}^{\prime}\right\|_{2}
$$

define the set $\mathcal{N}_{\varepsilon}$ where Lipschitz constant is well bounded:

$$
\mathcal{E}_{\text {Lip }}:=\left\{L \leq \frac{3 n \sqrt{p \theta \log n \log \theta^{-1}}}{M}\right\}
$$

then on event $\mathcal{E}_{\text {Lip }}$, for every $\boldsymbol{\beta}$ in $\mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$, there exists some $\boldsymbol{\beta}_{\varepsilon}$ in $\mathcal{N}_{\varepsilon, i}$ with $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_{\varepsilon}\right\|_{2} \leq \varepsilon$, thus

$$
\begin{equation*}
\left|\left(\sum_{i \in[n]} \overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})-n \mathbb{E} \overline{\boldsymbol{X}}(\boldsymbol{\beta})\right)-\left(\sum_{i \in[n]} \overline{\boldsymbol{X}}_{i}\left(\boldsymbol{\beta}_{\varepsilon}\right)-n \mathbb{E} \overline{\boldsymbol{X}}\left(\boldsymbol{\beta}_{\varepsilon}\right)\right)\right| \leq 2 L \varepsilon \leq \frac{2 c n \theta^{2}}{p \log n} \tag{B.420}
\end{equation*}
$$

On event $\mathcal{E}_{\text {Lip }} \cap \mathcal{E}_{\text {Net }}$, from B.417, B.419 and B.420), we can conclude that for every $\boldsymbol{\beta} \in \mathfrak{R}\left(\mathcal{S}_{\boldsymbol{\tau}}, \gamma\left(c_{\mu}\right)\right)$ and $i \in[n]$,

$$
\begin{equation*}
\sum_{i \in[n]} \overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta}) \leq \frac{24 c n \theta^{2}}{p \log n} \tag{B.421}
\end{equation*}
$$

as desired, where the error probability of $\mathcal{E}_{\text {Lip }}^{c}$ is bounded using Theorem B.1.3. which gives

$$
\begin{equation*}
\mathbb{P}\left[\mathcal{E}_{\text {Lip }}^{c}\right] \leq \mathbb{P}\left[\max _{j \in[n]}\left\|\boldsymbol{x}_{[ \pm p]+j}\right\|_{2}>3 \sqrt{p \theta \log n \log \theta^{-1}}\right] \leq 2 / n \tag{B.422}
\end{equation*}
$$

4. (Bound $\left.\mathbb{P}\left[\mathcal{E}_{\text {Net }}^{c}\right]\right)$ Wlog let us assume that $2 p$ divides $n$. By applying union bound and observing that $\overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})$ is independent of $\overline{\boldsymbol{X}}_{i+2 p}(\boldsymbol{\beta})$ for any $i \in[n]$, we split $\sum_{i} \overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})$ into $n / 2 p$ independent sums of r.v.s, we have

$$
\mathbb{P}\left[\mathcal{E}_{\text {Net }}^{c}\right] \leq 2 p\left|\mathcal{N}_{\varepsilon}\right| \cdot \mathbb{P}\left[\sum_{j=0}^{n / 2 p-1}\left(\overline{\boldsymbol{X}}_{2 p j}(\boldsymbol{\beta})-\mathbb{E}[\overline{\boldsymbol{X}}(\boldsymbol{\beta})]\right)>\frac{9 c n \theta^{2}}{p^{2} \log n}\right]
$$

where each summand has bounded variance and $L^{q}$-norm derived similarly as its expectation such that

$$
\mathbb{E} \overline{\boldsymbol{X}}_{i}(\boldsymbol{\beta})^{q} \leq 2 \cdot \mathbb{P}\left[\left\langle s_{i}[\boldsymbol{x}], \boldsymbol{\beta}\right\rangle \in[\lambda-B-M, \lambda+B+M]\right] \leq 2 \cdot \frac{1}{2 \lambda} \cdot 2(B+M) \leq \frac{4 c \theta^{2}}{p \log n}
$$

and apply Bernstein inequality Theorem B.10.4 with $\left(\sigma^{2}, R\right)=\left(4 c \theta^{2} /(p \log n), 1\right)$, obtains
$\mathbb{P}\left[\sum_{j=0}^{n / 2 p-1}\left(\overline{\boldsymbol{X}}_{2 p j}(\boldsymbol{\beta})-\mathbb{E}[\overline{\boldsymbol{X}}(\boldsymbol{\beta})]\right)>\frac{9 c n \theta^{2}}{p^{2} \log n}\right] \leq \exp \left[\frac{-\left(9 c n \theta^{2} / p^{2} \log n\right)^{2}}{2 c n \theta^{2} / p^{2} \log n+2\left(9 c n \theta^{2} / p^{2} \log n\right)}\right] \leq \exp \left[\frac{-4 c n \theta^{2}}{p^{2} \log n}\right]$, thus when $n=C p^{5} \theta^{-2} \log p$ :

