Supplementary Material for Geometry and Symmetry in Short-and-Sparse Deconvolution

Anonymous authors

A Notations

All vectors/matrices are written in bold font \( \mathbf{a} / \mathbf{A} \): indexed values are written as \( a_i, \mathbf{A}_{ij} \). Zeros or ones vectors are defined as 0 or 1, and \( i \)-th canonical basis vector defined as \( e_i \). 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 capital italic symbols \( I, J \) for subsets of \( [n] \). We abuse notation slightly and write \( [\pm p] = \{n - p + 1, \ldots, n-1, 0\} \) and \( [\pm p] = \{n - p + 1, \ldots, n-1, 0, 1, \ldots, p-1\} \). Index sets can be labels for vectors; \( \mathbf{a}_I \in \mathbb{R}^{|I|} \) denotes the restriction of the vector \( \mathbf{a} \) to coordinates \( I \). Also, we use check symbol for reversal operator on index set \( I = -I \) and vectors \( \mathbf{a}_I = \mathbf{a}_{-I} \).

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

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

\[
\mathbf{C}_\mathbf{a} \mathbf{b} = \mathbf{a} * \mathbf{b} \tag{A.1}
\]

The correlation between \( \mathbf{a} \) and \( \mathbf{b} \) can be also written in similar form of convolution operator which reverse one vector before convolution. Define two correlation matrices \( \mathbf{C}_{\mathbf{a}}^\ast \) and \( \tilde{\mathbf{C}}_{\mathbf{a}} \) as \( \mathbf{C}_{\mathbf{a}}^\ast e_j = s_j [\mathbf{a}] \) and \( \tilde{\mathbf{C}}_{\mathbf{a}} e_j = s_{-j} [\mathbf{a}] \). The two operators will satisfy

\[
\mathbf{C}_{\mathbf{a}}^\ast \mathbf{b} = \check{\mathbf{a}} * \mathbf{b}, \quad \tilde{\mathbf{C}}_{\mathbf{a}} \mathbf{b} = \mathbf{a} * \check{\mathbf{b}} \tag{A.2}
\]

B Geometry of \( \phi_\rho \) in Shift Space

Underlying our main geometric and algorithmic results is a relationship between the geometry of the function \( \phi_\rho \) and the symmetries of the deconvolution problem. In this section, we present Theorem B.1, which is a more general result of the main geometric theorem Theorem 3.1. We describe this relationship at a more technical level, by interpreting the gradient and hessian of the function \( \phi_\rho \) in terms of the shifts \( s_\ell [\mathbf{a}_0] \) and stating a key lemma which asserts that a certain neighborhood of the union of subspaces \( \Sigma_{\ell \phi_\rho} \) can be decomposed into regions of negative curvature, strong gradient, and strong convexity near the target solutions \( \pm s_\ell [\mathbf{a}_0] \).
B.1 Shifts and Correlations

The set $\Sigma_{4\theta_0}$ is a union of subspaces. Any point $a$ in one of these subspaces $S_\tau$ is a superposition of shifts of $a_0$:

$$a = \sum_{\ell \in \tau} \alpha_\ell S_\ell[a_0]. \quad (B.1)$$

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

$$a = \sum_{\ell \in \tau} \alpha_\ell S_\ell[a_0] + \sum_{\ell \notin \tau} \alpha_\ell S_\ell[a_0]. \quad (B.2)$$

The vector $\alpha$ can be viewed as the coefficients of a decomposition of $a$ into different shifts of $a_0$. This representation is not unique. For $a$ close to $S_\tau$, we can choose a particular $\alpha$ for which $\alpha_{\tau^c}$ 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 $a$ and the shifts of $a_0$: $\beta_\ell = \langle a, S_\ell[a_0] \rangle$. Since the columns of $C_{a_0}$ are the shifts of $a_0$, we can write

$$\beta = C^*_{a_0} \lambda a$$

$$= C^*_{a_0} \mu \nu C_{a_0} \alpha = : M \alpha. \quad (B.3)$$

The matrix $M$ is the Gram matrix of the truncated shifts: $M_{ij} = \langle \nu^* S_i[a_0], \nu^* S_j[a_0] \rangle$. When $\mu$ is small, the off-diagonal elements of $M$ are small. In particular, on $S_\tau$, we may take $\alpha_{\tau^c} = 0$, and $\beta \approx \alpha$, in the sense that $\beta_{\tau} \approx \alpha_{\tau}$ and the entries of $\beta_{\tau^c}$ are small. For detailed elaboration, see Appendix E.

B.2 Shifts and the Calculus of $\varphi_1$

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 $\varphi_\nu$ to the vectors $\alpha, \beta$ introduced above. Working with $\varphi_1$ 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 $\varphi_1$.

The function $\varphi_1$ has a relatively simple closed form:

$$\varphi_1(a) = -\frac{1}{2} \|S_\lambda [\bar{y} + a]\|_2^2. \quad (B.5)$$

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

Gradient: Sparsifying the Correlations $\beta$

Our goal is to understand the local minimizers of the function $\varphi_1$ over the sphere. The function $\varphi_1$ is differentiable. Clearly, any point $a$ at which its gradient (over the sphere) is nonzero cannot be a local minimizer. We first give an expression for the gradient of $\varphi_1$ over Euclidean space $\mathbb{R}^p$, and then extend it to the sphere $\mathbb{S}^{p-1}$. Using $y = a_0 + x_0$ and calculus gives

$$\nabla \varphi_1(a) = -\nu^* C_{a_0} \bar{C}_{x_0} S_\lambda \left[ \bar{C}_{x_0} C^*_{a_0} \nu a \right]$$

$$= -\nu^* C_{a_0} \bar{C}_{x_0} S_\lambda \left[ \bar{C}_{x_0} \beta \right]$$

$$= -\nu^* C_{a_0} \chi[\beta], \quad (B.6)$$
We show this rigorously below, in the proof of our main theorems. Here, we support this claim pictorially, by plotting the Riemannian gradient is large whenever $\chi[\beta]$ appears complicated. However, its effect is relatively simple:

$$\nabla \varphi_{\ell^1}(a) = \sum_{\ell} \chi[\beta]_{\ell} s_{\ell}[a_0].$$

(B.7)

The operator $\chi$ appears complicated. However, its effect is relatively simple: when $a_0$ is a long random vector, $\chi[\beta]$ acts like a soft thresholding operator on the vector $\beta$. That is,

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

(B.8)

We show this rigorously below, in the proof of our main theorems. Here, we support this claim pictorially, by plotting the $\ell$-th entry $\chi[\beta]_{\ell}$ as $\beta_\ell$ varies — see Figure 1 (middle left) and compare to Figure 1 (left). Because $\chi[\beta]$ suppresses small entries of $\beta$, the strongest contributions to $-\nabla \varphi_{\ell^1}$ in (B.7) will come from shifts $s_{\ell}[a_0]$ with large $\beta_\ell$. In particular, the Euclidean gradient is large whenever there is a single preferred shift $s_{\ell}[a_0]$, i.e., the largest entry of $\beta$ is significantly larger than the second largest entry.

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 $S^{p-1}$, which is measured by the Riemannian gradient

$$\text{grad}[\varphi_{\ell^1}](a) = P_{a^\perp} \cdot \nabla \varphi_{\ell^1}(a) = -P_{a^\perp} \cdot \sum_{\ell} \chi[\beta] s_{\ell}[a_0].$$

(B.9)

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

(i) **Negative gradient points to one particular shift**: there is a single preferred shift $s_{\ell}[a_0]$ so that the Euclidean gradient is large and

(ii) **a is not too close to any shift**: it is possible to move in the tangent space in the direction of this shift.\(^1\)

Since the tangent space consists of those vectors orthogonal to $a$, this is possible whenever $s_{\ell}[a_0]$ is not too aligned with $a$, i.e., $a$ is not too close to $s_{\ell}[a_0]$.

\(^1\)...so the projection of the Euclidean gradient onto the tangent space does not vanish.
Our technical lemma quantifies this situation in terms of the ordered entries of $\beta$. Write $|\beta_{(0)}| \geq |\beta_{(1)}| \geq \ldots$, with corresponding shifts $s_{(0)}[a_0], s_{(1)}[a_0], \ldots$. There is a strong gradient whenever $|\beta_{(0)}|$ is significantly larger than $|\beta_{(1)}|$, and $|\beta_{(1)}|$ is not too small compared to $\lambda$: in particular, when $\frac{1}{2} |\beta_{(0)}| > |\beta_{(1)}| > \frac{\lambda}{4 \log^2 p}$. In this situation, gradient descent drives $a$ toward $s_{(0)}[a_0]$, reducing $|\beta_{(1)}|, \ldots$, and making the vector $\beta$ sparser. We establish the technical claim that the (Euclidean) gradient of $\varphi_{\ell}$ sparsifies vectors in shift space in Appendix F.

**Hessian: Negative Curvature Breaks Symmetry**

When there is no single preferred shift, i.e., when $|\beta_{(1)}|$ is close to $|\beta_{(0)}|$, the gradient can be small. Similarly, when $a$ is very close to $\pm s_{(0)}[a_0]$, the gradient can be small. In either of these situations, we need to study the curvature of the function $\varphi$ to determine whether there are local minimizers.

Strictly speaking, the function $\varphi_{\ell}$ is not twice differentiable, due to the nonsmoothness of the soft thresholding operator $S_\lambda[t]$. 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 $\nabla^2 \varphi_{\ell}$, and so $\nabla^2 \varphi_{\ell}$ yields useful information about the curvature of $\varphi_{\rho}$.

As with the gradient, the Hessian is complicated, but becomes simpler when the sample size is large. The following approximation

$$\nabla^2 \varphi_{\ell}(a) \approx -\sum_\ell s_{\ell}[a_0] s_{\ell}[a_0]^* \left( \frac{\partial}{\partial \beta_\ell} x_{\ell}[\beta] \right)$$

(B.10)

with $I = \text{supp} \left( S_\lambda \left[ C_y \mu a \right] \right)$. We (formally) extend this expression to every $a \in \mathbb{R}^n$, terming $\nabla^2 \varphi_{\ell}$ the pseudo-Hessian of $\varphi_{\ell}$. 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 $\nabla^2 \varphi_{\ell}$, and so $\nabla^2 \varphi_{\ell}$ yields useful information about the curvature of $\varphi_{\rho}$.

Again, we corroborate this approximation pictorially – see Figure 2.

From this approximation, we can see that the quadratic form $v^* \nabla^2 \varphi_{\ell} v$ takes on a large negative value whenever $v$ is a shift $s_{\ell}[a_0]$ corresponding to some $|\beta_{\ell}| \geq \lambda$, or whenever $v$ is a linear combination of such shifts. In particular, if for some $j$, $|\beta_{(0)}|, |\beta_{(1)}|, \ldots, |\beta_{(j)}| \gg \lambda$, then $\varphi_{\ell}$ will exhibit negative curvature in any direction $v \in \text{span}(s_{(0)}[a_0], s_{(1)}[a_0], \ldots, s_{(j)}[a_0])$.

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

$$\text{Hess}[\varphi_{\ell}](a) = P_{a^\perp} \left( \nabla^2 \varphi_{\ell}(a) + (-\nabla \varphi_{\ell}(a), a) \cdot I \right) P_{a^\perp}.$$ (B.13)

measures the curvature of $\varphi_{\ell}$ over the sphere. The projection $P_{a^\perp}$ restricts its action to directions $v \perp a$ that are tangent to the sphere. The additional term $(-\nabla \varphi_{\ell}(a), a)$ accounts for the curvature of the sphere. This term is always positive. The net effect is that directions of strong negative curvature of $\varphi_{\ell}$ 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 $\varphi_{\ell}$ over the sphere:
We close this section by stating a key theorem, which makes the above discussion precise. We will show that if $\beta_i \gg \lambda$ the Euclidean hessian exhibits a strong negative component in the $s_i[a_0]$ direction. The Riemannian hessian exhibits negative curvature in directions spanned by $s_i[a_0]$ with corresponding $|\beta_i| \gg \lambda$ and positive curvature in directions spanned by $s_i[a_0]$ with $|\beta_i| \ll \lambda$. Middle: this creates negative curvature along the subspace $S_\tau$ and positive curvature orthogonal to this subspace. Right: our analysis shows that there is always a direction of negative curvature when $\beta_{(1)} > \frac{\|\beta(0)\|}{\lambda}$; conversely when $\beta_{(1)} \ll \lambda$ there is positive curvature in every feasible direction and the function is strongly convex.

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

$$\text{span}(s_{(0)}[a_0], s_{(1)}[a_0], \ldots, s_{(j)}[a_0])$$

of the corresponding shifts of $a_0$.

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

(iii) **Strong convexity around minimizers**: Around a minimizer $s_i[a_0]$, only a single entry $\beta_i$ is large. Any tangent direction $v \perp a$ is nearly orthogonal to the subspace $\text{span}(s_i[a_0])$, and hence is a direction of positive (Riemannian) curvature. The objective function $\varphi_\beta$ is strongly convex around the target solutions $\pm s_i[a_0]$.

Figure 2 visualizes these regions of negative and positive curvature, and the technical claim of positivity/negativity of curvature in shift space is presented in detail in Appendix G.

### B.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 $S_\tau$ can be covered by regions of negative curvature, large gradient, and regions of strong convexity containing target solutions $\pm s_i[a_0]$. Furthermore, at the boundary of this neighborhood, the negative gradient points back—retracts—toward the subspace $S_\tau$, due to the (directional) convexity of $\varphi_\beta$ away from the subspace.

To formally state the result, we need a way of measuring how close $a$ is to the subspace $S_\tau$. For technical reasons, it turns out to be convenient to do this in terms of the coefficients $\alpha$ in the representation

$$a = \sum_{i \in \tau} \alpha_i s_i[a_0] + \sum_{\ell \in \tau^c} \alpha_\ell s_\ell[a_0].$$

(B.14)

If $a \in S_\tau$, we can take $\alpha$ with $\alpha_{\tau^c} = 0$. We can view the energy $\|\alpha_{\tau^c}\|_2$ as a measure of the distance from $a$ to $S_\tau$. A technical wrinkle arises, because the representation (B.14) is not unique. We resolve this issue by
Theorem B.1

The aforementioned geometric properties hold over this set:

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

\[
\Sigma_\gamma^\theta := \bigcup_{|\tau| \leq \theta_0} \{ a : d_\alpha(a, S_\tau) \leq \gamma \}.
\]

The aforementioned geometric properties hold over this set:

**Theorem B.1 (Three subregions).** Suppose that \( y = a_0 \ast x_0 \) where \( a_0 \in S^{p-1} \) is \( \mu \)-shift coherent and \( x_0 \sim_{i.i.d.} BG(\theta) \in \mathbb{R}^n \) satisfying

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

for some constants \( c', c > 0 \). Set \( \lambda = 0.1/\sqrt{p_0} \theta \) in \( \varphi_\rho \) where \( \rho(x) = \sqrt{x^2 + \theta^2} \). There exist numerical constants \( C, c'', c''', c_1, c_4 > 0 \) such that if \( \delta \leq \frac{c'' \lambda^2 \theta}{p_0 \log^2 n} \) and \( n > Cp_0^2 \theta^{-2} \log p_0 \), then with probability at least \( 1 - c''/n \), for every \( a \in \Sigma_\gamma^\theta \), we have:

- **(Negative curvature):** If \( |\beta(1)| \geq \nu_1 |\beta(0)| \), then
  \[
  \lambda_{\min}(\text{Hess}[\varphi_\rho](a)) \leq -c_1 n \theta \lambda;
  \]

- **(Large gradient):** If \( \nu_1 |\beta(0)| \geq |\beta(1)| \geq \nu_2(\theta) \lambda \), then
  \[
  \|\text{grad}[\varphi_\rho](a)\|_2 \geq c_2 n \theta \frac{\lambda^2}{\log^2 \theta \lambda};
  \]

- **(Convex near shifts):** If \( \nu_2(\theta) \lambda \geq |\beta(1)| \), then
  \[
  \text{Hess}[\varphi_\rho](a) \succ c_3 n \theta P_{\alpha a};
  \]

- **(Retraction to subspace):** If \( \frac{\gamma}{2} \leq d_\alpha(a, S_\tau) \leq \gamma \), then for every \( \alpha \) satisfying \( a = \nu^{\prime \prime} C_{\alpha a} \zeta \), there exists \( \zeta \) satisfying \( \text{grad}[\varphi_\rho](a) = \nu^{\prime \prime} C_{\alpha a} \zeta \), such that
  \[
  \langle \zeta_\tau, \alpha_\tau \rangle \geq c_4 \| \zeta_\tau \|_2 \| \alpha_\tau \|_2 ;
  \]

- **(Local minimizers):** If \( a \) is a local minimizer,
  \[
  \min_{\sigma \in \{\pm 1\}} \| a - \sigma s_\ell(a_0) \|_2 \leq \frac{1}{2} \max \{ \mu, p_0^{-1} \},
  \]

where \( \nu_1 = \frac{4}{\lambda}, \nu_2(\theta) = \frac{1}{4 \log^2 \theta} \) and \( \gamma = \frac{c_{\text{poly}}(\sqrt{1/\theta}, \sqrt{1/\mu})}{\log^2 \theta \log^2 \mu} \cdot \frac{1}{\sqrt{p_0}} \).

**Proof.** See Appendix I.5.

The retraction property elaborated in (B.21) implies that the negative gradient at \( a \) points in a direction that decreases \( d_\alpha(a, S_\tau) \). This is a consequence of positive curvature away from \( S_\tau \). It essentially implies that the gradient is monotone in \( \alpha_\tau \)-space: choose any \( a \in S_\tau \cap S^{p-1} \), write \( \alpha \) to be its coefficient, and let \( \zeta \) be the coefficient of \( \text{grad}[\varphi_\rho](a) \). Then \( \alpha_\tau \approx 0, \zeta_\tau \approx 0 \) and

\[
\langle \zeta_\tau - \zeta_\tau, \alpha_\tau - \alpha_\tau \rangle \approx \langle \zeta_\tau - 0, \alpha_\tau - 0 \rangle = \langle \zeta_\tau, \alpha_\tau \rangle > 0.
\]

Our main geometric claim Theorem 3.1 is a direct consequence of Theorem B.1. Moreover, it suggests that as long as we can minimize \( \varphi_\rho \) within the region \( \Sigma_\gamma^\theta \), we will solve the SaS deconvolution problem.

---

6
C Provable Algorithm

In light of Theorem B.1, we introduced a two-part algorithm Algorithm 1, which first applies the curvilinear descent method to find a local minimum of \( \varphi_\rho \) within \( \Sigma^\gamma_{4p_0} \), followed by refinement algorithm that uses alternating minimization to exactly recover the ground truth. This algorithm exactly solves SAS deconvolution problem. In this section, we demonstrate Theorem C.1 and Theorem C.2, which are sufficient to prove Theorem 3.2 when jointly combined.

C.1 Minimization

There are three major issues in finding a local minimizer within \( \Sigma^\gamma_{4p_0} \). We want …

(i) **Initialization.** the initializer \( a^{(0)} \) to reside within \( \Sigma^\gamma_{4p_0} \).

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

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

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

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

\[
a^{(0)} = -P_y \varphi_\rho \left( P_{y} \left[ 0^{p_0-1}; y_0; \cdots; y_{p_0-1}; 0^{p_0-1} \right] \right) \\
= -P_y \varphi_\rho \left[ P_{y} \left( a_0 \ast x_0 \right) \right], \\
\approx -P_y \varphi_\rho \left[ P_{y} \left( a_0 \ast \bar{x}_0 \right) \right],
\]

is the normalized gradient vector from a chunk of data \( a^{(-1)} := P_{y} \left( a_0 \ast \bar{x}_0 \right) \) with \( \bar{x}_0 \) a normalized Bernoulli-Gaussian random vector of length \( 2p_0 - 1 \). Since \( \nabla \varphi_\rho \approx \nabla \varphi_\varphi \), expand the gradient \( \nabla \varphi_\varphi \) and rewrite the gradient \( \ell \left( a^{(-1)} \right) \) in shift space, we get

\[
-\nabla \varphi_\varphi \left( a^{(-1)} \right) \approx \mathbf{c}^0 \mathbf{c}_{a_0} \mathbf{C}_{x_0} \mathbf{S}_\lambda \left[ \mathbf{C}_{x_0} \mathbf{c}_{a_0} \mathbf{P}_{y} \left( a_0 \ast \bar{x}_0 \right) \right]
\]

\[
= \mathbf{c}^0 \mathbf{c}_{a_0} \mathbf{C}_{a_0} \mathbf{S}_\lambda \left[ \mathbf{C}_{a_0} \mathbf{P}_{y} \mathbf{C}_{a_0} \mathbf{x}_0 \right]
\]

\[
\approx \mathbf{c}^0 \mathbf{c}_{a_0} \mathbf{S}_\lambda \left[ \mathbf{x}_0 \right],
\]

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

\[
\max_{i \neq j} \left| \left( \mathbf{c}_{a_0}^0 \mathbf{s}_j \left[ a_0 \right], \mathbf{c}_{a_0}^0 \mathbf{s}_j \left[ a_0 \right] \right) \right| \leq \mu \ll 1. \tag{C.1}
\]

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

\[
a^{(0)} \approx P_y \mathbf{c}^0 \mathbf{c}_{a_0} \mathbf{S}_\lambda \left[ \mathbf{x}_0 \right], \tag{C.2}
\]

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

\[
a^{(0)} \in \Sigma^\gamma_{4p_0}. \tag{C.3}
\]

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

\[
\lambda \equiv \frac{c}{\sqrt{p_0 \theta}}. \tag{C.4}
\]
for some $c \in (0, 1)$.

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 $\Sigma_{4d^p_0}$ – in particular, negative curvature in symmetry-breaking directions and positive curvature away from $S_T$.

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

$$a^+ \leftarrow P_{S^p_{\rho-1}} [a - tg - t^2v] \tag{C.5}$$

For small $t$,

$$\varphi(a^+) \approx \varphi(a) + \langle g, \xi \rangle + \frac{1}{2} \xi^* H \xi. \tag{C.6}$$

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

- When $d_\alpha(a, S_T) \leq \gamma$, curvilinear steps move a small distance away from the subspace:

$$|d_\alpha (a^+, S_T) - d_\alpha (a, S_T)| \leq \frac{\gamma}{2}. \tag{C.7}$$

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

$$d_\alpha (a^+, S_T) \leq d_\alpha (a, S_T). \tag{C.8}$$

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

$$\forall k = 1, 2, \ldots, \quad a^{(k)} \in \Sigma_{4d^p_0}. \tag{C.9}$$

We conclude this section with the following theorem:

**Theorem C.1 (Convergence of retractive curvilinear search).** Suppose signals $a_0, x_0$ satisfy the conditions of Theorem B.1, $\theta > 10^3 c/p_0$ ($c > 1$), and $a_0$ is $\mu$-truncated shift coherent $\max_i \{ \langle s_i, a_0 \rangle, \epsilon^*_p s, s_j | a_0 \rangle \} \leq \mu$. Write $g = \text{grad}[\varphi_\rho](a)$ and $H = \text{Hess}[\varphi_\rho](a)$. When the smallest eigenvalue of $H$ is strictly smaller than $-\eta_c$, let $v$ be the unit eigenvector of smallest eigenvalue, scaled so $v^* g \geq 0$; otherwise let $v = 0$. Define a sequence $\{a^{(k)}\}_{k \in \mathbb{N}}$ where $a^{(0)}$ equals (?) and for $k = 1, 2, \ldots, K_1$:

$$a^{(k+1)} \leftarrow P_{S^p_{\rho-1}} [a^{(k)} - tg^{(k)} - t^2v^{(k)}] \tag{C.10}$$

with largest $t \in (0, \frac{\alpha_{\eta}}{\eta_0^2})$ satisfying Armijo steplength:

$$\varphi_\rho(a^{(k+1)}) < \varphi_\rho(a^{(k)}) - \frac{1}{2} \left( t \|g^{(k)}\|^2 + t^4 \eta_0 \|v^{(k)}\|^2 \right), \tag{C.11}$$

then with probability at least $1 - 1/c$, there exists some signed shift $\bar{a} = \pm s_i a_0$ where $i \in [\pm p_0]$ such that $\|a^{(k)} - \bar{a}\|^2 \leq \mu + 1/p$ for all $k \geq K_1 = \text{poly}(n, p)$. Here, $\eta_0 = c' n \delta \lambda$ for some $c' < c_1$ in Theorem B.1.