$$
\begin{equation*}
\mathbb{P}\left[\mathcal{E}_{\text {Net }}^{c}\right] \leq \exp \left[\log (2 p)+13 p \log \left(\frac{2 p \log n}{c \cdot c_{\lambda}}\right)-\frac{4 c n \theta^{2}}{p^{2} \log n}\right] \leq 1 / n \tag{B.423}
\end{equation*}
$$

as long as $\frac{C}{\log C}>10^{5} /\left(c^{2} \cdot c_{\lambda}\right)$.

## B. 10 Tools

Lemma B.10.1 (Tail bound for Gaussian r.v.). If $X \sim \mathcal{N}\left(0, \sigma^{2}\right)$, then its tail bound for $t>0$ can be

$$
\begin{equation*}
\mathbb{P}[X>t] \leq \frac{\sigma}{t \sqrt{2 \pi}} \exp \left(-\frac{t^{2}}{2 \sigma^{2}}\right) \tag{B.424}
\end{equation*}
$$

Lemma B.10.2 (Moments of the Gaussian random variables). If $X \sim \mathcal{N}\left(0, \sigma^{2}\right)$, then ifor all integer $p \geq 1$,

$$
\begin{equation*}
\mathbb{E}\left[|X|^{p}\right] \leq \sigma^{p}(p-1)!! \tag{B.425}
\end{equation*}
$$

Lemma B.10.3 (Gaussian concentration inequality). Let $\boldsymbol{x}=\left(\boldsymbol{x}_{1}, \ldots, \boldsymbol{x}_{n}\right)$ be a vector of $n$ independent standard normal variables. Let $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ be an L-Lipschitz function. Then for all $t>0$,

$$
\begin{equation*}
\mathbb{P}[|f(\boldsymbol{x})-\mathbb{E} f(\boldsymbol{x})| \geq t] \leq 2 \exp \left(-\frac{t^{2}}{2 L^{2}}\right) \tag{B.426}
\end{equation*}
$$

Lemma B.10.4 (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

$$
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left|\boldsymbol{x}_{k}\right|^{p}\right] \leq \frac{1}{2} \sigma^{2} R^{p-2} p!, \text { for all integers } p \geq 3
$$

Let $S \doteq \sum_{i=1}^{n} \boldsymbol{x}_{i}$, then for all $t>0$, it holds that

$$
\begin{equation*}
\mathbb{P}[|S-\mathbb{E}[S]| \geq t] \leq 2 \exp \left(-\frac{t^{2}}{2 n \sigma^{2}+2 R t}\right) \tag{B.427}
\end{equation*}
$$

Lemma B.10.5 ( $\varepsilon$-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 $\left|\mathcal{N}_{\varepsilon}\right| \leq(3 / \varepsilon)^{n}$.

Lemma B.10.6 (Hanson-Wright). [RV $\left.{ }^{+} 13\right]$ Let $\boldsymbol{x}_{1}, \ldots, \boldsymbol{x}_{n}$ be independent, subgaussian random variables with subgaussian norm $\sup _{p \geq 1} p^{-1 / 2}\left(\mathbb{E}\left|x_{i}^{p}\right|\right)^{1 / p} \leq \sigma$. Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, then for every $t>0$,

$$
\begin{equation*}
\mathbb{P}\left[\left|\boldsymbol{x}^{*} \boldsymbol{A} \boldsymbol{x}-\mathbb{E} \boldsymbol{x}^{*} \boldsymbol{A} \boldsymbol{x}\right| \geq t\right] \leq 2 \exp \left(-c \min \left(\frac{t^{2}}{64 \sigma^{4}\|\boldsymbol{A}\|_{F}^{2}}, \frac{t}{8 \sqrt{2} \sigma^{2}\|\boldsymbol{A}\|_{2}}\right)\right) \tag{B.428}
\end{equation*}
$$

Lemma B.10.7 (Maximum of separable convex function). Let $f: \mathbb{R}_{+} \rightarrow \mathbb{R}_{+}$be a convex function of the form $f(x)=x-s(x)$ with $s: \mathbb{R}_{+} \rightarrow \mathbb{R}_{+}$satisfying

$$
\frac{s(x)}{x} \leq \frac{s(y)}{y}, \text { for all } x \geq y>0
$$

Then for $n \in \mathbb{N}$ and $0<N \leq n L$,

$$
\begin{equation*}
\max _{0 \leq \boldsymbol{x} \leq L,\|\boldsymbol{x}\|_{1} \leq N} \sum_{i=1}^{n} f\left(\boldsymbol{x}_{i}\right) \leq N\left(1-\frac{s(L)}{L}\right) \tag{B.429}
\end{equation*}
$$

Proof. Since the feasible set is a convex polytope; the convex function $\sum_{i=1}^{n} f\left(\boldsymbol{x}_{i}\right)$ is maximized at a vertex, and that its vertices consist of 0 and permutations of the vector $[\underbrace{L, \ldots, L}_{\lfloor N / L\rfloor}, r, 0, \ldots, 0]$, where $r=N-\lfloor N / L\rfloor L \leq L$. Then the function value at the maximizing vector $\boldsymbol{x}_{*}$ can be derived as:

$$
\begin{aligned}
\sum_{i=1}^{n} f\left(\boldsymbol{x}_{* i}\right) & =\left\lfloor\frac{N}{L}\right\rfloor f(L)+f(r)=\frac{N-r}{L}(L-s(L))+(r-s(r)) \\
& =N\left(1-\frac{s(L)}{L}\right)+r\left(\frac{s(L)}{L}-\frac{s(r)}{r}\right) \leq N\left(1-\frac{s(L)}{L}\right)
\end{aligned}
$$


[^0]:    ${ }^{1}$ The Dirac measure $\boldsymbol{\delta}$ satisfies $\int \boldsymbol{y}(t) \boldsymbol{\delta}_{\boldsymbol{t}_{i}}(d t)=\boldsymbol{y}\left(t_{i}\right)$ for continuous and compactly supported $\boldsymbol{y}$ and has total variation $\int\left|\boldsymbol{\delta}_{t_{i}}\right|(d t)=$ 1 , so $\boldsymbol{y} * \boldsymbol{\delta}_{t}$ represents $\boldsymbol{y}$ with center at $t$.

[^1]:    ${ }^{1}$ The Dirac measure $\boldsymbol{\delta}$ satisfies $\int \boldsymbol{D}(\boldsymbol{w}) \boldsymbol{\delta}_{\boldsymbol{w}_{i}}(d \boldsymbol{w})=\boldsymbol{D}\left(\boldsymbol{w}_{i}\right)$ for continuous and compactly supported $\boldsymbol{D}$ and has total variation $\int\left|\boldsymbol{\delta}_{\boldsymbol{w}_{i}}\right|(d \boldsymbol{w})=1$, so $\boldsymbol{D} * \boldsymbol{\delta}_{\boldsymbol{w}}$ represents $\boldsymbol{D}$ with center at $\boldsymbol{w}$ Rud06]. As a functional, we write $\left\langle\boldsymbol{\delta}_{\boldsymbol{w}_{i}}, \cdot\right\rangle: L^{2}\left(\mathbb{R}^{2}\right) \rightarrow \mathbb{R}$ where $\left\langle\boldsymbol{\delta}_{\boldsymbol{w}_{i}}, \boldsymbol{D}\right\rangle=\boldsymbol{D}\left(\boldsymbol{w}_{i}\right)$.

[^2]:    ${ }^{2}$ The result of $\lambda_{\min }(\widetilde{\boldsymbol{G}})$ remains almost identical with other dense motif allocation on lattice such as rectangular grid.

[^3]:    ${ }^{3}$ With fixed density, imaging area is proportional to disc count, and the number of samples is (line count) $\times \sqrt{(\text { imaging area })}=$ $N \times \sqrt{k}$.

[^4]:    ${ }^{4}$ We invoke the canonical definition of inner product of $L^{2}$-space for both image and lines. For every images $\boldsymbol{Y}, \boldsymbol{Y}^{\prime} \in L^{2}\left(\mathbb{R}^{2}\right)$, we define $\left\langle\boldsymbol{Y}, \boldsymbol{Y}^{\prime}\right\rangle=\int \boldsymbol{Y}(\boldsymbol{w}) \boldsymbol{Y}^{\prime}(\boldsymbol{w}) d \boldsymbol{w}$; and for every lines $\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{R}}^{\prime} \in L^{2}(\mathbb{R} \times[m])$, we define $\left\langle\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{R}}^{\prime}\right\rangle=\sum_{i=1}^{m} \int \widetilde{\boldsymbol{R}}_{i}(t) \widetilde{\boldsymbol{R}}_{i}^{\prime}(t) d t$.