**Proof.** See Appendix J.2. \[\blacksquare\]
C.2 Local Refinement

In this section, we describe and analyze an algorithm which refines an estimate \( \bar{a} \approx a_0 \) of the kernel to exactly recover \((a_0, x_0)\). Set
\[
\begin{align*}
\mathbf{a}^{(0)} & \leftarrow \bar{a}, \quad \lambda^{(0)} \leftarrow C(p \theta + \log n)(\mu + 1/p), \quad I^{(0)} \leftarrow \text{supp}(S_\lambda \mathcal{C}_a^* y).
\end{align*}
\]
We alternatively minimize the Lasso objective with respect to \( a \) and \( x \):
\[
\begin{align*}
\mathbf{x}^{(k+1)} & \leftarrow \text{argmin}_x \frac{1}{2} \| a^{(k)} * x - y \|_2^2 + \lambda^{(k)} \sum_{i \notin I^{(k)}} |x_i|, \\
\mathbf{a}^{(k+1)} & \leftarrow P_{S^{p-1}} \left[ \text{argmin}_a \frac{1}{2} \| a * x^{(k+1)} - y \|_2^2 \right], \\
\lambda^{(k+1)} & \leftarrow \frac{1}{2} \lambda^{(k)}, \quad I^{(k+1)} \leftarrow \text{supp}(x^{(k+1)}).
\end{align*}
\]
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 \( a^{(k)} \) converges to one of the signed shifts of \( a_0 \) at a linear rate, in the sense that
\[
\min_{\sigma \in \pm 1, I \in [\pm p]} \| a^{(k)} - \sigma \cdot s_I(a_0) \|_2 \leq C' 2^{-k}.
\]

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 \( x_0 \) in a length-\( p \) window:
\[
\kappa_I := 6 \max \{ \theta p, \log n \},
\]
where the indexing and addition should be interpreted modulo \( n \). We will denote the support sets of true sparse vector \( x_0 \) and recovered \( x^{(k)} \) in the intermediate \( k \)-th steps as
\[
I = \text{supp}(x_0), \quad I^{(k)} = \text{supp}(x^{(k)}),
\]
then in the Bernoulli-Gaussian model, with high probability,
\[
\max_{\ell} | I \cap ([p] + \ell) | \leq \kappa_I.
\]
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 \( x_0 \) is densely occupied. In our main theorem statement, we preclude this possibility by putting an upper bound on signal length \( n \) with respect to window length \( p \) and shift coherence \( \mu \). We will assume
\[
(\mu + 1/p) \cdot \kappa_I^2 < c
\]
for some numerical constant \( c \in (0, 1) \).

Recall that (B.22) in Theorem B.1 provides that
\[
\| \bar{a} - a_0 \|_2 \leq (\mu + 1/p),
\]
which is sufficiently close to \( a_0 \) as long as (C.19) holds true. Here, we will elaborate this by showing a single iteration of alternating minimization algorithm (C.13)-(C.15) is a contraction mapping for \( a \) toward \( a_0 \).

To this end, at \( k \)-th iteration, write \( T = I^{(k)} \), \( J = I^{(k+1)} \) and \( \sigma^{(k)} = \text{sign} (x^{(k)}) \), then first observe that the solution to the reweighted Lasso problem (C.13) can be written as
\[
\begin{align*}
x^{(k+1)} &= t_J (L_J \mathcal{C}_{a^{(k)}}^* C a^{(k)} t_J)^{-1} t_J^* \left( \mathcal{C}_{a^{(k)}}^* C a_0 x_0 - \lambda^{(k)} P_{J^c} \sigma^{(k+1)} \right),
\end{align*}
\]
and the solution to least squares problem (C.14) will be
\[
a^{(k+1)} = \left( \lambda \mathbf{C}_{x}^* C_{x} \right)^{-1} \left( \lambda \mathbf{C}_{x}^* C_{x} \right) a_0 .
\] (C.23)

Here, we are going to illustrate the relationship between \(a^{(k+1)} - a_0\) and \(a^{(k)} - a_0\) using simple approximations. First, let us assume that \(a^{(k)} \approx a_0\), \(C^* a_0 \approx I\), and \(I \approx J \approx T\). Then (C.22) gives
\[
\begin{aligned}
x^{(k+1)} &\approx x_0, \\
(x^{(k+1)} - x_0) &\approx P_{I} \left( C^* a_0, x^{(k)} - C^* a_0, x_0 \right) \\
&\approx P_{I} \left[ C^* a_0, C_{x_0} \lambda (a_0 - a^{(k)}) \right],
\end{aligned}
\] (C.25)

which implies, while assuming \(C^* a_0, C_{x_0} \approx n \theta I\), that from (C.23):
\[
\begin{aligned}
(a^{(k+1)} - a_0) &\approx (n \theta)^{-1} \lambda \mathbf{C}_{x}^* C_{x} \lambda a_0 - \lambda \mathbf{C}_{x}^* C_{x} \lambda a_0 a_0 \\
&\approx (n \theta)^{-1} \lambda \mathbf{C}_{x}^* C_{x_0} \lambda (x_0 - x^{(k+1)}) \\
&\approx (n \theta)^{-1} \lambda \mathbf{C}_{x_0} C_{x} P_{I} \mathbf{C}_{x_0} \lambda (a^{(k)} - a_0),
\end{aligned}
\] (C.26)

Now since \(C^* a_0, P_{I} C_{x_0} \approx n \theta e_0 e_0^*\), this suggests that \((n \theta)^{-1} \lambda \mathbf{C}_{x_0} C_{x_0} P_{I} C_{x_0} \lambda (a^{(k)} - a_0)\) approximates a contraction mapping with fixed point \(a_0\), as follows:
\[
(n \theta)^{-1} \lambda \mathbf{C}_{x_0} C_{x_0} P_{I} C_{x_0} \lambda (a^{(k)} - a_0) \approx a_0 a_0^*.
\] (C.27)

Hence, if we can ensure all above approximation is sufficiently and increasingly accurate as the iterate proceeds, the alternating minimization essentially is a power method which finds the leading eigenvector of matrix \(a_0 a_0^*\)—and the solution to this algorithm is apparently \(a_0\). 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 C.2** (Linear rate convergence of alternating minimization). Suppose \(y = a_0 + x_0\) where \(a_0\) is \(\mu\)-shift coherent and \(x_0 \sim \mathcal{B}_G(\theta)\), then there exists some constants \(C, c, c_{\mu}\) such that if \((\mu + 1/p)\kappa_1^2 < c_{\mu}\) and \(n > C\theta^{-2}2^{2\log n}\), then with probability at least \(1 - c/n\), for any starting point \(a^{(0)}\) and \(\lambda^{(0)}, I^{(0)}\) such that
\[
\|a^{(0)} - a_0\|_2 \leq \mu + 1/p, \quad \lambda^{(0)} = 5 \kappa_1 (\mu + 1/p), \quad I^{(0)} = \text{supp} (C^* a_0 y),
\] (C.28)

and for \(k = 1, 2, \ldots\):
\[
\begin{aligned}
x^{(k+1)} &\leftarrow \text{argmin}_{x \in \mathcal{S}} \frac{1}{2} \|a^{(k)} - y\|_2^2 + \lambda^{(k)} \text{aux}_{I^{(k)}} |x_i|, \\
a^{(k+1)} &\leftarrow P_{I^{(k-1)}} \left[ \text{argmin}_{a \in \mathcal{S}} \frac{1}{2} \|a \cdot x^{(k+1)} - y\|_2^2 \right], \\
\lambda^{(k+1)} &\leftarrow \frac{1}{2} \lambda^{(k)}, \quad I^{(k+1)} \leftarrow \text{supp} (x^{(k+1)})
\end{aligned}
\] (C.29, 30, 31)

then
\[
\|a^{(k+1)} - a_0\|_2 \leq (\mu + 1/p) 2^{-k}
\] (C.32)

for every \(k = 0, 1, 2, \ldots\).

**Proof.** See Appendix K.3.

**Remark C.3.** The estimates \(x^{(k)}\) also converges to the ground truth \(x_0\) at a linear rate.
D Basic bounds for Bernoulli-Gaussian vectors

In this section, we prove several lemmas pertaining to the sparse random vector \( x_0 \sim_{i.i.d.} BG(\theta) \).

**Lemma D.1 (Support of \( x_0 \)).** Let \( x_0 \sim_{i.i.d.} BG(\theta) \) and \( I_0 = \text{supp}(x_0) \subseteq [n] \). Suppose \( n > 10\theta^{-1} \), then for any \( \varepsilon \in (0, \frac{1}{10}) \), with probability at least \( 1 - \varepsilon \) we have

\[
||I_0| - n\theta| \leq 2\sqrt{n\theta \log \varepsilon^{-1}}. \tag{D.1}
\]

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

\[
\forall t \in [2p] \setminus \{0\}, \quad \frac{1}{2}n\theta^2 \leq |I_0 \cap (I_0 + t)| \leq 2n\theta^2 \tag{D.2}
\]

where \( C \) is a numerical constant.

**Proof.** Let \( x_0 = \omega \cdot g \sim_{i.i.d.} BG(\theta) \), notice that the support of the Bernoulli-Gaussian vector \( x_0 \) is almost surely equal to the support of the Bernoulli vector \( \omega \). Applying Bernstein inequality Lemma N.4 with \( (\sigma^2, R) = (1, 1) \), then if \( n\theta > 10 \) we have

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

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

\[
|J_t| = \sum_{k \in [n]} \omega_k \omega_{k-t}, \tag{D.3}
\]

and define two subset \( J_{t1} \cup J_{t2} = J_t \), as follows:

\[
\begin{aligned}
J_{t1} &= I_0 \cap K_1, \quad K_1 := [n] \cap \{0, \ldots, t-1, 2t, \ldots, 3t-1, \ldots\} \\
J_{t2} &= I_0 \cap K_2, \quad K_2 := [n] \cap \{t, \ldots, 2t-1, 3t, \ldots, 4t-1, \ldots\}.
\end{aligned}
\tag{D.4}
\]

Here, the size of sets \( K_1, K_2 \) has two-side bounds \( 0.4n \leq (n - 2p) / 2 \leq |K_2| \leq |K_1| \leq (n + 2p) / 2 \leq 0.6n \), thus the size of sets \( J_{t1}, J_{t2} \) can be derived using Bernstein inequality Lemma N.4 with \( n > C\theta^{-2} \log p \) as

\[
\mathbb{P}\left[ \max_{t \in [2p] \setminus \{0\}} |J_{t1}| \geq n\theta^2 \right] = \mathbb{P}\left[ \max_{t \in [2p] \setminus \{0\}} \sum_{k \in K_1} \omega_k \omega_{k-t} \geq n\theta^2 \right]
\]

\[
\leq 2p \cdot \mathbb{P}\left[ \sum_{k \in K_1} \omega_k \omega_{k+1} \geq n\theta^2 \right]
\]

\[
\leq 2p \cdot \mathbb{P}\left[ \sum_{k \in K_1} \omega_k \omega_{k+1} - E \sum_{k \in K_1} \omega_k \omega_{k+1} \geq n\theta^2 - 0.6n\theta^2 \right]
\]

\[
\leq 4p \cdot \exp\left( \frac{- (0.4n\theta^2)^2}{2 \cdot 0.6n\theta^2 + 2 \cdot 0.4n\theta^2} \right) = \exp \left( \log(4p) - 0.08n\theta^2 \right)
\]

\[
\leq 1/n, \tag{D.5}
\]

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

\[
\mathbb{P}\left[ \min_{t \in [2p] \setminus \{0\}} |J_{t1}| \leq n\theta^2 / 4 \right] = \mathbb{P}\left[ \min_{t \in [2p] \setminus \{0\}} \sum_{k \in K_1} \omega_k \omega_{k-t} \leq n\theta^2 / 4 \right]
\]

\[
\leq 1/n.
\]

\[11\]
\[
\begin{align*}
&\leq 2p \cdot \mathbb{P} \left[ \sum_{k \in K_1} \omega_k \omega_{k+1} \leq n\theta^2/4 \right] \\
&\leq 2p \cdot \mathbb{P} \left[ \sum_{k \in K_1} \omega_k \omega_{k+1} - \mathbb{E} \sum_{k \in K_1} \omega_k \omega_{k+1} \leq n\theta^2/4 - 0.4n\theta^2 \right] \\
&\leq 4p \cdot \exp \left( -\frac{(0.15n\theta^2)^2}{2 \cdot 0.6n\theta^2 + 2 \cdot 0.15n\theta^2} \right) \\
&= \exp \left( \log(4p) - 0.0015n\theta^2 \right) \leq 1/n.
\end{align*}
\]

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

**Lemma D.2** (Norms of $x_0$). Let $x_0 \sim \text{i.i.d.} \; \text{BG}(\theta) \in \mathbb{R}^n$. If $n \geq 10\theta^{-1}$, then for any $\varepsilon \in (0, \frac{1}{10})$, with probability at least $1 - \varepsilon$,

\[
\left\| x_0 \right\|_1 - \sqrt{2/n} \theta \leq 2\sqrt{n\theta} \log \varepsilon^{-1}, \quad \left\| x_0 \right\|_2 - n\theta \leq 3\sqrt{n\theta} \log \varepsilon^{-1} \tag{D.6}
\]

**Proof.** To bound $\left\| x_0 \right\|_1$, using Bernstein inequality with $(\sigma^2, R) = (\theta, 1)$ and with $n\theta \geq 10$ we have

\[
\mathbb{P} \left[ \left\| x_0 \right\|_1 - \sqrt{2/n} \theta \geq 2\sqrt{n\theta} \log \varepsilon^{-1} \right] \leq 2 \exp \left( \frac{-4n\theta \log^2 \varepsilon^{-1}}{2n\theta + 4\sqrt{n\theta} \log \varepsilon^{-1}} \right) \leq \varepsilon
\]

Similarly for $\left\| x_0 \right\|_2^2$ from Gaussian moments **Lemma N.2**, we know the 2-norm $\sum_{i \in [n]} \mathbb{E} |x_{0i}|^4 = 3n\theta$ and $q$-norm $\sum_{i \in [n]} \mathbb{E} |x_{0i}|^{2p} \leq (n\theta)(2q-1)! \leq \frac{1}{2}(3n\theta)^{2q-2}q!$ for $q \geq 3$. Let $(\sigma^2, R) = (3\theta, 2)$ in Bernstein inequality form **Lemma N.4**, $n\theta \geq 10$ we have

\[
\mathbb{P} \left[ \left\| x_0 \right\|_2^2 - n\theta \geq 3\sqrt{n\theta} \log \varepsilon^{-1} \right] \leq 2 \exp \left( \frac{-9n\theta \log^2 \varepsilon^{-1}}{2(3n\theta) + 12\sqrt{n\theta} \log \varepsilon^{-1}} \right) \leq \varepsilon,
\]

completing the proof.

**Lemma D.3** (Norms of $x_0$ subvectors). Let $x_0 \sim \text{i.i.d.} \; \text{BG}(\theta) \in \mathbb{R}^n$ and $n > 10$, then with probability at least $1 - 3/n$, we have

\[
\max_{U = [2p] + j} \left\| P_U x_0 \right\|_2^2 \leq 2p\theta + 6 \left( \sqrt{p\theta} + \log n \right) \tag{D.7}
\]

and if $a_0$ is $\mu$-shift coherent and there exists a constance $c_\mu$ such that both $\theta^2 p < c_\mu$ and $\mu p^2 \theta < c_\mu$, then

\[
\max_{U = [p] + j} \left\| P_U [a_0 + x_0] \right\|_2^2 \leq p\theta + \log n. \tag{D.8}
\]

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

\[
\mathbb{P} \left[ \max_{U = [2p] + j} \left\| P_U x_0 \right\|_2^2 \geq 2p\theta + 6 \left( \sqrt{p\theta} + \log n \right) \right] \leq 2n \exp \left( -\frac{36 \left( \sqrt{p\theta} + \log n \right)^2}{6p\theta + 12 \left( \sqrt{p\theta} + \log n \right)} \right) \leq 2 \exp \left( \log n - \frac{36t^2}{6t^2 + 12t} \right) \leq \frac{2}{n}. \tag{D.9}
\]
For the second inequality, first we know calculate the expectation
\[
\mathbb{E} \| P_U [ a_0 \ast x_0 ] \|_2^2 = \mathbb{E} [ x_0^* C_{a_0}^* P_U C_{a_0} x_0 ]
\]
\[
= \theta \cdot \text{tr} ( C_{a_0}^* P_U C_{a_0} ) \| a_0 \|_2^2 + \theta \cdot \sum_{i=1}^{p-1} \| t^* s_i [ a_0 ] \|_2^2
\]
\[
= p \theta.
\]
Then apply Henson Wright inequality Lemma N.6 with \( \| C_{a_0}^* P_U C_{a_0} \|_F \leq p(1 + \mu p) \) and also \( \| C_{a_0}^* P_U C_{a_0} \|_2 = \| C_{a_0} \|_2 = 1 + \mu p \), we can derive
\[
\mathbb{P} \left[ \max_{U \in [p] \setminus J} \| P_U [ a_0 \ast x_0 ] \|_2^2 \geq p \theta + \log n \right] \leq n \exp \left( - \min \left\{ \frac{\log^2 n}{64 \theta^2 p (1 + \mu p)}, \frac{\log n}{8 \sqrt{2} \theta (1 + \mu p)} \right\} \right)
\]
\[
\leq \exp \left( \log n - \min \left\{ \frac{\log^2 n}{128 c_\mu}, \frac{\log n}{32 c_\mu} \right\} \right) \leq \frac{1}{n}
\]
when \( c_\mu < \frac{1}{32} \).
\[
\tag{D.11}
\]

**Lemma D.4 (Inner product between shifted \( x_0 \)).** Let \( x_0 \sim_{\text{i.i.d.}} \mathcal{B}(\theta) \in \mathbb{R}^n \). There exists a numerical constant \( C \) such that if \( n > C \theta^2 \log p + \theta \log^2 \theta^{-1} > 1 \), with probability at least \( 1 - 4/n \), the following two statements hold simultaneously:
\[
\max_{i \neq j \in [2p]} \langle s_i [x_0], s_j [x_0] \rangle \leq 6 \sqrt{n \theta^2 \log n};
\]
\[
\tag{D.12}
\]
and for \( x_i = |x_{0,i}| \in \mathbb{R}_+^n \) the vector of magnitudes of \( x_0 \),
\[
\max_{i \neq j \in [2p]} \langle s_i [x], s_j [x] \rangle \leq 4 n \theta^2.
\]
\[
\tag{D.13}
\]

**Proof.** We will start from proving \( \text{(D.13)} \). Write \( x = |g| \circ \omega \) where \( g / \omega \) are Gaussian/Bernoulli random vectors respectively. Let \( I_0 \) denote the support of \( \omega \) and \( t = |j - i| \) with \( 0 < t < p \). Then \( \text{(D.13)} \) can be written as summation of Gaussian r.v.s. on intersection of support set between shifts:
\[
\langle s_i [x], s_j [x] \rangle = \sum_{k \in I_0 \cap (I_0 + t)} |g_k| |g_{k-t}|
\]
\[
\tag{D.14}
\]
Define \( J_t := I_0 \cap (I_0 + t) = J_{t1} \cup J_{t2} \) same as \( \text{(D.4)} \). Notice that both \( \sum_{k \in J_{t1}} |g_k| |g_{k-t}| \) and \( \sum_{k \in J_{t2}} |g_k| |g_{k-t}| \) are sum of independent r.v.s.. We are left to consider the upper bound of \( \sum_{j \in J_{t1}} |g_j| |g_j'| \) where \( g, g' \) are independent Gaussian vectors. We condition on the following event
\[
\mathcal{E}_J := \{ \forall t \in [2p] \setminus \{0\}, \ n \theta^2 / 4 \leq |J_{t1}|, |J_{t2}| \leq n \theta^2 \},
\]
\[
\tag{D.15}
\]
which holds w.p. at least \( 1 - 2/n \) from Lemma D.1. Since \( \sum_{j \in J_{t1}} |g_j| |g_j'| \leq \| g_{J_{t1}} \|_2 \| g'_{J_{t1}} \|_2 \), we use Gaussian concentration Lemma N.3 and union bound to obtain
\[
\mathbb{P} \left[ \max_{t \in [2p] \setminus \{0\}} \left( \sum_{j \in J_{t1}} |g_j| |g_j'| > 2 |J_{t1}| \right) \leq 4p \cdot \mathbb{P} \left[ \| g_{J_{t1}} \|_2 \| g'_{J_{t1}} \|_2 - \mathbb{E} \| g_{J_{t1}} \|_2 \| g'_{J_{t1}} \|_2 > |J_{t1}| \right]
\]
\[
\leq 4p \cdot \mathbb{P} \left[ \| g_{J_{t1}} \|_2 - \mathbb{E} \| g_{J_{t1}} \|_2 > \sqrt{\| J_{t1} \| / 3} \right]
\]
\[
\leq 4p \exp \left( - \| J_{t1} \| / 9 / 2 \right) \leq 4p \exp \left( -n \theta^2 / 72 \right) \leq 1/n
\]
\[
\tag{D.16}
\]
where the last inequality is derived simply via assuming \( n = C\theta^{-2}\log p \) for some \( C > 10^4 \), such that

\[
C > 400 \cdot (4C)^{1/5} \implies C\log p > 400 \log((4C)^{1/5}p) \\
\implies C\log p > 72 \log(4Cp^5) > 72 \log(4Cp^2 \log^3 p) \\
\implies n\theta^2 > 72 \log(p \cdot 4C\theta^{-2} \log p) = 72 \log(4np).
\]

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

\[
\langle s_i[x], s_j[x] \rangle = \sum_{k \in J_{t1}} |g_k| |g_{k-i}^t| + \sum_{k \in J_{t2}} |g_k| |g_{k-t}^i| \leq 2(|J_{t1}| + |J_{t2}|) \leq 4n\theta^2. \tag{D.17}
\]

For (D.12) similarly condition on event \( E_j \), using Bernstein inequality Lemma N.4 with \((\sigma^2, R) = (1, 1)\):

\[
P \left[ \max_{t \in [2p] \setminus \{0\}} \left| \sum_{j \in J_{t1}} g_j g_{j-t} \right| > 3\sqrt{n\theta^2 \log n} \right] \leq p \cdot \exp \left( - \frac{9n\theta^2 \log n}{2 |J_{t1}| + 6 \sqrt{n\theta^2 \log n} } \right) \\
\leq p \cdot \exp \left( - \frac{9n\theta^2 \log n}{3n\theta^2} \right) \leq \frac{1}{n} \tag{D.18}
\]

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

\[
\left| \langle s_i[x_0], s_j[x_0] \rangle \right| \leq \sum_{k \in J_{t1}} |g_k| |g_{k-i}^t| + \sum_{k \in J_{t2}} |g_k| |g_{k-t}^i| \leq 6 \sqrt{n\theta^2 \log n}. \tag{D.19}
\]

Finally, both (D.17), (D.19) holds simultaneously with probability at least

\[
1 - 2/n - 1/n - 1/n = 1 - 4/n \tag{D.20}
\]

\[\square\]

**Lemma D.5 (Convolution of \( x_0 \)).** Given \( y = x_0 \ast a_0 \) where \( x_0 \sim_{i.i.d.} \mathbb{B}G(\theta) \in \mathbb{R}^n \) and \( a_0 \in \mathbb{R}^{p_0} \) 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 statements simultaneously hold:

\[
\|C_yt\|_2^2 \leq 3(1 + \mu p)n\theta \tag{D.21}
\]

and for all \( J \subseteq [n] \),

\[
\|P_J C_yt\|_2 \leq 14 |J| \left( (1 + \mu p)(p\theta + \log n) \right) \tag{D.22}
\]

**Proof.** Given any \( a \in \mathbb{S}^{p-1} \), write \( \beta = C_{a0}t \) where \( |\beta| \leq 2p \). Apply \( \|x_0\|_2^2 \leq 2n\theta \) from **Lemma D.2** by choosing \( \varepsilon = 1/n \), also \( \|\langle s_i[x_0], s_j[x_0]\rangle\| \leq 6 \sqrt{n\theta^2 \log n} \). From **Lemma D.4** we get:

\[
\|C_yt a\|_2^2 = \|C_{x_0} \beta\|_2^2 \leq \|\beta\|_2^2 \|x_0\|_2^2 + \sum_{i \neq j \in [p]} |\beta_i \beta_j \langle s_i[x_0], s_j[x_0]\rangle| \\
\leq |\beta|^2 \|x_0\|_2^2 + |\beta|_1^2 \max_{i \neq j \in [p]} \|\langle s_i[x_0], s_j[x_0]\rangle\| \\
\leq |\beta|^2 \cdot 2n\theta + \|\beta\|_2^2 \cdot 6 \sqrt{n\theta^2 \log n} \leq 3 |\beta|^2 n\theta
\]

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 \( \|P_j C_y t a\|_2^2 \): Simply apply Lemma D.3 and utilize norm bound of \( \|\beta\|_2^2 \), with probability at least \( 1 - 2/|n| \) we have:
\[
\|P_j C_y t a\|_2^2 = \sum_{i \in J} |(s_i[x_0], \beta)|^2 \leq |J| \max_{j \in [n]} \|P_j[x_0]\|_2^2 \|\beta\|_2^2 \leq |J| \cdot 14 (p\theta + \log n) \cdot \|\beta\|_2^2
\]

Finally apply Lemma E.4 and Gershgorin disc theorem obtain
\[
\|\beta\|_2^2 = \|C^*_a t a\|_2^2 \leq \|C^*_a\|_2^2 = \sigma_{\text{max}} (M) \leq 1 + \mu p. \tag{D.23}
\]

Remark D.6. When \( a_0 \) is a basis vector \( e_0 \), the result of Lemma D.5 gives upper bound of \( \|C_{x_0}\|_2 < 3n\theta \), whose lower bound can be derived similarly with \( \|C_{x_0}\|_2 \geq \frac{2}{3} n\theta \).

E Vectors in shift space

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

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

Definition E.1 (Sparsity-coherence condition). Let \( a_0 \in \mathbb{S}^{p_0 - 1} \) with shift coherence \( \mu \). We say that \( (a_0, \theta, |\tau|) \) satisfies the sparsity-coherence condition \( \text{SCC}(c_\mu) \) with constant \( c_\mu \) if
\[
\theta \in \left[ \frac{1}{p}, \frac{c_\mu}{4 \max \{|\tau|, \sqrt{p}\}} \right], \quad \mu \cdot \max \left\{ |\tau|^2, p^2 \theta^2 \right\} \cdot \log^2 \theta^{-1} \leq \frac{c_\mu}{4},
\]
where \( p = 3p_0 - 2 \).

Lemma E.2 (\( d_\alpha \) is a seminorm). For every solution subspace \( S_\tau \), the function \( d_\alpha (\cdot, S_\tau) : \mathbb{R}^p \to \mathbb{R}_+ \) defined as
\[
d_\alpha (a, S_\tau) = \inf \{ \|\alpha_{\tau^c}\|_2 | a = \mu^* C_a \alpha \}.
\]
is a seminorm, and for all \( a \in S_\tau \), \( d_\alpha (a, S_\tau) = 0 \).

Proof. It is immediate from definition that \( d(\cdot, S_\tau) \) is nonnegative and \( S_\tau \subseteq \{ a : d_\alpha (a, S_\tau) = 0 \} \). Subadditivity can be shown from simple norm inequalities and our definition of \( d_\alpha \), for all \( a_1, a_2 \) we have
\[
d_\alpha (a_1 + a_2, S_\tau) = \inf \{ \|\alpha_{\tau^c}\|_2 | a_1 + a_2 = \mu^* C_a \alpha \}
= \inf \{ \|\alpha_{1_{\tau^c}} + \alpha_{2_{\tau^c}}\|_2 | a_1 = \mu^* C_a \alpha_1, \quad a_2 = \mu^* C_a \alpha_2 \}
\leq \inf \{ \|\alpha_{1_{\tau^c}}\|_2 + \|\alpha_{2_{\tau^c}}\|_2 | a_1 = \mu^* C_a \alpha_1, \quad a_2 = \mu^* C_a \alpha_2 \}
= \inf \{ \|\alpha_{1_{\tau^c}}\|_2 | a_1 = \mu^* C_a \alpha_1 \} + \inf \{ \|\alpha_{2_{\tau^c}}\|_2 | a_2 = \mu^* C_a \alpha_2 \}
= d_\alpha (a_1, S_\tau) + d_\alpha (a_2, S_\tau).
\]
Similarly the absolute homogeneity, for any \( c \in \mathbb{R} \):
\[
d_\alpha (c \cdot a, S_\tau) = \inf \{ \|c \cdot \alpha_{\tau^c}\|_2 | c \cdot a = \mu^* C_a \alpha \} = \inf \{ \|c \cdot \alpha_{\tau^c}\|_2 | a = \mu^* C_a \alpha \}
\]
which completes the proof that \( d_\alpha \) is a seminorm.

**Definition E.3 (Widened subspace).** For subspace \( S_\tau \) let

\[
\mathcal{R}(S_\tau, \gamma(c_\mu)) := \{ a \in \mathbb{S}^{p-1} \mid d_\alpha(a, S_\tau) \leq \gamma \}
\]

(S E.3) denote its widening by \( \gamma \), in the seminorm \( d_\alpha \).

Our analysis works with a specific choice of width \( \gamma(c_\mu) \), which depends on the problem parameters \( a_0, \theta, |\tau| \) and a constant \( c_\mu \), via

\[
\gamma(c_\mu) = \frac{c_\mu}{4\log_\theta|\tau|} \min \left\{ \frac{1}{\sqrt{|\tau|}}, \frac{1}{\sqrt{\mu p}}, \frac{1}{\mu p \sqrt{\theta |\tau|}} \right\}
\]

(E.4)

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

\[
\begin{cases}
M_{ii} = 1 & i \in [-p_0 + 1, p_0 - 1] = [\pm p_0], \\
0 \leq M_{ii} \leq 1 & i \in [-2p_0 + 2, -p_0 ] \cup [p_0, 2p_0 - 2], \\
M_{ii} = 0 & \text{otherwise},
\end{cases}
\]

(E.5)

and the off-diagonal entries satisfy

\[
\begin{cases}
|M_{ij}| \leq \mu & 0 < |i - j| < p_0, \ {i \in [-p_0 + 1, p_0 - 1]} \cup \{j \in [-p_0 + 1, p_0 - 1]\} \\
|M_{ij}| < 1 & \{i, j \in [-2p_0 + 2, -p_0]\} \cup \{i, j \in [p_0, 2p_0 - 2]\} \\
0 & \text{otherwise}
\end{cases}
\]

(E.6)

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

\[
\begin{cases}
1 - \mu |\tau| \leq \sigma_{\min}(\mathbf{u}^* \mathbf{M} \tau) \leq \sigma_{\max}(\mathbf{u}^* \mathbf{M} \tau) \leq 1 + \mu |\tau| \\
\sigma_{\max}(\mathbf{u}^* \mathbf{M} \tau^c) \leq \mu \sqrt{|\tau|} \\
\sigma_{\max}(\mathbf{u}^* \mathbf{M} \tau^c) \leq 1 + \mu p
\end{cases}
\]

(E.7)

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

\[
\begin{bmatrix}
\mathbf{J} & \mu \cdot 1 & \mu \cdot 1 \\
\mu \cdot 1 & \mathbf{I} + \mu \cdot \mathbf{1}_o & \mu \cdot 1 \\
\mu \cdot 1 & \mu \cdot 1 & \mu \cdot 1
\end{bmatrix}
\]

(E.8)
Here, the center row/column vector is indexed at 0, the matrices $J, I, 1$ and $1_o$ are square and of size $(p_0 - 1)^2$. Among which, $I$ is the identity matrix, $1$ is the ones matrix whereas $1_o$ has all off diagonal entries equal 1. Also $|J|$ has property $|J_{i,j}| < 1$ for all $i, j$.

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

$$\sigma_{max} (\tau^*, M\tau^*) \leq \mu \sqrt{(2p_0 - 1) |\tau|} < \mu \sqrt{p |\tau|}.$$  \hspace{1cm} (E.9)

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

$$\sigma_{max} (\tau^*, M\tau^*) \leq \|C_{a_0}^* \tau^* C_{a_0}\|_2 = \|\tau^* C_{a_0} C_{a_0}^* \|_2 \leq 1 + \mu p$$ \hspace{1cm} (E.10)

where the third equality is derived via commutativity of convolution.

Lemma E.5 (Shift space vectors in widened subspace). Let $(a_0, \theta, |\tau|)$ satisfy the sparsity-coherence condition $\text{SCC}(c_\mu)$. Then for every $a \in \mathcal{R}(S_\tau, \gamma(c_\mu))$, every $\alpha$ satisfying $a = \tau^* C_{a_0} \alpha$ and $\|\alpha_\tau\|_2 \leq \gamma(c_\mu)$ has

$$\|\alpha_\tau\|_2 - 1| \leq c_\mu.$$ \hspace{1cm} (E.11)

moreover, $\beta = C_{a_0}^* a$ satisfies

$$1 - 3 c_\mu \leq \|\beta_\tau\|_2^2 \leq 1 + \frac{c_\mu}{|\tau| \log^2 \theta - 1},$$ \hspace{1cm} (E.12)

$$\|\beta_\tau\|_\infty \leq \frac{c_\mu}{|\tau| \log^2 \theta - 1} \hspace{1cm} (E.13)$$

$$\|\beta_\tau\|_2 \leq \frac{c_\mu}{|\tau| \log^2 \theta - 1} \min \left\{ \sqrt{\theta}, \gamma(c_\mu) \right\}.$$ \hspace{1cm} (E.14)

Proof. Write $-1/\log \theta = \theta_{\log}$ and $\gamma = \gamma(c_\mu)$ for convenience. First, by using bounds on $\gamma$ in (E.4) and $\mu |\tau| < c_\mu / 4$ and (E.15), we obtain:

$$\begin{cases}
\gamma \cdot \sqrt{1 + \mu p} \leq \gamma (1 + \sqrt{\mu p}) \leq c_\mu \theta_{\log}^2 / 2 \\
\gamma \cdot \sqrt{1 + \mu^2 p} \leq \gamma (1 + \sqrt{\mu^2 p}) \leq \frac{c_\mu \theta_{\log}^2}{4} \left( \frac{1}{\sqrt{|\tau|}} + \sqrt{\mu} \right) \leq c_\mu \theta_{\log}^2 / 2 \sqrt{|\tau|} \\
\gamma \cdot \mu \sqrt{p |\tau|} \leq \gamma \cdot \sqrt{\mu^2 p} \cdot \sqrt{\mu |\tau|} \leq c_\mu \theta_{\log}^2 / 4
\end{cases}$$ \hspace{1cm} (E.15)

Let $a = \tau^* C_{a_0} \alpha$ with $\|\alpha_\tau\|_2 < \gamma$. Utilize properties of $\tau^* C_{a_0}$ from Lemma E.4 and $\mu |\tau| < c_\mu / 4$ and (E.15), we have:

$$\|\alpha_\tau\|_2 \geq \|\tau^* C_{a_0} a\|_2^{-1} (\|a\|_2 - \|\tau^* C_{a_0} \alpha_\tau\|_2)$$

$$\geq \|\tau^* C_{a_0} a\|_2^{-1} (1 - \|\tau^* C_{a_0}\|_2^2)$$

$$\geq \frac{1}{\sqrt{1 + \mu |\tau|}} \left( 1 - \gamma \cdot \sqrt{1 + \mu p} \right) \geq \frac{1 - c_\mu / 2}{\sqrt{1 + c_\mu / 4}} \geq 1 - c_\mu,$$ \hspace{1cm} (E.16)

and similarly, the upper bound can be derived as:

$$\|\alpha_\tau\|_2 \leq \frac{1}{\sqrt{1 - \mu |\tau|}} \left( 1 + \gamma \cdot \sqrt{1 + \mu p} \right) \leq \frac{1 + c_\mu / 2}{\sqrt{1 - c_\mu / 4}} \leq 1 + c_\mu.$$ \hspace{1cm} (E.17)
The bound of $\|\beta_\tau\|^2_2$ can be simply obtained using $\mu |\tau| < c_\mu / 4$ and $\gamma$ bound from (E.15) as:

$$\|\beta_\tau\|^2_2 \leq \sigma_{\max}^2 (\ell^*_\tau C_{a_0} \ell) \leq 1 + \mu |\tau| \leq 1 + \frac{c_\mu \theta^2_{\log}}{|\tau|} \quad (E.18)$$

$$\|\beta_\tau\|^2_2 \geq (\sigma_{\min} (\ell^*_\tau M \ell_\tau) \|\alpha_\tau\|_2 - \sigma_{\max} (\ell^*_\tau M \ell_\tau) \|\alpha_\tau\|_2)^2 \geq \left( (1 - \mu |\tau|) (1 - c_\mu) - \sqrt{p |\tau| \cdot \gamma} \right) \geq 1 - 3c_\mu. \quad (E.19)$$

As for the upper bound of $\|\beta_\tau\|_\infty$, follow from (E.15), we have:

$$\|\beta_\tau\|_\infty \leq \|\ell^*_\tau M \alpha_\tau\|_\infty + \|\ell^*_\tau M \alpha_\tau\|_\infty \leq \mu \sqrt{|\tau|} \|\alpha_\tau\|_2 + \sqrt{1 + \mu^{2p}} \|\alpha_\tau\|_2 \leq \frac{c_\mu \theta^2_{\log} (1 + c_\mu)}{|\tau|} + \gamma \sqrt{1 + \mu^{2p}} \leq \frac{c_\mu \theta^2_{\log}}{\sqrt{|\tau|}} \quad (E.20)$$

the bound for $\|\beta_\tau\|_2$ requires two inequalities, we know

$$\|\beta_\tau\|_2 \leq \|\ell^*_\tau M \alpha_\tau\|_2 + \|\ell^*_\tau M \alpha_\tau\|_2 \leq \mu \sqrt{p |\tau|} \|\alpha_\tau\|_2 + (1 + p \mu) \|\alpha_\tau\|_2, \quad (E.21)$$

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

$$\leq \frac{\mu \sqrt{p \theta} |\tau|^{3/2}}{\theta |\tau|} (1 + c_\mu) + \gamma \sqrt{\mu \theta |\tau|} \leq \frac{2c_\mu \theta^2_{\log} + c_\mu \theta^2_{\log} + c_\mu \theta^2_{\log}}{4 \sqrt{\theta |\tau|}} \leq \frac{c_\mu \theta^2_{\log}}{\sqrt{\theta |\tau|}}. \quad (E.22)$$

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

$$\leq \frac{\gamma}{\theta |\tau|} \cdot \frac{\mu \sqrt{p \theta} |\tau|^{3/2}}{\theta |\tau|} (1 + c_\mu) + \gamma \mu \theta |\tau| \leq \frac{\gamma}{\theta |\tau|} \cdot \frac{4 \mu \sqrt{p \theta} |\tau|^{3/2}}{c_\mu \theta^2_{\log}} \cdot \max \left\{ \sqrt{|\tau|}, \sqrt{\mu p \theta |\tau|} \right\} \leq \frac{\gamma}{\theta |\tau|} \cdot \frac{\mu \sqrt{p \theta} |\tau|^{3/2}}{c_\mu \theta^2_{\log}} \cdot \max \left\{ \mu |\tau|^2 \cdot \sqrt{\mu p \theta |\tau|} \cdot \sqrt{\mu |\tau|} \right\} \leq \frac{\gamma}{\theta |\tau|} \cdot \left( \frac{c_\mu \theta^2_{\log}}{4} \cdot \frac{c_\mu \theta^2_{\log}}{4} \cdot \frac{c_\mu \theta^2_{\log}}{4} \right) \leq \frac{c_\mu \theta^2_{\log}}{\theta |\tau|}, \quad (E.23)$$

which completes the proof.

Corollary E.6 (|⟨β_τ, x_0, τ⟩| is small). Given $x_0 \sim_{i.i.d.} \text{BG}(\theta)$ in $\mathbb{R}^n$ and $|\tau|$, $c_\mu$ such that $(a_0, \theta, |\tau|)$ satisfies the sparsity-coherence condition SCC($c_\mu$). Write $\lambda = c_\lambda / \sqrt{|\tau|}$ with some $c_\lambda \geq 1 / 3$, then if $c_\mu \leq c_\mu / 25$,

$$\mathbb{P} \left[ \sum_{i \in \tau^c} \beta_i x_{0i} > \frac{\lambda}{10} \right] \leq 2 \theta, \quad \mathbb{P} \left[ \sum_{i} \beta_i x_{0i} > \frac{\lambda}{10} \right] \leq \theta |\tau| + 2 \theta. \quad (E.24)$$
Proof. We bound tail probability of the first result with Gaussian moments Lemma N.2 and Bernstein inequality Lemma N.4. Via Hölder’s inequality, \( \sum_{i \in \tau} E(\beta_i x_i)^q = E x_i^q \| \beta_{\tau} \|^q \leq (q-1)! \| \beta_{\tau} \|^2 \| \beta_{\tau} \|^q \), thus

\[
P \left[ \sum_{i \in \tau} \beta_i x_{i0} > \lambda/10 \right] \leq 2 \exp \left( \frac{-(\lambda/10)^2}{2 \theta \| \beta_{\tau} \|^2 + 2(\lambda/10) \| \beta_{\tau} \|_\infty} \right) \quad (E.25)
\]

Write \( \theta \log \frac{c_\mu}{\sqrt{|\tau|}} \), Lemma E.5 implies when \( c_\mu \leq \frac{c_\mu}{25} \), we have \( \theta \| \beta_{\tau} \|^2 \leq \frac{c_\mu \theta \| \beta_{\tau} \|^2}{|\tau|^2} \leq \frac{\theta \log \lambda^2}{25} \) and \( \| \beta_{\tau} \|_\infty \leq \frac{c_\mu \theta \log \lambda}{\sqrt{|\tau|}} \), therefore,

\[
(E.25) \leq 2 \exp \left( \frac{-\lambda^2/100}{2 \theta \log \lambda^2/25 + 2(\theta \log \lambda/25) \cdot (\lambda/10)} \right)
\leq 2 \exp (\log \theta) \leq 2 \theta \quad (E.26)
\]

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

\[
P \left[ \sum_{i} \beta_i x_{i0} > \lambda/10 \right] \leq P [ \| \beta_{\tau}^* x_{\tau} \| + \| \beta_{\tau}^* x_{\tau} \| > \lambda/10] 
\leq P [x_{\tau} \neq 0] + P [x_{\tau} = 0] \cdot P \| \beta_{\tau}^* x_{\tau} \| > \lambda/10] 
\leq \theta |\tau| + 2 \theta. \quad (E.27)
\]

Corollary E.7 (\( \langle \beta_{\tau \setminus \{0\}}, x_{0, \tau \setminus \{0\}} \rangle \) is small near shifts). Suppose that \( x_0 \sim \text{i.i.d. BG}(\theta) \) in \( \mathbb{R}^n \), and \( |\tau| \), \( c_\mu \) such that \( (a_0, \theta, |\tau|) \) satisfies the sparsity-coherence condition SCC\((c_\mu)\), then if \( c_\mu \leq \frac{1}{10} \), for any \( a \) such that \( |\beta_{\{1\}}| \leq \frac{\lambda}{4 \log \sigma_{\tau}} \), we have

\[
P \left[ \sum_{i \in \tau \setminus \{0\}} \beta_i x_{i0} > \frac{2 \lambda}{5} \right] \leq 2 \theta. \quad (E.28)
\]

Proof. For the last tail bound, write \( x = \omega \circ g \). Wlog define \( \beta_0 \) be the largest correlation \( \beta_{\{0\}} \), define random variables \( s' = \langle \beta_{\tau \setminus \{0\}}, x_{\tau \setminus \{0\}} \rangle \). Firstly most of the entries of \( x_\tau \) would be zero since via Bernstein inequality with \( \theta |\tau| < 0.1 \):

\[
P \left[ \sum_{i \in \tau} \omega_i > \log \theta^{-1} \right] \leq P \left[ \sum_{i \in \tau} \omega_i > \theta |\tau| + 0.9 \log \theta^{-1} \right] 
\leq \exp \left( \frac{-0.9^2 \log^2 \theta^{-1}}{2 (\theta |\tau| + 0.9 \log \theta^{-1}/3)} \right) \leq \theta \quad (E.29)
\]

thus with probability at least \( 1 - \theta \), we can write \( s' \) as a Gaussian r.v. with variation bounded as \( \mathbb{E} s'^2 \leq \mathbb{E} \left[ \sum_{i=1}^{\log \theta^{-1}} \beta_i g_i \right]^2 = \log \theta^{-1} \beta_{\{1\}}^2 \), then via Gaussian tail bound Lemma N.1:

\[
P \left[ |s'| > 0.4 \lambda \right] \leq P \left[ |g| > \frac{0.4 \lambda}{\sqrt{\log \theta^{-1} |\beta_{\{1\}}|}} \right] + P \left[ \sum_{i \in \tau} \omega_i > \log \theta^{-1} \right] 
\leq 2 \exp (-1.2 \log \theta^{-1}) + \theta \leq 2 \theta. \quad (E.30)
\]
The following lemma describes the behavior of the summands in the above expression:

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

\[
\mathbb{E}_\eta \left[ g \mathcal{S}_\lambda [b \cdot g + s] \right] = b \left( 1 - \text{erf}_b(\lambda, s) \right),
\]

where

\[
\text{erf}_b(\lambda, s) = \frac{1}{2} \text{erf} \left( \frac{\lambda + s}{\sqrt{2}|b|} \right) + \frac{1}{2} \text{erf} \left( \frac{\lambda - s}{\sqrt{2}|b|} \right).
\]

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

\[
\sigma \nu_2 \mathcal{S}_\lambda [b] \leq \sigma \mathbb{E}_\eta \left[ g \mathcal{S}_\lambda [b \cdot g] \right] \leq \sigma \nu_2 [b] + \varepsilon
\]

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

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

\[
\int_{-\infty}^{t} t' f(t')dt' = -f(t), \quad \int_{-\infty}^{t} t^2 f(t')dt' = \frac{1}{2} \text{erf} \left( \frac{t}{\sqrt{2}} \right) - tf(t)
\]

Integrating, we obtain

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

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

\[
\int_{t \geq \frac{\lambda}{b}} \left( bt^2 - Lt \right) f(t)dt = b \left[ \frac{1}{2} \text{erf} \left( \frac{L}{\sqrt{2}b} \right) + \frac{L}{b} \int f \left( \frac{L}{b} \right) \right] = \frac{b}{2} - \frac{b}{2} \text{erf} \left( \frac{L}{\sqrt{2}b} \right),
\]

and similarly for the second summand, which gives

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

20
For \( b < 0 \), alternatively we have

\[
\mathbb{E} \left[ gS_{\lambda} [-|b| \cdot g + s] \right] = -\mathbb{E}[gS_{\lambda} [b \cdot g - s]] = -|b|(1 - \text{erf}(\lambda, -s)) = b(1 - \text{erf}(\lambda, s)),
\]

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

\[
\min \left\{ 1 - \varepsilon, \frac{1 - \varepsilon}{\sqrt{\log(1/\varepsilon)}} \right\} \leq \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \leq \frac{2x}{\sqrt{\pi}} \tag{F.4}
\]

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

\[
\text{erf}(x) \geq 1 - e^{-x^2} \geq 1 - e^{\log \varepsilon} = 1 - \varepsilon
\]

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

\[
\text{erf}(x) \geq \frac{\text{erf}(T) - \text{erf}(0)}{T - 0} x + \text{erf}(0) \geq \frac{1 - \varepsilon}{\sqrt{\log(1/\varepsilon)}} x.
\]

Lastly plug (F.4) into (F.1) and apply condition \( |b| \leq 1 \) and \( \varepsilon < 1/4 \) we have

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

which completes the proof.