[^5]:    ${ }^{1}$ In this paper, the cyclic convolution $\boldsymbol{a}_{0} * \boldsymbol{x}_{0}$ assumes $\boldsymbol{a}_{0}$ to be zeropadded $\left[\boldsymbol{a}_{0}, \mathbf{0}^{n-p_{0}}\right]$ to length $n$.
    ${ }^{2}$ Our result can be applied to recovering direct convolutions. Let $\boldsymbol{y} \in \mathbb{R}^{p_{0}+n-1}$ be the direct convolution between $\boldsymbol{a}_{0} \in \mathbb{R}^{p_{0}}$ and $\boldsymbol{x}_{0} \in \mathbb{R}^{n}$, then $\boldsymbol{y}$ can also be expressed as circular convolution between $\boldsymbol{a}_{0}$ and $\left[\boldsymbol{x}_{0} ; \mathbf{0}^{p_{0}-1}\right]$.
    ${ }^{3}$ Such as matrix lifting relaxation ARR14 LLB16, in which $\boldsymbol{a}_{0}$ or $\boldsymbol{x}_{0}$ resides in random subspaces w/o shift symmetry.
    ${ }^{4}$ I.e., the portion of the shifted signal $s_{\ell}\left[\boldsymbol{a}_{0}\right]$ that falls in the window $\left\{0, \ldots, p_{0}-1\right\}$.

[^6]:    ${ }^{5}$ For a generic $\boldsymbol{a}$, we have $\left\langle s_{i}[\boldsymbol{a}], s_{j}[\boldsymbol{a}]\right\rangle \approx 0$ and hence $\|\boldsymbol{a} * \boldsymbol{x}\|_{2}^{2}=\boldsymbol{x}^{*} \boldsymbol{C}_{\boldsymbol{a}}^{*} \boldsymbol{C}_{\boldsymbol{a}} \boldsymbol{x} \approx \boldsymbol{x}^{*} \boldsymbol{I} \boldsymbol{x}=\|\boldsymbol{x}\|_{2}^{2}$.
    ${ }^{6}$ The objective $\varphi_{\ell^{1}}$ is not twice differentiable everywhere, and hence cannot be minimized using conventional second order methods.
    ${ }^{7}$ This particular surrogate is sometimes being named as the pseudo-Huber function.

[^7]:    ${ }^{8}$ This is purely a technical convenience. Our theory guarantees recovery of a signed shift $\left( \pm s_{\ell}\left[\boldsymbol{a}_{0}\right], \pm s_{-\ell}\left[\boldsymbol{x}_{0}\right]\right)$ of the truth. If $\boldsymbol{a}_{0}$ does not have unit norm, identical reasoning implies that our method recovers a scaled shift $\left(\alpha s_{\ell}\left[\boldsymbol{a}_{0}\right], \alpha^{-1} s_{-\ell}\left[\boldsymbol{x}_{0}\right]\right)$ with $\alpha= \pm \frac{1}{\left\|\boldsymbol{a}_{0}\right\|_{2}}$.
    ${ }^{9}$ The use of " $\approx$ " here suppresses constant and logarithmic factors.

[^8]:    ${ }^{10}$ The upper right panel of Figure 3.2 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 $\mu\left(a_{0}\right)$ 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.

[^9]:    ${ }^{11}$ 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, $-\boldsymbol{M a}=-\nabla \xi(\boldsymbol{a})$. The generalized power method, for minimizing a function $\varphi$ over the sphere consists of repeatedly projecting $-\nabla \varphi$ onto the sphere, giving the iteration $\boldsymbol{a} \mapsto-\boldsymbol{P}_{\mathbb{S} p-1} \nabla \varphi(\boldsymbol{a})$. 3.20 can be interpreted as one step of the generalized power method for the objective function $\varphi_{\rho}$.

[^10]:    ${ }^{12}$ In practice, we suggest setting $\lambda=c_{\lambda} / \sqrt{p_{0} \theta}$ with $c_{\lambda} \in[0.5,0.8]$.

[^11]:    ${ }^{13}$...so the projection of the Euclidean gradient onto the tangent space does not vanish.

[^12]:    ${ }^{1}$ Apply $\exp \left[-x^{2} / 2\right]>1-x^{2} / 2$