This lemma establishes when \( x_0 \) is separated, then \( \chi \) is soft thresholding operator on \( \beta \) with threshold about \( \lambda/2 \). This phenomenon extends beyond the separated case, as long as when \( x_0 \) is sufficiently sparse (when Definition E.1 holds). Recall that \( \chi : \mathbb{R}^n \to \mathbb{R}^n \) is defined as

\[
\chi[\beta] = \bar{C}_{x_0}S_{\lambda} \left[ \tilde{C}_{x_0}\beta \right]. \tag{F.5}
\]

The following lemma bounds its expectation:

**Lemma F.2** (Expectation of \( \chi(\beta) \)). Let \( x_0 \sim_{\text{i.i.d.}} \text{BG}(\theta) \) and \( \lambda > 0 \), then for every \( \alpha \in \mathbb{S}^{p-1} \) and every \( i \in [n], \) define the operator \( \chi \) as in (F.5), then

\[
n^{-1}\mathbb{E}_{\chi}[\beta]_i = \theta \beta_i (1 - E_{s_i} \text{erf}_{\beta_i}(\lambda, s_i)) \tag{F.6}
\]

where \( s_i = \sum_{j \neq i} \beta_j x_{0j} \). Suppose \( (\alpha_0, \theta, |\tau|) \) satisfies the sparsity-coherence condition \( \text{SCC}(c_\mu) \) and \( \lambda = c_\mu / \sqrt{|\tau|} \)

for some \( c_\mu > 1/5 \) and \( \sigma_i = \text{sign}(\beta_i) \), then there exists some numerical constant \( \tau \) such that if \( c_\mu \leq \tau \) then for every \( \alpha \in \mathbb{R}(S_\tau, \gamma(c_\mu)) \) and every \( i \in [n], \) (F.6) has upper bound

\[
\sigma_i n^{-1} \mathbb{E}_{\chi}[\beta]_i \leq \sigma_i n^{-1} \mathbb{E}_{\chi}[|\beta|_i] := \begin{cases} 4\theta^2 |\tau| |\beta_i|, & |\beta_i| < \nu_1 \lambda \\ \theta (|\beta_i| - \nu_1 \lambda/2), & |\beta_i| \geq \nu_1 \lambda \end{cases} \tag{F.7}
\]

and lower bound

\[
\sigma_i n^{-1} \mathbb{E}_{\chi}[\beta]_i \geq \sigma_i n^{-1} \mathbb{E}_{\chi}[|\beta|_i] := \theta S_{\nu_2 \lambda} ||\beta_i||, \tag{F.8}
\]

where \( \nu_1 = 1/ \left( 2 \sqrt{\log \theta^{-1}} \right), \nu_2 = \sqrt{2/\pi}. \)
Thus wlog let us consider pieces. and a random variable \( \chi \) gives like a soft thresholding operation, and for small \( \theta \), it reduces \( |\beta| \) by multiplying a very small number \( 4\theta |\tau| \ll 1 \). We rigorously prove this segmentation of \( \chi \) operator as follows:

**Proof.** First, since \( s_i[x_0] \equiv_d s_j[x_0] \),

\[
\chi[\beta]_i = e_i^\ast \tilde{C}_{x_0} S_{\lambda} \left[ \tilde{C}_{x_0} \beta \right] = \left\langle s_{-i}[x_0], S_{\lambda} \left[ x_0 * \tilde{\beta} \right] \right\rangle = \chi[s_{-i}[\beta]]_j
\]

Thus wlog let us consider \( i = 0 \) and write \( x \) as \( x_0 \). The random variable \( \chi[\beta]_0 \) can be written sum of random variables as:

\[
\chi[\beta]_0 = \left\langle x, S_{\lambda} \left[ \beta_0 x_0 + \sum_{\ell \neq 0} \beta_\ell s_{-\ell}[x] \right] \right\rangle = \sum_{j \in [n]} x_j S_{\lambda} \left[ \beta_0 x_j + \sum_{\ell \neq 0} \beta_\ell x_{j+\ell} \right],
\]

and a random variable \( Z_j(\beta) \) is defined as

\[
Z_j(\beta) = x_j S_{\lambda} \left[ \beta_0 x_j + \sum_{\ell \in [\pm 1]} \beta_\ell x_{j+\ell} \right],
\]

gives \( \chi[\beta]_0 = \sum_{j \in [n]} Z_j(\beta) \) as sum of r.v.s. of same distribution and thus \( n^{-1} \mathbb{E} \chi[\beta]_0 = \mathbb{E} Z_0(\beta) \). Define a random variable \( s_0 = \sum_{\ell \neq 0} \beta_\ell x_{\ell} \), which is independent of \( x_0 \). From **Lemma F.1**, we can conclude

\[
n^{-1} \mathbb{E} \chi[\beta]_0 = \mathbb{E}_{x_0, s_0} x_0 S_{\lambda} \left[ \beta_0 x_0 + s_0 \right] = \theta \beta_0 (1 - \mathbb{E}_{s_0} \text{erf}_{\beta_0}(\lambda, s_0))
\]

so that \((\ref{F.6})\) holds for \( i = 0 \), and hence for all \( i \).

1. (Upper bound of \( \mathbb{E}Z \)) Wlog assume \( \beta_0 \geq 0 \) and write \( Z = Z_0 \). We derive the upper bound on \( \mathbb{E}Z \) in two pieces.

---

**Figure 3:** A numerical example of \( \mathbb{E} \chi[\beta]_i \). We provide figures for the expectation of \( \chi \) when entries of \( x_0 \) are \( 2p \)-separated. Left: the yellow line is the function \( \beta_i \rightarrow \beta_i (1 - \text{erf}_{\beta_i}(\lambda, 0)) \) derived from (\ref{F.1}), and the blue/red lines are its upper/lower bound (\ref{F.3}) utilized in the analysis respectively. Right: functions of \( \beta_i \rightarrow \beta_i (1 - \text{erf}_{\beta_i}(\lambda, 0)) \) with different \( \lambda \), the section of function of \( \beta_i > \nu_2 \lambda \) are close to linear.
(1). First, since $\mathbb{E} x_0 S_\lambda \left[ 0 \cdot x_0 + s_0 \right] = 0$, we have
\[
\mathbb{E} Z(\beta) \leq \beta_0 \sup_{\beta \in [0,\beta_0]} \frac{d}{d\beta} \mathbb{E}_{x_0,s_0} x_0 S_\lambda \left[ \beta x_0 + s_0 \right]
\]
\[
= \theta \beta_0 \sup_{\beta \in [0,\beta_0]} \frac{d}{d\beta} \int_{|\beta g + s_0| > \lambda} g \left( \beta g + s_0 - \text{sign}(\beta g + s_0) \cdot \lambda \right) d\mu(g) d\mu(s_0)
\]
\[
= \theta \beta_0 \sup_{\beta \in [0,\beta_0]} \int_{|\beta g + s_0| > \lambda} g^2 \mathbf{1}_{|\beta g + s_0| > \lambda}
\]
\[
\leq \theta \beta_0 \sup_{\beta \in [0,\beta_0]} \int_{|\beta g + s_0| > \lambda} g^2 \left( \mathbf{1}_{|\beta g| > \frac{\lambda}{|1|}} + \mathbf{1}_{|s_0| > \frac{\lambda}{|s_0|}} \right)
\]
\[
\leq \theta \beta_0 \left( (E g^6)^{1/3} P \left[ |\beta_0 g| > (9\lambda/10)^{2/3} \right. \right. \left. \left. + P \left[ |s_0| > \lambda/10 \right] \right) \right) \quad (F.11)
\]
We bound the tail probability of $s_0$ using Corollary E.6 where
\[
P \left[ |s_0| > \lambda/10 \right] \leq \mathbb{P} \left[ |\sum_i \beta_i x_i| > \lambda/10 \right] \leq \theta |\tau| + 2\theta \leq 3 \theta |\tau| \quad \text{(F.12)}
\]
On the other hand, the first term in (F.11) can be derived by pdf of Gaussian r.v. Lemma N.1 as:
\[
(E g^6)^{1/3} P \left[ |\beta_0 g| > (9\lambda/10)^{2/3} \right] \leq \sqrt{15} \left( \frac{10\beta_0}{9\lambda \sqrt{2\pi}} \right)^{2/3} \exp \left( -\frac{\lambda^2}{4\beta_0^2} \right)
\]
\[
\leq \frac{3}{2} \left( \frac{\beta_0}{\lambda} \right)^{2/3} \exp \left( -\frac{\lambda^2}{4\beta_0^2} \right) \quad (F.13)
\]
Combine (E.26), (F.13), when $\beta_0 < \nu_1 \lambda$, we know $e^{\frac{-\lambda^2}{4\beta_0^2}} \leq e^{\log \theta \leq \theta |\tau|}$. The first type of upper bound $\mathbb{E} Z$ is derived as
\[
\forall \beta_0 \in [0,\nu_1 \lambda], \quad \mathbb{E} Z(\beta) \leq \beta_0 \left( \frac{3}{2} \nu_1^{2/3} \exp \left( -\frac{\lambda^2}{4\beta_0^2} \right) \right) \leq \theta^2 |\tau| \beta_0 \quad (F.14)
\]
(2). The second type of upper bound can be derived directly from Lemma F.1:
\[
\mathbb{E} Z(\beta) \leq \mathbb{E}_{x_0} x_0 S_\lambda \left[ \beta_0 x_0 + s_0 \right] \leq \mathbb{E}_{x_0} x_0 S_\lambda \left[ \beta_0 x_0 + \mathbb{E}_{s_0} |x_0| \mathbb{E}_{s_0} |s_0| \right]
\]
\[
\leq \theta \cdot \left( S_{\nu_1 \lambda} [\beta_0] + \varepsilon + \sqrt{2/\pi \cdot \mathbb{E} |s_0|} \right) \quad (F.15)
\]
where $\mathbb{E} |s|$ can be bounded with $|\beta|_2$ and $\theta |\tau| < c_\mu \theta_{\log}$ from Lemma E.5. When $c_\mu < \frac{1}{10}$, observe that
\[
|s| \leq \sqrt{\sum_i \mathbb{E} x_i^2 \beta_i^2} \leq \sqrt{\theta} \left( |\beta|_2 + |\beta|_2 \right) \leq \sqrt{\theta} \left( 1 + c_\mu \right) + \frac{c_\mu \theta_{\log}}{|\tau|} \leq \frac{2 c_\mu \theta_{\log}}{|\tau|} \quad (F.16)
\]
Now choose $\varepsilon = \theta \leq \frac{c_\mu \theta_{\log}}{|\tau|}$, so that $\nu_1' = \nu_1 = \frac{\sqrt{\theta}}{2} \varepsilon$ in (F.15). Since $c_\mu < \frac{1}{25}$ we gain
\[
\mathbb{E} Z(\beta) \leq \theta \left( S_{\nu_1 \lambda} [\beta_0] + \frac{2 c_\mu \theta_{\log}}{|\tau|} \right) \leq \theta \left( S_{\nu_1 \lambda} [\beta_0] + \frac{3 c_\mu \theta_{\log}}{|\tau|} \right) \leq \theta \left( S_{\nu_1 \lambda} [\beta_0] + \frac{\sqrt{\theta_{\log}} \lambda}{5} \right) \leq \theta \left( S_{\nu_1 \lambda} [\beta_0] + \frac{1}{2} \nu_1 \lambda \right) \quad (F.17)
\]
(3). Combine both (F.14) and (F.17), we can thus conclude that
\[
\mathbb{E} Z(\beta) \leq \begin{cases} 4 \beta^2 |\tau| \beta_0 & \beta_0 \leq \nu_1 \lambda \\ \theta (\beta_0 - \frac{\nu_1 \lambda}{2}) & \beta_0 > \nu_1 \lambda \end{cases} \quad (F.18)
\]
2. (Lower bound of $\mathbb{E}Z$) On the other hand, for the lower bound for $\mathbb{E}Z$, use the fact that $\text{erf}_\beta(\lambda, s)$ is concave in $s_0$, we have

$$\mathbb{E}Z(\beta) = \mathbb{E}_{s_0} \mathbb{E}_{\mathbf{x}_0} \mathbb{E}_{\mathbf{x}_0} \mathcal{S}_{\lambda} [\beta_0 \mathbf{x}_0 + s_0]$$

$$= \theta \cdot \mathbb{E}_{s_0} \left[ \beta_0 - \frac{\beta_0}{2} \cdot \text{erf} \left( \frac{\lambda - s_0}{\sqrt{2} |\beta_0|} \right) - \frac{\beta_0}{2} \cdot \text{erf} \left( \frac{\lambda + s_0}{\sqrt{2} |\beta_0|} \right) \right]$$

$$\geq \theta \left( \beta_0 - \beta_0 \cdot \text{erf} \left( \frac{\lambda}{\sqrt{2} |\beta_0|} \right) \right) \geq \theta \cdot S_{\lambda} \mathbb{E} [\beta_0] =: \mathbb{E}Z(\beta).$$

(F.19)

The proof of $\beta_0 < 0$ is in the same vein. For cases of $i \neq 0$, since $\chi[\beta]_i \equiv \chi[\beta]_0 = 0$, replace $\beta_0$ with $\beta_i$ we obtain the desired result.

Another convenient fact of $\mathbb{E}\chi[\beta]_i$ is that it is monotone increasing w.r.t. $|\beta_i|$. The monotonicity is clear in Figure 3; it is demonstrated rigorously with the following lemma:

**Lemma F.3** (Monotonicity of $\mathbb{E}\chi[\beta]$). Suppose $\mathbf{x}_0 \sim \text{i.i.d.} \, \mathcal{B}G(\theta)$ in $\mathbb{R}^n$, and $|\tau|, c_\tau$ such that $(a_0, \theta, |\tau|)$ satisfies the sparsity-coherence condition $\text{SCC}(c_\mu)$. Define $\lambda = c_\lambda \sqrt{|\tau|}$ in $\varphi_\mu$, where $c_\lambda \in [0, \frac{1}{2}]$, then there exists some numerical constant $\tau > 0$, such that $c_\mu < \tau$, the expectation $\mathbb{E}|\chi[\beta]|_i$ is monotone increasing in $|\beta_i|$. In other words, if $|\beta_i| > |\beta_j|$ then

$$\sigma_i \mathbb{E} |\chi[\beta]|_i \geq \sigma_j \mathbb{E} |\chi[\beta]|_j$$

(20)

where $\sigma_i = \text{sign}(\beta_i)$.

The proof first operate simple calculus and then followed by studying cases of $|\beta_i| > |\beta_j|$ when either it is smaller are larger then $\lambda$.

**Proof.** 1. (Monotonicity by gradient negativity) Wlog assume $\beta_i > \beta_j > 0$, and from Lemma F.2 we can write $(n\theta)^{-1}\mathbb{E}\chi[\beta]_i = \beta_i (1 - \mathbb{E}_{s_i} \text{erf}_\beta(\lambda, s_i))$. Consider $t \in [0, 1]$ and define $\ell(t) = t\beta_i - t\beta_j$. Write the random variable $s_{ij} = \sum_{t \neq i,j} \beta_k x_t$. Define $h$ as a function of $t$ such that

$$h(t) = \mathbb{E}_{x, s_{ij}} \left[ (\beta_i - \beta_j) (1 - \text{erf}_{1-t}(\lambda, (\beta_i + \ell(t))x + s_{ij})) \right] = \mathbb{E}_{x, s_{ij}} \left[ (\beta_i - \ell(t)) (1 - \text{erf}_{1-t}(\lambda, (\beta_j + \ell(t))x + s_{ij})) \right].$$

(F.21)

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

$$\text{erf}_\beta(\lambda, s_{ij}) = \int_0^{\lambda + s_{ij}} f(z) \, dz + \int_0^{\lambda - s_{ij}} f(z) \, dz,$$

and use chain rule:

$$h'(t) = \mathbb{E}_{x, s_{ij}} \left[ (\beta_j - \beta_i) (1 - \text{erf}_{1-t}(\lambda, (\beta_j + \ell(t))x + s_{ij})) \right]$$

$$- (\beta_i - \ell(t)) \cdot \frac{d}{dt} \left( \frac{\lambda + x \cdot (\beta_j + \ell(t)) + s_{ij}}{\beta_i - \ell(t)} \right) \cdot f \left( \frac{\lambda + x \cdot (\beta_j + \ell(t)) + s_{ij}}{\beta_i - \ell(t)} \right)$$

$$- (\beta_i - \ell(t)) \cdot \frac{d}{dt} \left( \frac{\lambda - x \cdot (\beta_j + \ell(t)) - s_{ij}}{\beta_i - \ell(t)} \right) \cdot f \left( \frac{\lambda - x \cdot (\beta_j + \ell(t)) - s_{ij}}{\beta_i - \ell(t)} \right)$$

$$= (\beta_j - \beta_i) \mathbb{E}_{x, s_{ij}} \left[ 1 - \text{erf}_{1-t}(\lambda, (\beta_j + \ell(t))x + s_{ij}) \right]$$

$$+ \left( \frac{\lambda + x (\beta_j + \ell(t)) + s_{ij}}{\beta_i - \ell(t)} \right) \cdot f \left( \frac{\lambda + x (\beta_j + \ell(t)) + s_{ij}}{\beta_i - \ell(t)} \right).$$

24
Consider the term only related to \( z_{\lambda_+} \), condition on cases that it is either positive or negative, observe that

\[
\mu_+ := E_{x,s_{ij}|z_{\lambda_+}>0} \left[ \int_0^{z_{\lambda_+}} f_z (z) dz - 2 \lambda_+ f_{z_{\lambda_+}} \right] = 0, \\
\mu_- := E_{x,s_{ij}|z_{\lambda_-}>0} \left[ \int_0^{z_{\lambda_-}} f_z (z) dz - 2 \lambda_- f_{z_{\lambda_-}} \right] \leq 0,
\]

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

\[
\begin{align*}
- \int_0^v f_z (z) dz + v \cdot f(v) &= 0, \\
\frac{dv}{d\tau} \left\{ - \int_0^v f_z (z) dz + v \cdot f(v) \right\} &= -f(v) + f'(v) < 0 & v > 0;
\end{align*}
\]

and similarly for \( z_{\lambda_-} \):

\[
\begin{align*}
\mu_- := E_{x,s_{ij}|z_{\lambda_-}>0} \left[ \int_0^{z_{\lambda_-}} f_z (z) dz - 2 \lambda_- f_{z_{\lambda_-}} \right] &\leq 0, \\
\mu_+ := E_{x,s_{ij}|z_{\lambda_+}>0} \left[ \int_0^{z_{\lambda_+}} f_z (z) dz - 2 \lambda_+ f_{z_{\lambda_+}} \right] &\leq 0.
\end{align*}
\]

then combine every term to (F.22) using tower property and from assumption \( \beta_j - \beta_i < 0 \) we obtain

\[
(F.22) \leq (\beta_j - \beta_i) \left( 1 - \mathbb{P} \left[ z_{\lambda_+} > 0 \right] \cdot \mu_+ \right) - \mathbb{P} \left[ z_{\lambda_-} > 0 \right] \cdot \mu_- + E_{x,s_{ij}} \left[ x (f_{z_{\lambda_+}} - f_{z_{\lambda_-}}) \right] \leq (\beta_j - \beta_i) \left( 1 - \min \left\{ \frac{\mathbb{P} \left[ z_{\lambda_+} > 0 \right]}{2}, \frac{E \left[ z_{\lambda_+} \right]}{\sqrt{2\pi}} \right\} \right) - \min \left\{ \frac{\mathbb{P} \left[ z_{\lambda_-} > 0 \right]}{2}, \frac{E \left[ z_{\lambda_-} \right]}{\sqrt{2\pi}} \right\} - \frac{\theta}{\sqrt{2\pi}} \cdot E \left[ x \right], (F.23)
\]

where \( g \) is standard Gaussian r.v.

2. (Cases of varying \( \beta_i = \beta_j \)) Let \( c \lambda < \frac{1}{4} \). Suppose \( \beta_i - \ell(t) \leq \frac{1}{4 \sqrt{|\tau|}} \). Recall that \( \| \beta \|_2^2 \geq 1 - 3c \mu \). We are going to show there is at least one of the entry \( \beta_* \in \{ \beta_i \}_{i \in \tau \setminus \{ j \}} \) \( \cup \{ \beta_j + \ell(t) \} \) is greater than \( \frac{0.85}{\sqrt{|\tau|}} \). First, if both \( i, j \notin \tau \), the lower bound is immediate since \( \beta_*^2 = \| \beta_* \|_2^2 > 1 - 3c \mu > \frac{1}{4 \sqrt{|\tau|}} \). On the other hand if at least one of \( i, j \) is in \( \tau \) and all other \( \beta \tau \) entries are small where \( \| \beta \tau \setminus \{ i, j \} \|_2^2 \leq \| \beta \|_2^2 \geq \frac{1}{4 \sqrt{|\tau|}} \), then we know via norm inequalities,

\[
(\beta_i + \beta_j)^2 > \beta^2_i + \beta^2_j > \| \beta \|_2^2 - (1 - 1) \| \beta \tau \setminus \{ i, j \} \|_2^2 \geq \frac{1}{1 - 3c \mu}, (F.24)
\]

which implies if \( c \mu < \frac{1}{10 \mu} \),

\[
\beta_* = \beta_j + \ell(t) = (\beta_i + \beta_j) - (\beta_i - \ell(t)) \geq \frac{\sqrt{1 - 3c \mu}}{\sqrt{|\tau|}} - \frac{1}{4 \sqrt{|\tau|}} \geq \frac{0.72}{\sqrt{|\tau|}}. (F.25)
\]

In this case, adopt result from Corollary E.6 such that \( \mathbb{P} \left[ | \sum_{i \in \tau} \beta_i x_i | > \lambda / 10 \right] \leq 0.01 \), we have

\[
\mathbb{P} [ z_{\lambda_-} > 0 ] = 1 - \mathbb{P} [ x (\beta_j + \ell(t)) + s_{ij} < -\lambda ]
\]
We can express the (pseudo) curvature (B.10) in direction \( v \in S^{p-1} \) in terms of the correlation \( \gamma = C_{a_0}^* \nu \) between \( v \) and \( a_0 \), giving

\[
v^* \tilde{\nabla}^2 \varphi_{\ell^1}(a) v = -\gamma^* \tilde{C}_{x_0} P_\ell \tilde{C}_{x_0} \gamma,
\]

where

\[
I(a) = \text{supp} \left( S_\lambda \left[ \tilde{C}_{x_0} C_{a_0}^* \ell \right] a \right) = \left\{ i \in [n] \mid x_0 \ast \tilde{b}_i > \lambda \right\}.
\]
The $i$-th diagonal entry of $\tilde{C}_{\mathbf{x}_0} P_{(a)} \tilde{C}_{\mathbf{x}_0}$ is

$$-e_i^T \tilde{C}_{\mathbf{x}_0} P_{(a)} \tilde{C}_{\mathbf{x}_0} e_i = - \left\| P_{(a)} \tilde{C}_{\mathbf{x}_0} e_i \right\|_2^2 = - \left\| P_{(a)} s_{-i} [x_0] \right\|_2^2,$$

which is the core component for us to study the curvature of objective $\varphi_\ell$. We illustrate the expectation of diagonal term of Hessian in Lemma G.2 and Corollary G.3, whose figure of visualized $\left\| P_{(a)} s_{-i} [x_0] \right\|_2^2$ is shown in Figure 3. Lastly, we also prove the off-diagonal terms $e_i^T \tilde{C}_{\mathbf{x}_0} P_{(a)} \tilde{C}_{\mathbf{x}_0} e_j$ of Hessian is likely inconsequential in calculation of curvature in Lemma G.4.

We expect the Hessian to have stronger negative component in the $s_i [a_0]$ direction as $\left\| P_{(a)} s_{-i} [x_0] \right\|_2^2$ becomes larger. This term can be tremendously simplified when $x_0$ is very sparse: suppose all entries of its support $I_0$ are separated by at least $2p−1$ samples, then by implementing the definition of support from (G.1), we can derive

$$- \left\| P_{(a)} s_{-i} [x_0] \right\|_2^2 = - \sum_{j \in I_0} x_{0j}^2 1 \{ |\sum_j \beta_j x_{0j} | > \lambda \} \sum_{j \in I_0} g_j^2 1 \{ |\beta_j g_j | > \lambda \},$$

where $1$ is the indicator function and $g_j$ are independent standard Gaussian r.v.s. In expectation, the summands in (G.3) acts like a smoothed logic function on entry $\beta_i$:

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

$$\mathbb{E}_g \left[ g^2 1 \{ |\beta g + s | > \lambda \} \right] = 1 - \text{erf}_b (\lambda, s) + f_b (\lambda, s),$$

where

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

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

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

$$\mathbb{E}_g^2 1 \{ |\beta g | > \lambda \} \approx \begin{cases} 1, & |b| = 1 \\ 1/2, & |b| = \sqrt{2/\pi} \lambda \\ 0, & |b| = 0 \end{cases}.$$

Relate (G.6) to (G.3), when $|\beta_i|$ is close to 1 then we expect $- \frac{1}{n \theta} \left\| P_{\ell} s_{-\ell} [x_0] \right\|_2^2$ to be close to $-1$, and it increases to 0 as $|\beta_i|$ decreases, suggests that the Euclidean Hessian at point $a$ has stronger negative component at $s_i [a_0]$ direction if $|\langle a, s_i [a_0] \rangle|$ is larger. See Figure 4 for a numerical example. This phenomenon can be extend beyond the idealistic separating case as follows:

**Lemma G.2 (Expected Hessian diagonals).** Let $x_0 \sim_{\text{i.i.d.}} \mathcal{BG}(\theta)$ and $\lambda > 0$, define the set $I (a)$ in (G.1), write $s_i = \sum_{\ell \neq i} \beta_i x_{0\ell}$, then for every $a \in \mathbb{S}^{p−1}$ and $i \in [n]$:

$$n^{-1} \mathbb{E} \left[ \left\| P_{(a)} s_{-i} [x_0] \right\|_2^2 \right] = \theta \left[ 1 - \mathbb{E}_s \text{erf}_{\beta_i} (\lambda, s_i) + \mathbb{E}_s f_{\beta_i} (\lambda, s_i) \right].$$

**Proof.** Write $x_0$ as $a$. Observe that $y * \bar{a} = x_0 * \bar{a} = \sum_{\ell} \beta_{\ell} s_{-\ell} [x_0]$. Thus for any $j \in [n]$ and $i \in [\pm p]$:

$$(y * \bar{a})_{j-i} = \left( \beta_i s_{-i} [x] + \sum_{\ell \neq i} \beta_{\ell} s_{-\ell} [x] \right)_{j-i} = \beta_j x_j + \sum_{\ell \neq i} \beta_{\ell} x_{j+\ell-i} + \beta_i x_j + s_j,$$

27
where \( x_j \) is independent of \( s_j \), and both \( x_j, s_j \) are symmetric and identically distributed for all \( j \in [n] \). Rewrite the random variable using (G.1) as

\[
\| P_{(a)} s_{-i}[x_0] \|_2^2 = \| P_{(a)} \sum_{j \in [n]} (x_{0j} e_{j-i}) \|_2^2 = \sum_{j \in [n]} x_{0j}^2 1_{\{ |s| > \lambda \}}
\]

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

\[
\mathbb{E} \| P_{(a)} s_{-i}[x_0] \|_2^2 = n \theta \cdot \mathbb{E} g_2^2 1_{\{ |\beta| + s_0 | > \lambda \}} = n \theta \cdot \mathbb{E} (1 - \text{erf}_{\lambda}(\lambda, s_0) + f_{\beta}(\lambda, s_0))
\]

where \( s_i = \sum_{\ell \neq i} x_{0\ell} / \beta_i \) with \( x_{0\ell} \sim \text{i.i.d.} \) \( \text{BG}(\theta) \), yielding the claimed expression. \( \blacksquare \)

When the signal length of \( y \) is sufficiently large, then \( i \)-th diagonal term for Hessian \( \| P_{(a)} s_{-i}[x_0] \|_2^2 \) will be close enough to its expected value.

**Corollary G.3 (Large sample deviation of curvature).** Suppose \( x_0 \sim \text{i.i.d.} \) \( \text{BG}(\theta) \) in \( \mathbb{R}^n \), and \( k, c_\mu \) such that \((a_0, \theta, k)\) satisfies the sparsity-coherence condition \( \text{SCC}(c_\mu) \). Define \( \lambda = c_\lambda \sqrt{n} \) in \( \varphi_\theta \) for some \( c_\lambda > 1/5 \), then there exists some numerical constant \( C, c, \tau > 0 \), such that if \( n \geq C p^4 \theta^{-1} \log p \) and \( c_\mu \leq \tau \), then with probability at least \( 1 - 3/n \), for every \( a \in \cup_{|\tau|} \mathcal{L}(S, \gamma(c_\mu)) \) and every \( i \in [n] \), we have:

\[
n^{-1} \| P_{(a)} s_{-i}[x_0] \|_2^2 - n^{-1} \mathbb{E} \| P_{(a)} s_{-i}[x_0] \|_2^2 \leq c \theta / p \tag{G.9}
\]

**Proof.** See Appendix L.2. \( \blacksquare \)

The off-diagonal entries of Hessian in general are much smaller then the diagonal entries; however, it affects the region near sign shifts of \( a_0 \) 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 \( |\beta_{(0)}| \) is large and the rest is small.
Lemma G.4 (Hessian off-diagonal term near solution). Suppose \( x_0 \sim_{i.i.d.} \mathbb{B} \theta \) in \( \mathbb{R}^n \), and \( k, c_{\mu} \) such that \((a_0, \theta, k)\) satisfies the sparsity-coherence condition SCC \((c_{\mu})\). Let \( \lambda = c_{\mu} / \sqrt{k} \) with \( c_{\mu} > 1/5 \), then there exists some numerical constant \( C, \tau > 0 \) such that if \( n \geq C \theta^{-1} \log p \) and \( c_{\mu} \leq \tau \), then with probability at least \( 1 - 4/n \), for every \( a \in \bigcup_{|\tau| \leq k} \mathcal{H}(\mathbf{S}_\tau, \gamma(c_{\mu})) \), where \( |\beta(1)| \leq \frac{1}{4 \log \theta - 1} \lambda \) and every \( i \neq j \in [\pm p] \setminus \{0\} \), we have
\[
|s_j| \cdot |s_i|^* \cdot P_{1(a)} \cdot |s_j|^* \cdot |s_i| < 8n \theta^3 \tag{G.10}
\]

Proof. Write \( \theta_0 = -1 / \log \theta \) and \( x_0 = x = \omega \circ g \). Wlog let \( \beta_0 \) be the largest correlation \( \beta(0) \). Define random variables \( s' = \langle \beta_{\tau \setminus \{0, i, j\}}, x_{\tau \setminus \{0, i, j\}} \rangle \). Firstly via Corollary E.7 we have \( P[|s'| > 0.4 \lambda] \leq 2 \theta \); also define \( s = \langle \beta_{\tau \setminus \{0, i, j\}}, x_{\tau \setminus \{0, i, j\}} \rangle \), and base on Corollary E.6 we have \( P[|s| > \lambda / 10] \leq 2 \theta \). Expand the \((-i, j)-th\) cross term with \( \theta < 0.1 \) we have:
\[
E|s_j| \cdot |s_i|^* \cdot P_{1(a)} \cdot |s_j|^* \cdot |s_i| = \sum_{|k| \in [n]} |x_k + \alpha x_{k+i} x_{k+i} + \alpha x_{k+i+j} + s + s'| \lambda \n = \theta\lambda \cdot E|g_j|^2 |\{\beta_{\tau \setminus \{0, i, j\}}, x_{\tau \setminus \{0, i, j\}} \rangle | \lambda \n \leq \theta^2 \cdot E|g_j|^2 (2 (\lambda / 4)^4) + P[|s| > 0.1 \lambda] + P[|s'| > 0.4 \lambda]) \n \leq \theta^2 \cdot (\exp(-\log^2 \theta - 1) + \theta + 2 \theta + 2 \theta) \n \leq 6n \theta^2. \tag{G.11}
\]

Write (G.10) as two summation of independent random variables with \( t = j - i \) by separating sum into two sets \( J_{11}, J_{12} \) defined in (D.4) with both \( |J_{11}|, |J_{12}| < n \theta^2 \) with probability at least \( 1 - 2/n \) from Lemma D.1
\[
E|s_j| \cdot |s_i|^* \cdot P_{1(a)} \cdot |s_j|^* \cdot |s_i| = \sum_{(k-i) \in J_{11}} \sum_{(k-i) \in J_{12}} |g_k| |g_k+t| = \sum_{(k-i) \in J_{11}} |g_k| |g_k+t| + \sum_{(k-i) \in J_{12}} |g_k| |g_k+t|,
\]
whose first summands can be upper bounded with high probability via Bernstein inequality Lemma N.4 with \((\sigma^2, R) = (1, 1)\) and writes \( C := \bigcup_{|\tau| \leq k} \mathcal{H}(\mathbf{S}_\tau, \gamma(X)) \cap \{a \mid |\beta(1)| \leq \frac{1}{4 \log \theta - 1} \lambda\} \), then we have
\[
P\left[ \max_{a \in C} \sum_{(k-i) \in J_{11}} |g_k| |g_k+t| - \sum_{(k-i) \in J_{12}} |g_k| |g_k+t| \geq n \theta^3 \right] \leq 4p^2 \cdot \exp\left( \frac{-n^2 \theta^6}{2 |J_{11}| + 2n \theta^3} \right) \leq \exp\left( 8 \log p - \frac{-n^2 \theta^6}{3n \theta^2} \right) \leq \exp\left( \frac{-n^2 \theta^4}{10} \right) \leq \frac{1}{n} \tag{G.12}
\]
when \( n = C \theta^{-1} \log p \) with \( C > 10^4 \) and \( \theta \log^2 \theta^{-1} \geq 1/p \). Thus for all \( i \neq j \in [\pm p] \setminus \{0\} \) and \( a \) satisfies our condition of lemma, from (G.11) and (G.12) we can conclude:
\[
|s_j| |s_i|^* \cdot P_{1(a)} \cdot |s_j|^* \cdot |s_i| \leq \sum_{J_{11}} \sum_{J_{12}} |g_k| |g_k+t| = 2n \theta^3 \leq 8n \theta^3
\]
which holds with probability at least \( 1 - 2/n - 2 \cdot 1/n = 1 - 4/n \) base on Lemma D.1 and (G.12).
H 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 $\rho(x) = \sqrt{x^2 + \delta^2}$ in the main script, and is general for a class of function class defined in Definition H.2 that is smooth and well approximates $\ell^1$ when the proper smoothing parameter $\delta$ is chosen from the result of Lemma H.6. In particular we ask the regularizer $\rho_\delta(x)$ to be uniformly bounded to $|x|$ by $\delta/2$:

$$\forall x \in \mathbb{R}, \quad |\rho_\delta(x)| \leq \delta/2 \tag{H.1}$$

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

$$\| \text{prox}_{\lambda \rho} [\tilde{a} \ast y] - \text{prox}_{\lambda \rho} [\tilde{a} \ast y] \|_2 \to 0, \tag{H.2}$$
$$\| \nabla \varphi_{\ell^1}(a) - \nabla \varphi_{\rho_\delta}(a) \|_2 \to 0, \tag{H.3}$$
$$\| \tilde{\nabla}^2 \varphi_{\ell^1}(a) - \tilde{\nabla}^2 \varphi_{\rho_\delta}(a) \|_2 \to 0. \tag{H.4}$$

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

The marginal minimizer over $x$ in (??) can be expressed in terms of the proximal operator [BC11] of $\rho$ at point $\tilde{a} \ast y$:

$$\text{prox}_{\lambda \rho} [\tilde{a} \ast y] = \arg \min_{x \in \mathbb{R}^n} \left\{ \lambda \rho(x) + \frac{1}{2} \| x \|_2^2 - \langle a \ast x, y \rangle \right\}. \tag{H.5}$$

Plugging in, we obtain

$$\varphi_\rho(a) = \lambda \rho (\text{prox}_{\lambda \rho} [\tilde{a} \ast y]) + \frac{1}{2} \| \tilde{a} \ast y - \text{prox}_{\lambda \rho} [\tilde{a} \ast y] \|_2^2 - \frac{1}{2} \| \tilde{a} \ast y \|_2^2 + \frac{1}{2} \| y \|_2^2 \tag{H.6}$$

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

$$\varphi_\rho(a) = \epsilon(\lambda \rho)(\tilde{a} \ast y) - \frac{1}{2} \| \tilde{a} \ast y \|_2^2 + \frac{1}{2} \| y \|_2^2 \tag{H.7},$$

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

Fact H.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 $\epsilon(\lambda f)(z) = \lambda f (\text{prox}_{\lambda f} [z]) + \frac{1}{2} \| z - \text{prox}_{\lambda f} [z] \|_2^2$ is Fréchet differentiable with $\nabla \epsilon(\lambda f)(z) = z - \text{prox}_{\lambda f} [z]$.

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

$$\nabla \varphi_\rho(a) = -\iota^* \tilde{C}_y \text{prox}_{\lambda \rho} \left[ \tilde{C}_y t a \right], \tag{H.8}$$
$$\nabla^2 \varphi_\rho(a) = -\iota^* \tilde{C}_y \nabla \text{prox}_{\lambda \rho} \left[ \tilde{C}_y t a \right] \tilde{C}_y t. \tag{H.9}$$

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

$$\text{grad}[\varphi_\rho](a) = -P_a \iota^* \tilde{C}_y \text{prox}_{\lambda \rho} \left[ \tilde{C}_y t a \right], \tag{H.10}$$
$$\text{Hess}[\varphi_\rho](a) = -P_a \left( \iota^* \tilde{C}_y \nabla \text{prox}_{\lambda \rho} \left[ \tilde{C}_y t a \right] \tilde{C}_y t - \langle \nabla \varphi_\rho(a), a \rangle I \right) P_a^+. \tag{H.11}$$

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:
will show that it approximates This optimization problem is strongly convex, and so the minimizer 

\[ \text{prox}_{\lambda \rho}(x) \]

condition and since 
\[ \rho \]

Lemma H.4

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

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

\[ 0 \leq \text{sign}(z) (S_\lambda^{\sqrt{\delta}}(z) - S_\lambda(z)) \leq \sqrt{\delta}, \quad \forall z \in \mathbb{R} \]

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

\[ |\nabla S_\lambda^{\sqrt{\delta}}(z) - \nabla S_\lambda(z)| \leq \sqrt{\delta}/B, \quad ||z| - \lambda| \geq B. \quad (H.13) \]

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

Lemma H.4 (Proximal operator for smoothed \( \ell^1 \)). Suppose \( \rho \) is a \( \delta \)-smoothed \( \ell^1 \) function, then \( z_i \mapsto \text{prox}_{\lambda \rho}(z_i) \) is a \( \sqrt{\delta} \)-smoothed soft threshold function.

Proof. We know that

\[ x_z := \text{prox}_{\lambda \rho}(z) = \arg \min_{x \in \mathbb{R}^n} \lambda \rho(x) + \frac{1}{2} ||x - z||^2. \quad (H.14) \]

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

\[ x_{zi} = (\text{Id} + \lambda \nabla \rho_i)^{-1}(z_i). \quad (H.15) \]
Since $\rho_i$ is convex and even, $\nabla \rho_i$ is monotone increasing and odd. By inverse function theorem, we know that strict monotonicity and differentiability of $\text{Id} + \lambda \nabla \rho_i$ implies its inverse is differentiable and is a strictly monotone increasing odd function. Furthermore, it implies $\nabla x_{zi}$ has the form

$$\nabla x_{zi} = \nabla_x (\text{Id} + \lambda \nabla \rho_i)^{-1}(x_i) = \frac{1}{\lambda \nabla^2 \rho_i(x_{zi})} < 1. \quad \text{(H.16)}$$

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

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

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

Consider a positive $x$ with $\nabla \rho_i(x) > 1 + \varepsilon$ for some $\varepsilon > 0$, by convexity if $\bar{x} > x$ then $\nabla \rho_i(\bar{x}) > 1 + \varepsilon$, 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 $x \geq \sqrt{\lambda \delta}$:

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

To prove (H.12), when $0 \leq z_i \leq \lambda$, then $S_\lambda[z_i] = 0$ and $x_{zi} \leq \sqrt{\lambda \delta}$ since if $x_{zi} > \sqrt{\lambda \delta}$, by (H.17):

$$\lambda \nabla \rho_i(x_{zi}) + x_{zi} > \lambda (1 - \sqrt{\delta/\lambda}) + \sqrt{\lambda \delta} = \lambda \geq z_i$$

then $x_{zi}$ violates the stationary condition in (H.15), resulting $0 \leq x_{zi} - S_\lambda[z_i] \leq \sqrt{\lambda \delta}$ whenever $0 \leq z_i \leq \lambda$.

Likewise in the case of $z_i \geq \lambda$ where $S_\lambda[z_i] = z_i - \lambda$, (H.17) provides:

$$\begin{align*}
\forall x_{zi} > z_i - \lambda + \sqrt{\lambda \delta}, & \quad \lambda \nabla \rho_i(x_{zi}) + x_{zi} > \lambda (1 - \sqrt{\delta/\lambda}) + z_i - \lambda + \sqrt{\lambda \delta} = z_i \\
\forall x_{zi} < z_i - \lambda, & \quad \lambda \nabla \rho_i(x_{zi}) + x_{zi} < \lambda + z_i - \lambda = z_i
\end{align*}$$

again violates (H.15) and therefore (H.12) holds for all $z_i \in \mathbb{R}$.

Lastly (H.13) is a direct result of (H.12). For all $z_i \leq \lambda - B$, recall that $\nabla x_{zi}$ is monotone increasing in $z_i$:

$$\nabla x_{zi} \leq \min_{y \in [\lambda - B, \lambda]} \nabla x_{yi} \leq \frac{x_{zi} - x_{(\lambda - B)j}}{\lambda - (\lambda - B)} \leq \frac{(\sqrt{\lambda \delta} + S_\lambda[\lambda]) - S_\lambda[\lambda - B]}{B} = \frac{\sqrt{\lambda \delta}}{B};$$

and similarly for all $z_i > \lambda + B$:

$$\nabla x_{zi} \geq \max_{y \in [\lambda, \lambda + B]} \nabla x_{yi} \geq \frac{x_{(\lambda + B)j} - x_{zi}}{(\lambda + B) - \lambda} \geq \frac{S_\lambda[\lambda + B] - (S_\lambda[\lambda] + \sqrt{\lambda \delta})}{B} = 1 - \frac{\sqrt{\lambda \delta}}{B},$$

implies (H.13) holds.

Based on (H.9)-(H.10) and denote $C_{y \cdot a} = \tilde{a} + y$, the only differences of Riemannian gradient and Hessian between $\varphi_{\rho}$ and $\varphi_{\rho_i}$ comes from the difference of $\text{prox}_{\lambda || \cdot ||}$ and $\text{prox}_{\lambda || \cdot ||_{\ell_1}}$. Thus for the purpose of obtaining good geometric approximation of $\varphi_{\rho}$ with that of objective $\varphi_{\rho_i}$, we may apply both Definition H.3 and Lemma H.4, together suggest if $\rho$ is a $\delta$-smoothed $\ell^1$ function, then the $i$-th entry of $\text{prox}_{\lambda || \cdot ||}$ will be $\sqrt{\lambda \delta}$-close to the authentic soft thresholding function $S_\lambda[\tilde{a} + y]$, and its gradient $\nabla \text{prox}_{\lambda || \cdot ||}$ is $\sqrt{\lambda \delta}/B$-close to $\nabla S_\lambda[\tilde{a} + y]$ as long as $(\tilde{a} + y)_i$ is not close to $\pm \lambda$ by distance $B$.

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

32
Lemma H.5 (Gradients discontinuity entries). For each $a \in \mathbb{S}^{p-1}$, let
\[ J_B(a) := \left\{ i \mid \left( C_y a \right)_i \in [-\lambda - B, -\lambda + B] \cup [\lambda - B, \lambda + B] \right\}. \tag{H.18} \]
Suppose the subspace dimension is at most $k$ and signal $y$ satisfies Definition E.1. Let $\lambda = c_\lambda / \sqrt{k}$ and $B \leq c' \lambda^2 / p \log n$ for some $c_\lambda, c' \in (0, 1)$, then there is a numerical constant $\tilde{C} > 0$ such that if $n \geq C \mu^2 \theta^{-2} \log p$, then with probability at least $1 - 3/n$, for every $a \in \cup_{|r| \leq k} \mathcal{R}(\mathcal{S}_r, \gamma(c))$, we have
\[ |J_B(a)| \leq \frac{2c_n \theta^2}{p \log n} \tag{H.19} \]

Proof. See Appendix L.3.

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

Lemma H.6 ($\varphi^\ell$ approximates $\varphi^\emptyset$). Suppose $x_0 \sim_{i.i.d.} \text{BG}(\theta)$ in $\mathbb{R}^n$, and $k, \epsilon_k$ such that $(a_0, \theta, k)$ satisfies the sparsity-coherence condition $\text{SCC}(c_k)$. Let $\rho \in \mathbb{R}^n \to \mathbb{R}$ be a $\delta$-smoothed $\ell^1$ function with
\[ \lambda = \frac{c\lambda}{\sqrt{k}}, \quad \delta \leq \frac{c'\theta^8}{p^2 \log^2 \lambda} \tag{H.20} \]
with some $c', c_\lambda \in (0, 1)$, then there is a numerical constant $C, \tau > 0$ such that if $n \geq C \rho^2 \theta^{-2} \log p$ and $c_k \leq \tau$, then with probability at least $1 - 10/n$, the following statements hold simultaneously for every $a \in \cup_{|r| \leq k} \mathcal{R}(\mathcal{S}_r, \gamma(c_k))$:

1. The coefficients have norm difference
\[ \left\| \left( \left( \left. \proxt{\lambda^*}{\ell^1} \right|_{[\pm \rho]} \right) \overset{\mathbf{C}_{x_0}}{\bigl(} \left[ a \ast y \bigr] - \left[ \left. \proxt{\lambda}{\ell^1} \right|_{[\pm \rho]} \right) \overset{\mathbf{C}_{x_0}}{\bigl(} \left[ \left[ a \ast y \bigr] \right\|_2 \leq c' \theta n^4. \tag{H.21} \right. \]

2. The gradient has norm difference
\[ \left\| \nabla \varphi^\ell (a) - \nabla \varphi^\emptyset (a) \right\|_2 \leq c' \theta n^4. \tag{H.22} \]

3. The (pseudo) Riemannian curvature difference is bounded in all directions $v \in \mathbb{S}^{p-1}$ via
\[ \forall v \in \mathbb{S}^{p-1}, \quad \left\| \left( \left. \text{Hess}[\varphi^\ell] \right|_{[\pm \rho]} (a) - \left. \text{Hess}[\varphi^\emptyset] \right|_{[\pm \rho]} (a) \right) v \right\| \leq 200 c' \theta n^2. \tag{H.23} \]

Proof. 1. (Coefficients) From Lemma H.4, the proximal $\delta$-smoothed $\ell^1$ function satisfies
\[ \left| S_{\lambda} [a \ast y] - S_{\lambda} [a \ast y] \right|_2 < \sqrt{\lambda \delta} \quad \forall a \in [n]. \]
Since the support of coefficient vectors are contained in $[\pm \rho]$, using simple norm inequality:
\[ \left\| \left. \proxt{\lambda}{\ell^1} \right|_{[\pm \rho]} \overset{\mathbf{C}_{x_0}}{\bigl(} \left[ a \ast y \bigr] - \left. \proxt{\lambda}{\ell^1} \right|_{[\pm \rho]} \overset{\mathbf{C}_{x_0}}{\bigl(} \left[ a \ast y \bigr] \right\|_2 \leq \sqrt{\lambda \delta n} \cdot \left\| \left. \proxt{\lambda}{\ell^1} \right|_{[\pm \rho]} \overset{\mathbf{C}_{x_0}}{\bigl(} \left[ a \ast y \bigr] \right\|_2. \tag{H.24} \right. \]
Apply Lemma D.5 by replacing $a_0$ with standard basis $e_0$ and extend support of $a$ to $a_0$, notice that in this case we have $\mu = 0$. Condition on the event
\[ \left\| \left. \proxt{\lambda}{\ell^1} \right|_{[\pm \rho]} \overset{\mathbf{C}_{x_0}}{\bigl(} \right\|_2 \leq \left\| \left. \proxt{\lambda}{\ell^1} \right|_{[\pm \rho]} \overset{\mathbf{C}_{x_0}}{\bigl(} \right\|_2 \leq \sqrt{3(1 + 2 \mu \rho) n \theta} \leq \sqrt{3 n \theta}, \]
and we gain
\[ (H.24) \leq \sqrt{\lambda \delta n} \cdot \sqrt{3 n \theta} \leq n \sqrt{3 \lambda \theta \delta} \leq c' \theta n^4. \]
2. (Gradient) From definition of Riemannian gradient (H.9) and apply similar norm bound of (H.24), and condition on the following events of Lemma D.5 holds, obtain
\[ \| \nabla \phi_j (v) - \nabla \phi (v) \|_2 \leq \sqrt{dn} \cdot \| v^* \bar{C} y \|_2 \leq n \sqrt{3 \lambda \theta (1 + \mu p) \delta} \leq c'n \theta^4. \] (H.25)

3. (Hessian) For every realization of \( J_B (a) \) from \( a \in \{ \tau \leq \tilde{R} (S_T, \gamma (\epsilon_\mu)) \} \), base on Lemma H.5, condition on the event such that
\[ B \leq \frac{c' \lambda \theta^2}{p \log n}, \quad |J| \leq \frac{2c'n \theta^2}{p \log n}; \] (H.26)
and rewrite \( J_B (a) \) as \( J \). Also condition on the event using Lemma D.5 and \( (1 + \mu p) \theta \log \theta^{-1} < 1 \)
\[ \| v^* \bar{C} y \|_2 \leq \sqrt{3n}, \quad \| v^* \bar{C} y \|_2 \leq \sqrt{8 |J| p \log n}, \] (H.27)
then the difference of Hessian (H.10), in direction \( v \in S_{p-1} \) can be bounded as
\[
\| v^* \left( \text{Hess} [\phi_j] (a) - \text{Hess} [\phi] (a) \right) v \| \\
\leq v^* v^* \bar{C} y \left( P_I (a) - \text{diag} \left[ \nabla S_k^2 \left[ \bar{C} y \bar{a} \right] \right] \right) \bar{C} y v + \| \nabla \phi_j (a) - \nabla \phi (a) \|_2 \] (H.28)
where \( I (a) \) is defined in (G.1). Let \( D = P_I (a) - \text{diag} \left[ \nabla S_k^2 \left[ \bar{C} y \bar{a} \right] \right] \) and notice that \( D \) is a diagonal matrix, which suggests (H.28) can be decomposed using
\[
(P_J + P_{J'}) D (P_J + P_{J'}) = P_J D P_J + P_{J'} D P_{J'},
\]
where, from with property of \( \sqrt{\delta} \)-smoothed \( \ell^1 \) function Lemma H.4:
\[
\max_j |P_J D P_J|_{jj} \leq 1, \quad \max_j |P_{J'} D P_{J'}|_{jj} \leq \sqrt{\lambda \delta / B}.
\]
Finally, once again apply \( \delta \) bound from (H.20) and bounds for \( B, |J|, y \) from (H.26)-(H.27), we gain
\[
(H.28) \leq \left| v^* \bar{C} y P_J \right|_2^2 + \frac{\sqrt{\lambda \delta}}{B} \left| v^* \bar{C} y \right|_2^2 + \| \nabla \phi_j (a) - \nabla \phi (a) \|_2 \\
\leq 8 |J| p \log n + \frac{3n \sqrt{\lambda \delta}}{B} + c'n \theta^2 \\
\leq 8 \cdot \frac{2c'n \theta^2}{p \log n} \cdot p \log n + \frac{3n (c^4 \lambda^2 \theta^8 / p^2 \log^2 n)^{1/2}}{c' \lambda \theta^2 / p \log p} + c'n \theta^2 \\
\leq 200c' n \theta^2,
\]
where all above result holds with probability at least \( 1 - 10/n \) from Lemma H.5 and Lemma D.5.

I Analysis of geometry

In this section we prove major geometrical result in Theorem B.1. This lemma consists of three parts of geometry of \( \phi_e \), including the negative curvature region Corollary I.2, large gradient region Corollary I.4, strong convexity region near shift Corollary I.6, and retraction to subspace Corollary I.8, which are respectively base on geometric properties of \( \phi_j \) in Lemma I.1, Lemma I.3, Lemma I.5 and Lemma I.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 \( S_{(i,j)} \), as follows:
I. Negative curvature

For any $a \in S^{p-1}$ near the subspace $S_\tau$ such that the entries of leading correlation vector $\beta_{(0)}, \beta_{(1)}$ have balanced magnitude, the Hessian of $\varphi_\rho(a)$ exhibits negative curvature in the span of $s_{(0)}(a_0), s_{(1)}(a_0)$. We will first demonstrate the pseudo negative curvature of $\varphi_{\ell_1}$ in Lemma I.1, then show $\varphi_\mu$ approximates $\varphi_{\ell_1}$ in terms of Hessian in Corollary I.2 when $\rho$ is properly defined as in Appendix H.

**Lemma I.1** (Negative curvature for $\varphi_{\ell_1}$). Suppose that $x_0 \sim_{i.i.d.} BG(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(a_0, \theta, k)$ satisfies the sparsity-coherence condition $SCC(c_\mu)$. Set $\lambda = c_\lambda/\sqrt{k}$ in $\varphi_{\ell_1}$ with $c_\lambda \in \left[\frac{1}{5}, \frac{1}{4}\right]$. There exist numerical
constants $C, c, c', \tau > 0$ such that if $n > C p^5 \theta^{-2} \log p$, and $c_\mu \leq \tau$, then with probability at least $1 - c'/n$ the following holds at every $a \in \cup_{|\tau| \leq k} \mathcal{R}(S, \gamma(c_\mu))$ satisfying $|\beta(1)| \geq \frac{5}{4} |\beta(0)|$: for $v \in S_{(0),(1)} \cap S_t^p \cap a^\perp$, 
\[ v^\ast \text{Hess}[\varphi_{\ell}](a)v \leq -c n \theta \lambda. \]  
(I.1)

**Proof.** First of all the regional condition \[ \frac{|\beta(0)|}{|\beta(1)|} \leq \frac{5}{4} \] provides a two side bound for the two leading $\beta$'s 
\[ 0.79 \geq \frac{|\beta(0)|}{\sqrt{\beta(0)^2 + \beta(1)^2}} \geq |\beta(0)| \geq |\beta(1)| \geq \frac{4}{5} |\beta(0)| \geq \frac{4}{5} \cdot \frac{|\beta(2)|}{\sqrt{|\beta(2)|}} \geq 0.79 \]  
(I.2)

Set $J = \{(0), (1)\}$, choose $v = \iota^\ast C_{a_i} \iota^\ast \gamma$ with $\|v\|_2 = 1$ then $\|\gamma\|_2 - 1 \leq \mu$. There exists such $v$ satisfies condition above with $a \perp v$ by choosing $\gamma$ as 
\[ a^\ast v = a^\ast C_{a_i} \iota^\ast \gamma = \gamma(0) \beta(0) + \gamma(1) \beta(1) = 0, \]  

hence $\gamma(0) = |\gamma(0)| |\beta(0)| \leq \frac{5}{4}$, This implies $\gamma(0)^2 \geq \frac{10}{25} \gamma(1)^2 \geq \frac{10}{25} (1 - \mu - \gamma(0))$ where $\mu \leq \frac{\mu}{4} \leq \frac{1}{105}$, it gives the lower bound of $\gamma(0)$ as 
\[ \gamma(0) \geq \frac{(1 - \mu) \cdot 16}{25 + 16} \geq 0.385 \]  
(I.3)

1. (Expand the Hessian) The (pseudo) curvature along direction $v$ is written as 
\[ v^\ast \text{Hess}[\varphi_{\ell}](a)v = v^\ast \nabla^2 \varphi_{\ell} \cdot (a)v = \langle \nabla \varphi_{\ell} \cdot (a), a \rangle = -\gamma^\ast \iota^\ast M \tilde{C}_x P_{(a)} \tilde{C}_x M_{i,j} \gamma + \beta^\ast \chi[\beta] \]  
(I.4)

expand the first term of (I.4) we obtain 
\[ -\gamma^\ast \iota^\ast M \tilde{C}_x P_{(a)} \tilde{C}_x M_{i,j} \gamma \]
\[ = -\gamma^\ast \iota^\ast M (P(0) + P(1) + P_{i,j}) \tilde{C}_x P_{(a)} \tilde{C}_x (P(0) + P(1) + P_{i,j}) M_{i,j} \gamma \]
\[ \leq - \sum_{i \in J} \left| P_{(a)} \tilde{C}_x e_i \right|_2^2 \left( e^\ast_i M_{i,j} \gamma \right)^2 + 2 \sum_{(i,j) \in \{A, J\} \atop (i,j) \neq (0),(1))} \left| e^\ast_i \tilde{C}_x P_{(a)} \tilde{C}_x e_j \right| \left( \left| e^\ast_i M_{i,j} \gamma \right| \left( \left| e^\ast_j M_{i,j} \gamma \right| \right) \right) \]
\[ \leq - \sum_{i \in J} \left| P_{(a)} \tilde{C}_x e_i \right|_2^2 \left| \gamma(0) - \mu \right|^2 \]
\[ + 2 \max_{i \neq j \in \{A\}} \left| e^\ast_i \tilde{C}_x P_{(a)} \tilde{C}_x e_j \right| \left( \left| \iota^\ast_j M_{i,j} \gamma \right|_1 \left| \iota^\ast_j M_{i,j} \gamma \right|_1 + \left( \left| \gamma(0) \right| + \mu \right) (\left| \gamma(1) \right| + \mu) \right) \]
(I.5)

Consider the following events 
\[ E_{\text{cross}} := \left\{ \forall a \in S_t^p, \max_{i \neq j \in \{A\}} \left| e^\ast_i \tilde{C}_x P_{(a)} \tilde{C}_x e_j \right| < 4 n \theta \right\} \]
\[ E_{\text{ncurv}} := \left\{ \forall a \in \mathcal{R}(S, \gamma(c_\mu)), \min_{i \in J} \left| P_{(a)} s_i \ast \left| x \right| \right|_2 \geq n \theta (1 - \mathbb{E}_{a_i}(\lambda, s_i) + \mathbb{E}_{a_i}(\lambda, s_i)) - c_n \right\}, \]
(I.6)

and from Lemma E.4 we know 
\[ \|\iota^\ast_j M_{i,j}\gamma\|_1 \leq \|\gamma\|_1 + 2 \mu \leq 1.5, \quad \|\iota^\ast_j M_{i,j}\gamma\|_1 \leq \mu p \|\gamma\|_1 \leq 1.5 \mu p, \]
on the event $E_{\text{cross}} \cap E_{\text{ncurv}}$, we have 
\[ -\gamma^\ast \iota^\ast M \tilde{C}_x P_{(a)} \tilde{C}_x M_{i,j} \gamma \]
where we write $\sigma_i = \text{sign}(\beta_i)$ as:

$$E_{\mathcal{I}} := \left\{ \sigma_i \chi[|\beta_i|] \leq \begin{cases} n \theta \cdot |\beta_i| (1 - E_{s_i, \text{erf}_\beta}(\lambda, s_i)) + \frac{c_{n \theta}}{p} \cdot \mu, & \forall i \in \tau \\ n \theta \cdot |\beta_i| 4 |\tau| + \frac{c_{n \theta}}{p} \cdot \mu, & \forall i \in \tau^c \end{cases} \right\},$$

and use both $\|\beta\|_1 \leq \frac{c_{n \theta}}{\sqrt{\tau}}$, $\|\beta\|_2^2 \leq \frac{c_{n \theta}}{\sqrt{\tau}}$. On this event we have

$$\beta^* \chi[|\beta|] \leq n \theta \cdot \sum_{i \in \tau} \beta_i^2 (1 - E_{s_i, \text{erf}_\beta}(\lambda, s_i)) + 4n \theta^2 |\tau| \|\beta\|_1^2 + \frac{c_{n \theta} n \theta}{p} \|\beta\|_1$$

$$\leq n \theta \cdot \sum_{i \in \tau} \beta_i^2 (1 - E_{s_i, \text{erf}_\beta}(\lambda, s_i)) + \frac{5c_{n \theta} n \theta}{\sqrt{|\tau|}}. \quad \text{(I.9)}$$

2. (Lower bound $E_{f_{\beta_i}}$) Combine the first term from each of the (I.7) and (I.9). Use $\mu \leq c_{n \theta} \leq \frac{1}{300}$ and (I.3) to obtain $(|\gamma_i| - \mu)^2 > 0.38$, we have

$$\frac{1}{n \theta} (g_1(\beta) + g_2(\beta)) \leq - \sum_{i \in J} \left[ (|\gamma_i| - \mu)^2 - \beta_i^2 \right] (1 - E_{s_i, \text{erf}_\beta}(\lambda, s_i))$$

$$+ \sum_{i \in \tau \setminus J} \beta_i^2 (1 - E_{s_i, \text{erf}_\beta}(\lambda, s_i)) - 0.38 \sum_{i \in J} E_{s_i, f_{\beta_i}(\lambda, s_i)} \quad \text{(I.10)}$$

now use Taylor expansion $^2$ for $f_{\beta_i}$, and apply the upper bound where $E_{s_i^2} \leq \theta \|\beta\|_2 \leq \theta \left( 1 + \frac{c_{n \theta}}{\sqrt{\tau}} + \frac{c_{n \theta}}{\sqrt{\tau}} \right) \leq \frac{3c_{n \theta}}{|\tau|}$

$$E_{s_i, f_{\beta_i}(\lambda, s_i)} \geq E_{s_i, \frac{1}{\sqrt{2\pi}} \left( \frac{2\lambda}{|\beta_i|} - \frac{\lambda^3}{|\beta_i|^3} \left( 1 + \frac{3s_i^2}{\lambda^2} \right) \right)}$$

$$\geq \frac{1}{\sqrt{2\pi}} \left( \frac{2\lambda}{|\beta_i|} - \frac{1}{|\beta_i|^3} \left( \lambda^3 + \frac{9c_{n \theta} \lambda}{|\tau|} \right) \right),$$

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

$$\begin{cases} f'(\beta) = 0 \implies 2\lambda \beta_i^2 = 3\lambda \left( \lambda^2 + \frac{9c_{n \theta}}{|\tau|} \right) \\ f''(\beta) = \frac{1}{|\beta_i|^3} \left( 4\lambda - \frac{12\lambda}{|\beta_i|^2} \left( \lambda^2 + \frac{9c_{n \theta}}{|\tau|} \right) \right) = \frac{1}{|\beta_i|^3} \left( 4\lambda - \frac{12\lambda}{|\beta_i|^2} \right) < 0 \end{cases}$$

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

$$0.38 \sum_{i \in J} E_{s_i, f_{\beta_i}(\lambda, s_i)} \geq 0.3 \min_{\beta = \frac{0.79}{\sqrt{|\tau|}}} f(\beta)$$

$^2$ Apply $\exp[-x^2/2] > 1 - x^2/2$
where Lemma F.2 provides the upper bound for $\gamma$ then calculate the constant for the second term in (I.14) by writing

$$
(1 - \mathbb{E}_{s(1)} \operatorname{erf}_{\lambda}(\lambda, s(1))) \geq (1 - \mathbb{E}_{s(0)} \operatorname{erf}_{\lambda}(\lambda, s(0)))
$$

then combine (I.11)-(I.12) and use

$$
2 \leq (1 + \kappa)(1 + 1) \kappa
$$

when $\beta(0) = (|\gamma(0)| - \mu)^2 - \eta$ for some $\eta > 0$. With monotonicity Lemma F.3, which implies:

$$
(1 - \mathbb{E}_{s(0)} \operatorname{erf}_{\lambda}(\lambda, s(0))) \geq (1 - \mathbb{E}_{s(1)} \operatorname{erf}_{\lambda}(\lambda, s(1)))
$$

then combine (I.11)-(I.12) and use $\mu \leq \frac{c{\mu}}{4\sqrt{r}}$ from Lemma E.5

$$
(10) \leq - \left( \left( |\gamma(0)| - \mu \right)^2 - \beta(0)^2 - \eta \right) (1 - \mathbb{E}_{s(0)} \operatorname{erf}_{\lambda}(\lambda, s(0)))
$$

On the other hand, when $\beta(0) \geq (|\gamma(0)| - \mu)^2 > 0.38$, combining (I.11)-(I.12) gives:

$$
(10) \leq \left( \|\beta\|_2^2 - \|\gamma\|_2^2 + 2\mu \|\gamma\|_1 \right) + \left( \left( |\gamma(0)| - \mu \right)^2 - \beta(0)^2 \right) \mathbb{E}_{s(0)} \operatorname{erf}_{\lambda}(\lambda, s(0))
$$

\[\begin{align*}
&+ \left( \left( |\gamma(1)| - \mu \right)^2 - \sum_{i \in \tau(0)} \beta_i^2 \right) \mathbb{E}_{s(1)} \operatorname{erf}_{\lambda}(\lambda, s(1)) - 0.38 \sum_{i \in J} \mathbb{E}_{s(1)} \operatorname{erf}_{\lambda}(\lambda, s(1))
\end{align*}\]

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

$$
\mathbb{E}_{s(1)} \operatorname{erf}_{\lambda}(\lambda, s(1)) = 1 - \frac{1}{n\theta(\beta(1))} \mathbb{E}_{X_1} \operatorname{erf}_{\lambda}(\lambda, s(1)) \leq 1 - \frac{\sigma(1)}{n\theta(\beta(1))} \mathbb{E}_{X_1} \operatorname{erf}_{\lambda}(\lambda, s(1))
$$

then calculate the constant for the second term in (I.14) by writing $\kappa = \frac{\sigma(1)}{\theta(0)} = \frac{\beta(0)}{\beta(0)} \leq \frac{5}{4}$, which provides

$$
\gamma^2(1) \leq \left( \frac{1 + \mu}{\kappa + 1} \right)^2 \text{ and } \beta^2(0) \leq \frac{\beta \gamma^2(1)}{\pi^2 + 1} \text{ where } \mu < \frac{\pi}{\kappa + 1} \text{ and by applying } |\beta(0)| > \frac{4}{3} \beta(0) \geq 0.3, \text{ we have}
$$

$$
\left( \frac{\gamma^2(1)}{\beta(1)} - 1 \right) + c_{\mu} + \beta^2(0) \leq \frac{\kappa}{\kappa(\kappa + 1)} \beta^2(0) + \kappa \beta(0) + \frac{\mu + c_{\mu}}{0.3}
$$

38
\[ \leq \frac{\kappa^2 - 1}{\sqrt{\kappa^2 + 1}} + \kappa \left( \| \beta \|_2^2 - 1 \right) + 4.2c_\mu \leq 0.36 + 6c_\mu, \]  
(I.16)

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

\[
\begin{align*}
\text{(I.10)} & \leq \frac{2c_\mu}{\sqrt{\| \tau \|}} + \sqrt{\frac{2}{\pi}} \left( \tau_{(1)}^2 - 1 + c_\mu + \beta_{(0)}^2 \right) \frac{\lambda}{|\beta_{(1)}|} - 0.6\lambda \\
& \leq \frac{2c_\mu}{\sqrt{\| \tau \|}} + \sqrt{\frac{2}{\pi}} \left( 0.36\lambda + \frac{6c_\mu c_\lambda}{0.3} \right) - 0.6\lambda \\
& \leq \frac{4c_\mu}{\sqrt{\| \tau \|}} - 0.3\lambda \\
& \leq -\frac{1}{3}\lambda \\
\end{align*}
\]

(I.17)

3. (Collect all results) Combine the components of pseudo Hessian (I.7), (I.9) with bounds for \( g_1 + g_2 \) from (I.13) and (I.17), and use Lemma E.5 which provides both \( \mu p \theta |\tau| < \frac{\lambda}{4} \) and \( \theta |\tau| < \frac{\lambda}{4} \) where \( c_\mu < \frac{1}{36} \) and \( c_\lambda \geq \frac{1}{3} \), we can obtain:

\[
v^*Hess_{\beta_\alpha}(a)v \leq g_1(\beta) + g_2(\beta) + \frac{7c_\mu n\theta}{\sqrt{\| \tau \|}} + (18\mu p + 8) n\theta^2
\]

\[
\leq n\theta \left( \frac{4c_\mu}{\sqrt{\| \tau \|}} - 0.3\lambda \right) + n\theta \cdot \frac{7c_\mu}{\sqrt{\| \tau \|}} + n\theta \cdot \frac{6.5c_\mu}{\| \tau \|}
\]

\[
\leq \frac{n\theta}{\sqrt{|\tau|}} (0.059 - 0.06) \leq -0.001n\theta\lambda
\]

Finally, the curvature is negative along \( v \) direction with probability at least

\[
1 - \mathbb{P} [ \mathcal{E}_{\text{cross}} ] = \mathbb{P} [ \mathcal{E}_{\text{ncurv}} ] = \mathbb{P} [ \mathcal{E}_{\text{curv}} ].
\]

Lemma D.4 Corollary G.3 Corollary F.4

Similarly for objective \( \varphi_{\rho} \), we have that

**Corollary I.2** (Negative curvature for \( \varphi_{\rho} \)). Suppose that \( x_0 \sim_{i.d.} BG(\theta) \) in \( \mathbb{R}^n \), and \( k, c_\mu \) such that \( (a_0, \theta, k) \)
satisfies the sparsity-coherence condition SCC(\( c_\mu \)). Define \( \lambda = c_\lambda / \sqrt{\| \tau \| \} \) in \( \varphi_{\rho} \) where \( c_\lambda \in \left[ \frac{1}{3}, \frac{1}{2} \right] \), then there exists some numerical constants \( C, c, c', c'' \), \( \epsilon > 0 \) such that if \( \rho \) is \( \delta \)-smoothed \( \ell^1 \) function where \( \delta \leq c'' \lambda \theta^8 / \rho^2 \log^2 n \), \( n > C\rho^5 \theta^{-2} \log \rho \) and \( c_\mu \leq \epsilon \), then with probability at least \( 1 - c'/n \), for every \( \alpha \in \cup_{|\tau| \leq k \rho} \mathcal{S}_\tau, \gamma(\gamma_\alpha) \) satisfying \( |\beta_{(1)}| \geq \frac{1}{3} |\beta_{(0)}| \): for \( v \in \mathcal{S}_{(\alpha,1)} \cap \mathbb{S}^{p-1} \cap a^+ \),

\[
v^*Hess[\varphi_{\rho}](a)v \leq -cn\theta\lambda
\]

(I.20)

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

\[
v^*Hess[\varphi_{\rho}](a)v \leq -cn\theta\lambda + 200' n\theta^2 \leq -cn\theta\lambda/2
\]

(I.21)

**I.2 Large gradient**

For any \( \alpha \in \mathbb{S}^{p-1} \) near subspace and the second largest correlation \( \beta_{(1)} \) much smaller then the first correlation \( \beta_{(0)} \), while not being near 0, the negative gradient of \( \varphi_{\rho}(a) \) will point at the largest shift. We show this in Lemma I.3, and the \( \varphi_{\rho} \) version in Corollary I.4 when \( \rho \) is properly defined as in Appendix H.
Lemma I.3 (Large gradient for $\varphi_{\ell^1}$). Suppose that $x_0 \sim_{\text{i.i.d.}} \text{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(a_0, \theta, k)$ satisfies the sparsity-coherence condition SCC($c_\mu$). Define $\lambda = c_\mu / \sqrt{k}$ in $\varphi_{\ell^1}$ with some $c_\mu \in \left[\frac{\ell}{2}, \frac{\ell}{4}\right]$, then there exists some numerical constants $C, c', c, \tau > 0$, such that if $n > C p^5 \theta^{-2} \log p$ and $c_\mu \leq \tau$, then with probability at least $1 - c' / n$, for every $a \in \mathcal{R}(S_{\tau}, \gamma(c_\mu))$ satisfying $\frac{4}{5} |\beta(0)| > |\beta(1)| > \frac{1}{4 \log \theta - \tau} \lambda$, we gain:

$$\left\langle \sigma_0^* \iota s_0(a_0), -\text{grad}[\varphi_{\ell^1}(a)] \right\rangle \geq c n \theta \left(\log^{-2} \theta^{-1}\right) \lambda^2$$

where $\sigma_i = \text{sign}(\beta_i)$.

Proof. 1. (Properties for $\alpha, \beta$) Define $\theta_{\text{log}} = \frac{1}{\log \theta - \tau}$, we first derive upper bound on the dominant entry $|\beta(0)|$ as follows. Write the geodesic distance between $a$ and $\iota^* s_i(a_0)$ as a function of $\beta_i$ as $d_S(a, \pm \iota^* s_i(a_0)) = \cos^{-1}(\beta_i)$, then by triangle inequality we have:

$$d_S(a, \pm \iota^* s_i(a_0)) \geq d_S(\pm \iota^* s_i(a_0), \iota^* s_i(a_0)) - d_S(a, \iota^* s_i(a_0))$$

$$\implies \cos^{-1} \pm \beta(0) \geq \cos^{-1} \mu - \cos^{-1} |\beta(1)|$$

$$\implies \pm \beta(0) \leq \cos \left(\cos^{-1} \mu - \cos^{-1} |\beta(1)|\right) = \mu |\beta(1)| + \sqrt{(1 - \mu^2) \left(1 - \beta^2(1)\right)}$$

$$\leq 1 - \frac{1}{2} \left(|\beta(1)| - \mu\right)^2.$$

Use the regional condition $|\beta(1)| \geq \frac{\theta_{\text{ub}}}{4} \lambda$ and since $\mu |\tau|^{3/2} < \frac{c_{\mu} \theta_{\text{log}}}{\theta_{\text{ub}} \lambda}$ from Definition E.1, implies

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

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

$$|\beta(0)| \geq \frac{4}{5} |\beta(1)| =: \beta_{\text{lb}}.$$  

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

$$\beta(0) = \alpha(0) + \iota^* e(0) M \alpha(\gamma(0))$$

$$\implies |\alpha(0) - \beta(0)| \leq \mu \sqrt{|\tau|} \|\alpha_{\tau}\|_2 + \mu \sqrt{|\tau|} \|\alpha_{\tau^c}\|_2$$

$$\leq \frac{c_{\mu} \theta^2_{\text{log}} (1 + c_\mu)}{4 |\tau|} + \mu \sqrt{|\tau|} \leq \frac{c_{\mu} \theta^2_{\text{log}}}{|\tau|}.$$  

and therefore $|\alpha(0)| \leq |\beta(0)| + \frac{c_{\mu} \theta^2_{\text{log}}}{|\tau|} \leq 1 - 0.49 \left(\frac{\theta_{\text{ub}}}{4} \lambda\right)^2 + \frac{c_{\mu} \theta^2_{\text{log}}}{|\tau|} < 1.$

2. (Upper bound of $\beta^* \chi[\beta]$) Define a piecewise smooth convex upper bound $h$ for $\beta^* \chi[\beta]$ as:

$$h(\beta_i) := \begin{cases} \beta^2_i - \nu_1 \lambda |\beta_i| & |\beta_i| \geq \nu_1 \lambda \\ \frac{1}{2} \beta^2_i & |\beta_i| \leq \nu_1 \lambda \end{cases}$$

then Lemma N.7 tells us since $\|\beta_{\tau \setminus \{0\}}\|_{\infty} \leq \beta(1)$:

$$\sum_{i \in \tau \setminus \{0\}} h(\beta_i) \leq \|\beta_{\tau \setminus \{0\}}\|_2 \left(1 - \frac{\nu_1 \lambda \beta(1)}{2 \beta^2(1)}\right) \leq \left(1 + \frac{c_{\mu} \theta^2_{\text{log}}}{|\tau|} \right) \left(1 - \frac{\nu_1 \lambda}{2 \beta^2(1)}\right)$$

$$\leq \left(1 - \frac{\nu_1 \lambda}{2 \beta(1)}\right) \left(1 - \beta^2(0)\right) + \frac{c_{\mu} \theta^2_{\text{log}}}{|\tau|},$$

40
then condition on the following event using Corollary F.4,

\[ \mathcal{E}_\tau := \left\{ \beta, \chi[\beta]_i \leq \begin{cases} \frac{n \theta \cdot h(\beta_i) + \frac{c_\mu \theta}{p^{1/2}} |\beta_i|}{\sqrt{\tau}}, & \forall i \in \tau \setminus (0) \\ \frac{n \theta \cdot 4^2 \theta |\tau| + \frac{c_\mu \theta}{p^{3/4}} |\beta_i|}{\sqrt{\tau}}, & \forall i \in \tau^{-} \end{cases} \right\}, \]

which provides the upper bound of \( \beta^* \chi[\beta] \) by applying \( 5p > \log^{8/3}(p \log^2 p) > (\theta_{\text{log}}^2)^{1/3} \) from lower bound of \( \theta \) from Definition E.1, \( \| \beta^* \|_2 \leq \frac{c_\mu \theta \log}{\sqrt{\tau}} \) from Lemma E.5, \( |\tau| \leq \sqrt{\tau} \) from lemma assumption and let \( c_\mu < \frac{1}{100} \):

\[
\beta^* \chi[\beta] \leq \chi[\beta](0) + \sum_{i \in \tau \setminus (0)} \beta_i \chi[\beta]_i + \langle \beta_{\tau^{-}}, \chi[\beta]_{\tau^{-}} \rangle
\]

\[
\leq \chi[\beta](0) + \sum_{i \in \tau \setminus (0)} \theta \cdot \sum_{i \in \tau \setminus (0)} h(\beta_i) + 4 \theta^2 |\tau| \| \beta_{\tau^{-}} \|_2^2
\]

\[
+ \frac{c_\mu \theta}{p^{1/2}} \left( \sqrt{\tau} \| \beta_{\tau^{-}} \|_2 + \sqrt{\tau} \| \beta_{\tau_{\text{log}}} \|_2 \right)
\]

\[
\leq \chi[\beta](0) + \sum_{i \in \tau \setminus (0)} \theta \cdot \eta \left( 1 - \beta^2(0) \right) + \theta \cdot \frac{c_\mu \theta^2}{\| \tau \|_2^2} + \frac{4 \theta^2 |\tau| c_\mu \theta^2}{\theta |\tau|_2^2}
\]

\[
+ \frac{c_\mu \theta}{p^{1/4} |\tau|} \left( 1 + \frac{c_\mu \theta \log}{p \sqrt{\theta} |\tau|} \right)
\]

\[
\leq \chi[\beta](0) + \sum_{i \in \tau \setminus (0)} \eta \left( 1 - \beta^2(0) \right) + \frac{4 \theta^2 |\tau| c_\mu \theta^2}{\theta |\tau|_2^2}
\]

(1.26)

where \( \eta = 1 - \frac{c_\mu \theta \log}{2 \beta(1)} \).

3. (Align the gradient with \( \ell^* s(0)[a_0] \)) Based on the definition \( \beta \), since \( \beta(0) = \langle a, \ell^* s(0)[a_0] \rangle \), we can expect that the negative gradient is likely aligned with direction toward one of the candidate solution \( \pm \ell^* s(0)[a_0] \). Wlog assume that both \( \beta(0), \beta(1) \) are positive, then expand the gradient and use incoherent property for \( a_0 \) Lemma E.4 we have:

\[
\langle \ell^* s(0)[a_0], -\text{grad}_{\chi[\beta]} [a] \rangle = \langle \ell^* s(0)[a_0], \ell^* C_{a_0} \chi[\beta] - \beta^* \chi[\beta] \alpha \rangle
\]

\[
\geq \left( \chi[\beta](0) - \beta^* \chi[\beta] \alpha(0) \right) - \mu \| \chi[\beta](0) - \beta^* \chi[\beta] \alpha(0) \|_1,
\]

(1.27)

where \( \setminus (0) \) is an abbreviation of the complement set \( \{ \pm 2p_n \} \setminus (0) \). The latter part of (1.27) has an upper bound using bounds of \( \beta^* \chi[\beta] < \frac{3 \theta^2}{2}, \| \chi[\beta] \|_2 < \frac{3 \theta^2}{20} \) from (1.62), and \( \| \chi[\beta] \|_2 \leq n \theta \| \beta_{\tau(0)} \|_2 \) in event \( \mathcal{E}_\tau \), we obtain:

\[
\mu \| \chi[\beta](0) - \beta^* \chi[\beta] \alpha(0) \|_1
\]

\[
\leq \mu \left( \sqrt{\tau} \| \chi[\beta] \|_2 + \beta^* \chi[\beta] \sqrt{\tau} \| \alpha_{\tau^{-}} \|_2 + 3 \mu \sqrt{|\gamma_2|} \right)
\]

\[
\leq \mu \left( \sqrt{\tau} \| \chi[\beta] \|_2 + \beta^* \chi[\beta] \sqrt{\tau} \| \alpha_{\tau^{-}} \|_2 + 3 \mu \sqrt{|\gamma_2|} \right)
\]

\[
\leq \frac{c_\mu \theta^2}{4 |\tau|} \left( 2 \left( 1 + c_\mu \right) \right) \left( 1 + \frac{3}{2} c_\mu \right)
\]

(1.28)
On the other hand, the former term of (I.27) possesses a lower bound using (I.25)-(I.26), \( \chi[\beta](0) > n\theta \left( \beta(0) - \frac{\nu}{2} \lambda - \frac{c_\mu}{5} \right) \geq n\theta \left( \beta(0) - 0.51\nu_1\lambda \right) \) and \( \alpha(0) \leq 1: \\
\chi[\beta](0) - \beta^2 \chi[\beta] \alpha(0) \\
\geq (1 - \alpha(0) \beta(0)) \chi[\beta](0) - n\theta \cdot \left[ \eta \left( 1 - \beta^2(0) \right) + \frac{6c_\mu \theta^2 \log}{|\tau|} \right] \alpha(0) \\
\geq n\theta \left[ 1 - \left( \beta(0) + \frac{c_\mu \theta^2 \log}{|\tau|} \right) \beta(0) \right] (\beta(0) - 0.51\nu_1\lambda) \\
- n\theta \left[ \eta \left( 1 - \beta^2(0) \right) \beta(0) + \frac{c_\mu \theta^2 \log}{|\tau|} + \frac{6c_\mu \theta^2 \log}{\alpha(0)} \right] \\
\geq n\theta \left[ \left( 1 - \beta^2(0) \right) \left( \beta(0) - 0.51\nu_1\lambda \right) - \frac{c_\mu \theta^2 \log}{|\tau|} \beta(0) \right] \\
- \left( 1 - \beta^2(0) \right) \eta \beta(0) - \eta \frac{c_\mu \theta\log}{|\tau|} \left( 1 - \beta^2(0) \right) - \frac{6c_\mu \theta^2 \log}{|\tau|} \\
\geq n\theta \left[ \left( 1 - \beta^2(0) \right) \left( (1 - \eta) \beta(0) - 0.51\nu_1\lambda \right) - \frac{c_\mu \theta^2 \log}{|\tau|} \left( (1 - \eta) \beta^2(0) + 7 \right) \right], \tag{I.29} \\

combine (I.27) with (I.28)-(I.29) and \( \eta > 0 \), we have \\
\( (I.27) \geq n\theta \left[ \left( 1 - \beta^2(0) \right) \left( (1 - \eta) \beta(0) - 0.51\nu_1\lambda \right) - \frac{c_\mu \theta^2 \log}{|\tau|} \left( (1 - \eta) \beta^2(0) + 7 \right) \right] \\
- n\theta \cdot \frac{c_\mu \theta^2 \log}{|\tau|} \left( 0.5 + c_\mu - 0.5\beta(0) \right) \\
\geq n\theta \left[ \left( 1 - \beta^2(0) \right) \left( \frac{\nu_1 \lambda}{2\beta(1)} \beta(0) - 0.51\nu_1\lambda \right) - \frac{8c_\mu \theta^2 \log}{|\tau|} \right] Big, \tag{I.30} \\

4. (Lower bound of \( f(\beta) \)) Given a fixed \( \beta(1) \), the cubic function \( f(\beta(0)) \) has zeros set \( \beta(0) \in \{ \pm 1, 1.02\beta(1) \} \) and has negative leading coefficient. Combine with the condition of \( \beta(0) \in \{ \beta_{lb}, \beta_{ub} \} \) from (I.23)-(I.24), we can observe that \\
\( \beta(0) \in [\beta_{lb}, \beta_{ub}] = \left[ \frac{5}{4} \beta(1), 1 - 0.49\beta^2(1) \right] \subseteq \left[ 1.02\beta(1), 1 \right], \)

therefore the cubic term is always positive and minimizer is either one of the boundary point. When \( \beta(0) = \beta_{lb} \), 
use \( (1 + \frac{25}{16}) \beta^2(1) < 1.01 \), and use \( \nu_1\lambda < \frac{\theta\log}{2\sqrt{|\tau|}} \leq \frac{1}{2\sqrt{7}} \), since \(|\tau| \geq 2\), we have: \\
\( f(\beta_{lb}) \geq (1 - \beta^2_{lb}) \left( \frac{\nu_1 \lambda}{2\beta(1)} \beta_{lb} - 0.51\nu_1\lambda \right) \geq (1 - 0.616) \cdot \left( \frac{5}{8} - 0.51 \right) \nu_1\lambda 

42
\[
\frac{1}{10\sqrt{2}}\nu_1 \lambda \geq \frac{\theta^2_{\log}}{32} \lambda^2.
\]

(I.31)

On the other hand when \(\beta_{(0)} = \beta_{ub}\):

\[
f(\beta_{ub}) \geq (1 - \beta_{ub}^2) \left( \frac{\nu_1 \lambda}{\beta_{ub}} - 0.51 \nu_1 \lambda \right) \\
\geq 0.49 \beta_{ub}^2 \cdot \left( \frac{\nu_1 \lambda}{\beta_{ub}} - 0.49 \beta_{ub}^2 \right) - 0.51 \nu_1 \lambda,
\]

which is a cubic function of \(\beta_{(1)}\) with negative leading coefficient, whose zeros set is \([-0.73, 0, 2.81]\). Thus it minimizes at the boundary points of \(\beta_{(1)} \in \left[\frac{\lambda_{(1)}}{4 \log \theta_{-1}}, 1\right] \subset [0, 2.81]\), thus assign \(\beta_{(1)} = \frac{\lambda_{(1)}}{4 \log \theta_{-1}}\), we have:

\[
f(\beta_{ub}) \geq 0.49 \left( \frac{\lambda_{(1)}}{4 \log \theta_{-1}} \right)^2 \cdot \left( \frac{1}{2} \left( 1 - 0.49 \left( \frac{\lambda_{(1)}}{4 \log \theta_{-1}} \right)^2 \right) - 0.51 \nu_1 \lambda \right) \\
\geq \frac{1}{6} \left( \frac{\lambda_{(1)}}{4 \log \theta_{-1}} \right)^2 \geq \frac{\theta^2_{\log}}{96} \lambda^2.
\]

(I.32)

Finally combine (I.30) with the lower bound of cubic function (I.31)-(I.32) together with condition \(c_{\mu} < \frac{\theta^2_{\log}}{800}\) and \(\nu_1 = \sqrt{\frac{\theta_{\log}}{2}}\), obtain

\[
\langle \tau^* s(0)[a_0], -\text{grad}_{\varphi_{\nu}}[a] \rangle \geq n \theta \cdot \left( \min \{ f(\beta_{ub}), f(\beta_{ub}) \} - \frac{8c_{\mu} \theta_{\log}^2}{|\tau|} \right) \\
\geq n \theta \left( \frac{\theta_{\log}^2 c_{\mu}^2}{96 |\tau|} - \frac{8 \theta_{\log}^2 c_{\mu}^2}{800 |\tau|} \right) \geq 6 \times 10^{-3} n \theta_{\log}^2 c_{\mu}^2.
\]

(I.33)

The proof for the case where \(\beta_{(0)}\) negative can be derived in the same manner.

As a consequence, we have that

**Corollary I.4** (Large gradient for \(\varphi_{\rho}\)). Suppose that \(x_0 \sim_{\text{i.i.d.}} \text{BG}(\theta)\) in \(\mathbb{R}^n\), and \(k, c_{\mu}\) such that \((a_0, \theta, k)\) satisfies the sparsity-coherence condition \(\text{SCC}(c_{\mu})\). Define \(\lambda = c_{\mu}/\sqrt{k}\) in \(\varphi_{\rho}\) with \(c_{\lambda} \in \left[\frac{1}{5}, \frac{1}{4}\right]\), then there exists some numerical constants \(C, c, c', c'', \tau > 0\) such that if \(\rho\) is \(\delta\)-smoothed \(\ell^1\) function where \(\delta \leq c' \lambda \delta^3 / p^2 \log^2 n\) with \(n > C p^3 \theta^{-2} \log p\) and \(c_{\mu} \leq \tau\), then with probability at least \(1 - c' / n\), for every \(a \in \cup_{|\tau| \leq k} \mathcal{R}(S_{\theta}, \gamma(c_{\mu}))\) satisfying \(\frac{1}{5} |\beta_{(0)}| > |\beta_{(1)}| > \frac{1}{4 \log \theta_{-1}}\),

\[
\langle \sigma(0) \tau^* s(0)[a_0], -\text{grad}[\varphi_{\rho}](a) \rangle \geq cn \theta \left( \log^{-2} \theta^{-1} \right) \lambda^2
\]

where \(\sigma_i = \text{sign}(\beta_i)\).

**Proof.** Choose \(\tau^* s(0)[a_0]\) as in Lemma I.3, and apply (H.22) from Lemma H.6 with the constant multiplier of \(\delta\) satisfies \(c'' < c'/4\), then utilize \(\theta |\tau| \log^2 \theta^{-1} < c_{\mu}\) from Definition E.1 we have

\[
\langle \sigma(0) \tau^* s(0)[a_0], -\text{grad}[\varphi_{\rho}](a) \rangle \geq cn \theta (\log^{-2} \theta^{-1}) \lambda - c'' n \theta^2 \geq cn \theta (\log^{-2} \theta^{-1}) \lambda / 2
\]

(I.35)

**I.3 Convex near solutions**

For any \(a \in \mathbb{S}^{p-1}\) near subspace and the second largest correlation \(\beta_{(1)}\) smaller then \(\frac{1}{4 \log \theta_{-1}}\), then \(\varphi_{\rho}\) will be strongly convex at \(a\). We show this in Lemma I.5, and the \(\varphi_{\rho}\) version in Corollary I.6 when \(\rho\) is properly defined as in Appendix H.
Lemma I.5 (Strong convexity of $\varphi_{\ell_1}$ near shift). Suppose that $x_0 \sim_{i.i.d.} \text{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(a_0, \theta, k)$ satisfies the sparsity-coherence condition $\text{SCC}(c_\mu)$. Define $\lambda = c_\lambda / \sqrt{k}$ in $\varphi_{\ell_1}$, with $c_\lambda \in \left[ \frac{1}{2}, \frac{1}{4} \right]$, then there exists some numerical constants $C, c, c' > 0$ such that if $n > Cp^2 \theta^{-2} \log p$ and $c_\mu \leq \tau$, then with probability at least $1 - c'/n$, for every $a \in \bigcup_{|\tau| \leq k} \mathcal{R}(S_\tau, \gamma(c_\mu))$ satisfying $|\beta(\ell)| \leq \frac{1}{3 \log \theta - \tau} \lambda$, for all $v \in \mathbb{S}^{p-1} \cap v^\perp$,
\[
v^* \tilde{\text{Hess}}[\varphi_{\ell_1}](a)v > cn\theta;
\]

furthermore, there exists $\bar{a}$ as an local minimizer such that
\[
\min_i \| \bar{a} - s_i[a_0]\|_2 \leq \frac{1}{2} \max \{\mu, p^{-1}\}.
\]

Proof. 1. (Expectation of $\chi$ near shifts) We will write $x$ as $x_0$ throughout this proof. When $a$ is near one of the shift, the $\chi$ operator shrinks all other smaller entries of correlation vector $\beta_{\ell_1}$ in an even larger shrinking ratio. Firstly we can show $|\langle \beta_{\ell_1}(0), x_{\ell_1}(0) \rangle|$ is no larger then $\lambda/2$ with probability at least $1 - 4\theta$, since
\[
\mathbb{P} \left[ |\langle \beta_{\ell_1}(0), x_{\ell_1}(0) \rangle| > \frac{\lambda}{2} \right] \leq \mathbb{P} \left[ |\langle \beta_{\ell_1}(0), x_{\ell_1}(0) \rangle| > \frac{2\lambda}{5} \right] + \mathbb{P} \left[ |\langle \beta_{\ell_1}, x_{\ell_1} \rangle| > \frac{\lambda}{10} \right] \leq 4\theta
\]
via Corollary E.6 and Corollary E.7. Now recall from Lemma F.2 and the derivation of (F.10)-(F.11), we know for every $i \neq (0)$,
\[
\sigma_i \mathbb{E} \chi_i \beta_i = n\theta |\beta_i| \mathbb{E}_{s_i} [1 - \text{erf}(\lambda, s_i)]
\]
\[
\leq n\theta |\beta_i| \mathbb{E}_{s_i, x_i} \left[ g^2 \mathbb{E}_{|\beta|, g, s_i} \{ |\beta + g_0 x_{\ell_1}(0) + g_0 x_{\ell_1}(0), 0| > \lambda \} \right]
\]
\[
\leq n\theta |\beta_i| \left( \mathbb{E} g^2 \mathbb{E}_{|\beta|, g} \{ |\beta + g_0 x_{\ell_1}(0) > \lambda/2 \} + \mathbb{P} \left[ |x_{\ell_1}(0)| \neq 0 \right] \right)
\]
\[
+ \mathbb{P} \left[ |\langle \beta_{\ell_1}(0), x_{\ell_1}(0) \rangle| > \frac{\lambda}{2} \right] \leq n\theta |\beta_i| \left( (\mathbb{E} g^2)^{1/2} \mathbb{P} \left[ |\beta(\ell_1)| > \frac{\lambda}{2} \right]^{1/2} + \theta + 4\theta \right)
\]
\[
\leq n\theta |\beta_i| \left( \exp(-\log^2 \theta^{-1}) + 5\theta \right)
\]
\[
\leq 6n\theta^2 |\beta_i|
\]
where the third inequality is derived using union bound; the fourth inequality is the result of (I.38), and
the fifth inequality is derived from Gaussian tail bound lemma N.1.

2. (Local strong convexity) Let $\gamma = C_{a,\ell}^* v\nu$, for any $\| v \|_2 = 1$ we have $\| \gamma \|_2^2 \leq 1 + \mu p$. Furthermore:
\[
|\gamma(0)| = |\langle \nu^* s_{(0)}(a_0), v \rangle| = |\langle P_{a,\ell}^* v^* s_{(0)}(a_0), v \rangle| = |\langle \nu^* s_{(0)}(a_0) - \beta(0) a, v \rangle|
\]
\[
\leq \| \nu^* s_{(0)}(a_0) - \beta(0) a \|_2 \leq \sqrt{1 - \beta_{(0)}^2};
\]

Consider any such $v$, the pseudo Hessian can be lower bounded as
\[
v^* \tilde{\text{Hess}}[\varphi_{\ell_1}](a)v = -\gamma^* \tilde{C}_x P_{\ell}(a) \tilde{C}_x^* \gamma
\]
\[
\geq -\gamma_{(0)}^2 \| P_{\ell}(a) \tilde{C}_x e_{(0)} \|_2^2 - \sum_{i \neq (0)} \| P_{\ell}(a) \tilde{C}_x e_i \|_2^2 \gamma_i^2
\]
\[
- 2 \sum_{i \neq j} |e_i^* \tilde{C}_x P_{\ell}(a) \tilde{C}_x e_j| |\gamma_i| |\gamma_j|
\]
\[
\geq - \left( 1 - \beta_{(0)}^2 \right) \|x\|^2 - \max_{i \neq (0)} \left\| P_{i(a)}^{s-i} [x] \right\|^2 \|\gamma\|^2 - 2 \max_{i \neq j} \left\{ e_i^T \tilde{C}_x P_{i(a)} \tilde{C}_x e_j \right\} \|\gamma\|^2,
\]

where the second term is bounded by its expectation derived in Lemma G.2, and utilize \( \mathbb{P} \| s_i \| > \lambda / 2 < 4\theta \) from (I.38), \( \mathbb{E} \chi \) from (I.39) and regional condition \( |\beta(1)| \leq \frac{\lambda}{4 \log \theta} \) to acquire

\[
\mathbb{E} \| P_{i(a)}^{s-i} [x] \|^2 = n \theta \left[ 1 - E_{c} \operatorname{erf} \beta_i (\lambda, s_i) + E_{a}, f \beta_i (\lambda, s_i) \right] \\
\leq \frac{\mathbb{E} \chi \|\beta\|}{|\beta|} + n \theta \cdot \left( \max_{|s_i| \leq \frac{\lambda}{2}} f \beta_i (\lambda, s_i) + \mathbb{P} \left[ |s_i| > \frac{\lambda}{2} \right] \right) \\
\leq 6n \theta^2 + \frac{2n \theta}{\sqrt{2\pi}} \max_{|s_i| \leq \frac{\lambda}{2}} \left( \frac{\lambda}{|\beta|} \cdot \exp \left[ -\frac{(\lambda - |s_i|)^2}{2|\beta|^2} \right] \right) + 4n \theta^2 \\
\leq 10n \theta^2 + n \theta \cdot \log \theta^{-1} \exp \left( -2 \log \theta^{-1} \right) \\
\leq 11n \theta^2,
\]

and define the events \( E_{\|x\|^2}, E_{\text{cross}} \) and \( E_{\text{pcurv}} \) as follows:

\[
\begin{align*}
E_{\text{pcurv}} &:= \left\{ \forall a \in \cup_{|r| \leq k} \mathcal{R}(\mathcal{S}_r, \gamma(c_\mu)), \| P_{i(a)}^{s-i} [x] \|^2 \leq 11n \theta^2 + \frac{\zeta_n \theta}{\eta} \right\} \\
E_{\text{cross}} &:= \left\{ \forall a \in \cup_{|r| \leq k} \mathcal{R}(\mathcal{S}_r, \gamma(c_\mu)), |\beta(1)| \leq \frac{\lambda}{4 \log \theta \eta}, \max_{i \neq j \in [\pm \eta]} \left\{ e_i^T \tilde{C}_x P_{i(a)} \tilde{C}_x e_j \right\} \leq 8n \theta^3 \right\} \\
E_{\|x\|^2} &:= \left\{ \|x\|^2 \leq n \theta + 3 \sqrt{n \theta \log n} \right\}.
\end{align*}
\]

For the Hessian term, on the event \( E_{\text{pcurv}} \cap E_{\text{cross}} \cap E_{\|x\|^2} \), and use all \( \mu \eta \theta^2, \mu \eta \theta |r| \) and \( \theta \sqrt{\eta} \) are all less then \( \frac{\zeta_n \theta}{4 \log \theta \eta} \), from Lemma E.5, and from lemma assumption with sufficiently large \( C \) we have \( n > \theta^{-1} 36 \log^2 n \), thus \( v^* \nabla^2 \varphi_{\ell}(a) v \) can be lower bounded from (I.41) as

\[
v^* \nabla^2 \varphi_{\ell}(a) v \geq - \left( 1 - \beta_{(0)}^2 \right) \left( n \theta + 3 \sqrt{n \theta \log n} \right) - (1 + \mu \eta) \left( 11n \theta^2 + \frac{\zeta_n \eta}{\eta} \right) - 8 \eta (1 + \mu \eta) \cdot 8n \theta^3 \\
\geq - \frac{1}{2} n \theta \cdot (1 - \beta_{(0)}^2) - n \theta \cdot \left( \frac{11 \mu \eta}{4} + \frac{\zeta_n \eta}{\eta} + \frac{64 \mu \eta}{4} + \frac{64 \mu \eta}{4} \right) \\
\geq - \frac{1}{2} n \theta \cdot (1 - \beta_{(0)}^2 + 20 \mu \eta).
\]

The bounds of \( \beta^\ast \chi[\beta] \) can be derive on the event whose expectation is drawn from Lemma F.2 and (I.39) as

\[
E_{\chi} := \left\{ \begin{align*}
\sigma_i \chi[\beta_i] &\geq n \theta \mathcal{S}_v \lambda \|\beta_i\| - \frac{\zeta_n \eta}{\eta}, & \forall i \in [\pm \eta] \\
\sigma_i \chi[\beta_i] &\leq 6n \theta^2 |\beta_i| + \frac{\zeta_n \eta}{\eta \gamma}, & \forall i \neq (0)
\end{align*} \right\},
\]

then use \( \|\beta\|_{1} \leq 1 + \frac{\lambda \eta}{4 \log \theta \eta} \leq \frac{\lambda \eta}{2} \), implies:

\[
\beta^\ast \chi[\beta] \geq n \theta |\beta(0)| \left( |\beta(0)| - \nu_2 \lambda \right) - \xi \|\beta\|_{1} \frac{n \theta}{\eta} \\
\geq n \theta \left( \beta_{(0)}^2 - \sqrt{\frac{2}{\pi}} \lambda - \frac{\xi}{\eta} \right) \\
\geq n \theta \left( \beta_{(0)}^2 - \lambda \right).
\]
Finally via the regional condition $|\beta(1)| \leq \frac{\lambda}{4\log \theta}$, the absolute value of leading correlation
\[
\beta_0^2 \geq \|\beta\|_2^2 - |\tau| |\beta(1)| \geq 1 - 2c_\mu - 0.1^2 > 0.9,
\] (I.46)
then we collect all above results and obtain:
\[
v^*\tilde{\text{Hess}}[\varphi_\ell](a) v = v^*\nabla^2 \varphi_\ell(a) v - \beta^* \xi[a]\geq \left(1.5\beta_0^2 - 0.5 - \lambda - 20c_\mu\right)n\theta \geq 0.3n\theta,
\] (I.47)
with probability at least
\[
1 - \mathbb{P}\left[\mathcal{E}_\text{cross}\right] - \mathbb{P}\left[\mathcal{E}_\text{curv}\right] - \mathbb{P}\left[\mathcal{E}_{\|\xi\|_2}\right] - \mathbb{P}\left[\mathcal{E}_\mu\right] \geq 1 - 2c'/n.
\] (I.48)

3. (Identify local minima) Wlog let $a_*$ be a local minimum where its gradient is zero that is close to $a_0$. The strong convexity (I.47), provides the upper bound on $\|a_* - a_0\|_2$ via
\[
\varphi_\ell(a_*) \geq \varphi_\ell(a_0) + \langle a_* - a_0, \text{grad}[\varphi_\ell](a_0) \rangle + \frac{0.3}{2} n\theta \|a_* - a_0\|_2^2
\]
\[
\implies \|\text{grad}[\varphi_\ell](a_0)\|_2 \geq 0.15n\theta \|a_* - a_0\|_2
\] (I.49)
Thus we only require to bound the gradient at $a_0$, whose coefficients $\alpha = e_0$ and correlation $\beta$ has properties $\beta_0 = 1$ and $\|\beta_0\|_\infty \leq \mu$ hence $\|\beta\|_\infty \leq \sqrt{2\mu}$. Expand the gradient term and condition on $\mathcal{E}_\chi$, since $\mu p^2 \theta^2 \leq \frac{c_\mu}{4}$ and $\theta < \frac{1}{4\sqrt{p}}$, we can upper bound the gradient at $a_0$ as
\[
\|\text{grad}[\varphi_\ell](a_0)\|_2 = \|v^* C_{a_0} \chi[\beta - \beta^* \chi[\beta] e_0]\|_2 \leq \|v^* C_{a_0}\|_2 \|\chi[\beta] e_0\|_2
\]
\[
\leq \sqrt{1 + \mu P}(6n\theta \|\beta_0\|_2 + n\theta \cdot \frac{c_\mu}{p^{1/2}} \cdot \sqrt{2p})
\]
\[
\leq n\theta \sqrt{1 + \mu P}(6\mu \sqrt{2p} \cdot \theta + \frac{2c_\mu}{p})
\]
\[
\leq n\theta \left(3c_\mu \mu + 6\mu \cdot \sqrt{2p} \cdot p\theta + \frac{2c_\mu}{p} + \frac{2c_{s\ell} \sqrt{2p}}{p}\right)
\]
\[
\leq 7\sqrt{c_{s\ell}} n\theta \cdot \max\left\{\mu, \frac{1}{p}\right\}.
\] (I.50)
Thus we conclude that with sufficiently small $c_{s\ell}$:
\[
\|a_* - a_0\|_2 \leq 50\sqrt{c_{s\ell}} \max\left\{\mu, p^{-1}\right\} \leq \frac{1}{2} \max\left\{\mu, p^{-1}\right\},
\] (I.51)
and we complete the proof by generalize this result from minima near $a_0$ to any of its shifts $s_\ell a_0$.

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

**Corollary 1.6** (Strong convexity of $\varphi_\rho$ of near shift). Suppose that $x_0 \sim_{i.i.d.} \text{BG}(\theta)$ in $\mathbb{R}^n$, and $k, c_\mu$ such that $(a_0, \theta, k)$ satisfies the sparsity-coherence condition $\text{SCC}(c_\mu)$. Define $\lambda = c_\lambda / \sqrt{k}$ in $\varphi_\rho$ with $c_\lambda \in [\frac{1}{4}, \frac{1}{2}]$, then there exists some numerical constant $C, c, c', c'' > 0$ such that if $\rho$ is $\delta$-smoothed $\ell^2$ function where $\delta \leq c' \lambda \theta^p / p^2 \log^2 n$ and $n > Cp^p \theta^{-2} \log p$ and $c_\mu \leq \tau$, then with probability at least $1 - c'/n$, for every $a \in \cup_{|\tau| \leq \delta}(S_\tau, \gamma(c_\mu))$ satisfying $|\beta(1)| < \nu_1 \lambda$ for all $v \in S_{p-1} \cap a^\perp$,
\[
v^*\tilde{\text{Hess}}[\varphi_\rho](a) v > cn\theta;
\] (I.52)
furthermore, there exists $a$ as an local minimizer such that
\[
\min \|a - s_\ell a_0\|_2 \leq \frac{1}{2} \max \left\{\mu, p^{-1}\right\}
\] (I.53)
Proof. The strong convexity (I.52) is derived by combining (I.36) and (H.23) by letting constant multiplier of $\delta$ satisfies $c^{1/4} < 10^{-3}c$. On the other hand the local minimizer near solution (I.53) is derived via combining (I.49), (H.21) and utilize both $\theta \sqrt{k} < c_{\mu}$ and $\mu p^2 \theta^2 < c_{\mu}$ such that:

\[
\|\text{grad}[\varphi_{p}](a)\|_2 \leq \|\ell \ast C_{a_0}\|_2 \|\chi[\beta] - \tilde{C}_{\alpha_{0}} S^{c}_{\delta} \tilde{C}_{y} a\|_2 + \|\ell \ast C_{a_0}\|_2 \|\chi[\beta]_0\|_2 \\
\leq \sqrt{1 + \mu p \cdot n \theta^3} + \sqrt{c_{\mu} n \theta} \cdot \max \{\mu, p^{-1}\} \\
\leq \sqrt{8 n \theta \sqrt{c_{\mu} \cdot \max \{\mu, p^{-1}\}}}
\]

(I.54)

I.4 Retraction toward subspace

As in Figure 6, the function value grows in direction away from subspace $S_{\tau}$, we will illustrate this phenomenon by proving the negative gradient direction $-g$ will point toward the subspace $S_{\tau}$. To show this, we prove for every coefficients of $a$ as $a_{\tau}$, there exists coefficients of $g$ as $\zeta$ satisfies

\[
\left\langle \alpha_{\tau}, (\varphi_{\ell})(a) \right\rangle > c \|\alpha_{\tau}\|_2 \|\zeta_{\tau}\|_2
\]

whenever $d_{\tau}(a, S_{\tau}) \in \left[\gamma, \gamma\right]$. Apparently, the gradient will decrease $d_{\tau}(a, S_{\tau})$, hence being addressed as retractive toward subspace $S_{\tau}$. This retractive phenomenon is true for gradient of both $\varphi_{\ell}$ and $\varphi_{p}$.

Lemma I.7 (Retraction of $\varphi_{\ell}$ toward subspace). Suppose that $a_{0} \sim \text{i.i.d.}$ in $\mathbb{R}^{n}$, and $k, c_{\mu}$ such that $d_{\tau}(a_{0}, \theta, k)$ satisfies the sparsity-coherence condition SCC($c_{\mu}$). Define $\gamma = c_{\mu}/\sqrt{k}$ in $\varphi_{\ell}$ with $c_{\mu} \in \left(0, 1\right]$, then there exists some numerical constants $C, c, \tau > 0$ such that for $n_{\lambda} > C p^2 \theta^2 \log p$ and $c_{\mu} \leq \tau$, then with probability at least $1 - c' / n$, for every $a \in \cup_{|\tau| \leq k} \mathbb{H}(S_{\tau}, \gamma(c_{\mu}))$ such that if

\[
d_{\tau}(a, S_{\tau}) \geq \gamma(c_{\mu})/2
\]

then for every $a$ satisfying $a = \ell \ast C_{a_0} \alpha$, there exists some $\zeta$ satisfying $\text{grad}[\varphi_{\ell}](a) = \ell \ast C_{a_0} \zeta$ that

\[
\left\langle \zeta_{\tau}, \alpha_{\tau} \right\rangle \geq \frac{1}{4 \theta} \|\zeta_{\tau}\|_2^2.
\]

(I.57)

Proof. Write $\gamma = \gamma(c_{\mu})$. Recall the gradient can be derived as

\[
\text{grad}[\varphi_{\ell}](a) = -P_{a} \ast \ell \ast C_{a_0} \chi[\beta] = (a a^{*} - I) \ast C_{a_0} \chi[\beta] = \ell \ast C_{a_0} (b \ast \chi[\beta]) (a \ast - \chi[\beta]),
\]

for every $a$ satisfies $a = \ell \ast C_{a_0} \alpha$. Now via Corollary F.4, condition on the event:

\[
E_{\chi} := \left\{ \sigma_{i}, \chi[\beta]_i \leq \left\{ n \theta \cdot |\beta|, \frac{c_{\mu} n \theta}{p}, \frac{n \theta}{|\beta|}, 4 \theta |\tau| + \frac{c_{\mu} n \theta}{p}, \forall i \in \tau, \sigma_{i}, \chi[\beta]_i \geq n \theta \cdot S_{\delta} \sqrt{2/\pi \lambda} \right\} \right\},
\]

and on this event, utilize Lemma E.5, bounds of $b \ast \chi[\beta]$ and $\|\chi[\beta]_\tau\|_2$ can be derived with $c_{\mu} < \frac{1}{100}$ as:

\[
b \ast \chi[\beta] \leq n \theta \left( \|\beta\|_2, 4 \theta |\tau|, \|\beta\|_2 + c_{\mu} \right) \geq n \theta \left( 1 + c_{\mu} + 4 c_{\mu}^2 + c_{\mu} \right) \leq \frac{2}{3} n \theta
\]

(I.60)

\[
b \ast \chi[\beta] \geq n \theta \left( \|\beta\|_2, -\sqrt{2/\pi \lambda} \|\beta\|_1 + c_{\mu} \right) \geq n \theta \left( 1 - 4 c_{\mu} - \sqrt{2/\pi \lambda} c_{\mu} - c_{\mu} \right) \geq \frac{1}{3} n \theta
\]

(I.61)

\[
\|\chi[\beta]_\tau\|_2 \leq 4 n \theta^2 |\tau|, \|\beta\|_2 + \frac{c_{\mu} n \theta}{p}, \sqrt{2} \leq n \theta \left( 4 c_{\mu} \gamma + c_{\mu} \gamma \right) \leq \frac{1}{20} n \theta \gamma.
\]

(I.62)

Let $\alpha(g) = b \ast \chi[\beta] \ast - \chi[\beta]$, derive

\[
\left\langle \alpha(g), \alpha_{\tau} \right\rangle = \frac{1}{4 \theta} \|\alpha(g)_{\tau}\|_2^2
\]

47
which, and on event (I.59) and Lemma H.6, we know

\[
\alpha^\tau = \beta^* \chi^l[\beta] \alpha - \chi^l[\beta],
\]

for some constant \( c_1 > 0 \). Now given any \( \alpha \) satisfies \( \alpha = \lambda^* C_{\alpha_0} \alpha \), the gradient of both objective can be derived as

\[
\text{grad}[\lambda^l](\alpha) = -P_{\lambda^l} \lambda^* C_{\alpha_0} \alpha = (aa^* - I) \lambda^* C_{\alpha_0} \chi^l[\beta]
\]

\[
= \lambda^* C_{\alpha_0} (\beta^* \chi^l[\beta] \alpha - \chi^l[\beta]),
\]

In the same spirit, define the coefficient of each gradient vector

\[
\zeta^l = \beta^* \chi^l[\beta] \alpha - \chi^l[\beta],
\]

As a consequence, we have that \( (I.64) \) is minimized at \( \beta^* \chi[\beta] = \frac{1}{2} n \theta \). Plugging in,

\[
(I.63) \quad \geq \frac{3}{2} n \theta \| \alpha^{\tau} \|_2^2 - \frac{1}{20} n \theta \gamma \| \alpha^{\tau} \|_2 - \frac{1}{1000} n \theta \gamma^2
\]

then again this is a quadratic function of \( \| \alpha^{\tau} \|_2 \) with positive leading coefficient and zeros at \( \{ 0, \frac{8}{60} \theta \} \), thus

\[
(I.64) \quad \geq \frac{3}{2} n \theta \| \alpha^{\tau} \|_2^2 - \frac{1}{20} n \theta \gamma \| \alpha^{\tau} \|_2 - \frac{1}{1000} n \theta \gamma^2 > 0
\]

which concludes our proof.

As a consequence, we have that

Corollary I.8 (Retraction of \( \varphi_p \) toward the subspace). Suppose that \( x_0 \sim \text{i.i.d.} \) \( \text{BG}(\theta) \) in \( \mathbb{R}^n \), and \( k, c_\mu \) such that \( (a_0, \theta, k) \) satisfies the sparsity-coherence condition \( \text{SCC}(c_\lambda) \). Define \( \lambda = c_\lambda / \sqrt{|k|} \) in \( \varphi_p \) with \( c_\lambda \in (0, \frac{1}{\tau}) \), then there exists some numerical constants \( C, c, c', c'' \), \( \gamma > 0 \) such that if \( \rho \) is \( \delta \)-smoothed \( \ell^1 \) function where \( \delta \leq c'' \lambda \theta^8 / \rho^2 \log n \) and \( n > C\rho^5 \theta^{-2} \log p \) and \( c_\mu \leq \tau \), then with probability at least \( 1 - c' / n \), for every \( \alpha \in \cup_{|\tau| \leq k^4} \mathcal{R}(S_\tau, \gamma(c_\mu)) \) such that

\[
d_\alpha(a, S_\tau) \geq (c_\mu) / 2
\]

then for every \( \alpha \) satisfying \( \alpha = \lambda^* C_{\alpha_0} \alpha \), there exists some \( \zeta \) satisfying \( \text{grad}[\varphi_p](\alpha) = \lambda^* C_{\alpha_0} \zeta \) that

\[
(\zeta^{\tau}, \alpha^{\tau}) \geq \frac{1}{60 \theta} \| \zeta^{\tau} \|_2^2.
\]

Proof. Write \( \gamma = (c_\mu) \). Define

\[
\chi^l[\beta] = \tilde{C}_{x_0} S_\lambda [\tilde{a} * y], \quad \chi_p[\beta] = \tilde{C}_{x_0} S_\lambda^\perp [\tilde{a} * y],
\]

which, and on event (I.59) and Lemma H.6, we know

\[
\beta^* \chi^l[\beta] \leq \frac{3}{4} n \theta,
\]

\[
\| \chi^l[\beta] \|_2 \leq \frac{1}{\gamma} n \theta \gamma,
\]

\[
\| \chi^l[\beta] - \chi_p[\beta] \|_2 \leq c_1 n \theta^4,
\]

for some constant \( c_1 > 0 \). Now given any \( \alpha \) satisfies \( \alpha = \lambda^* C_{\alpha_0} \alpha \), the gradient of both objective can be derived as

\[
\text{grad}[\varphi_p](\alpha) = -P_{\lambda^l} \lambda^* C_{\alpha_0} \alpha = (aa^* - I) \lambda^* C_{\alpha_0} \chi^l[\beta]
\]

\[
= \lambda^* C_{\alpha_0} (\beta^* \chi^l[\beta] \alpha - \chi^l[\beta]),
\]

\[
\text{grad}[\varphi_p](\alpha) = -P_{\lambda^l} \lambda^* C_{\alpha_0} \alpha = (aa^* - I) \lambda^* C_{\alpha_0} \chi_p[\beta]
\]

\[
= \lambda^* C_{\alpha_0} (\beta^* \chi_p[\beta] \alpha - \chi_p[\beta]),
\]

(1.73)
\[ \zeta_\rho = \beta^* \chi_\rho[\beta] \alpha - \chi_\rho[\beta], \]  
which, by norm inequality from (I.68)-(I.70) and Lemma I.7, we can derive
\[ \|\zeta_\ell - \zeta\|_2 \leq \|(I - \alpha \beta^\ast)(\chi_\rho[\beta] - \chi_\ell[\beta])\|_2 \leq c_1 n \theta^4, \]  
(1.75)
\[ \|((\zeta_\ell)^\ast, \alpha)\|_2 \geq \|\beta^* \chi_\ell[\beta]\|_2 - \|\chi_\ell[\beta]\|_2 \geq \frac{1}{\gamma} n \theta^3, \]  
(1.76)
\[ \langle (\zeta_\ell)^\ast, \alpha \rangle \geq \frac{1}{\gamma^2} \|((\zeta_\ell)^\ast, \alpha)\|_2^2, \]  
(1.77)
where the first inequality is derived by observing \((I - \alpha \beta^\ast)\) is a projection operator, such as:
\[ \beta^* \alpha = \alpha^\ast \alpha C_{\alpha^\ast \alpha} \alpha = \alpha^\ast \alpha = 1, \]
\[ (I - \alpha \beta^\ast)^2 = I - 2 \alpha \beta^\ast + \alpha (\beta^* \alpha) \beta^\ast = I - \alpha \beta^\ast. \]
Now we are ready to derive (1.57):
\[ \langle (\zeta_\ell)^\ast, \alpha \rangle \geq \langle (\zeta_\ell)^\ast, \alpha \rangle - \|\alpha \|_2 \|\zeta_\ell - \zeta\|_2 \]
\[ \geq \frac{1}{4\theta^2} \|((\zeta_\ell)^\ast, \alpha)\|_2^2 - c_1 n \theta^4 \gamma \]
\[ \geq \frac{1}{12 \theta^4} \|((\zeta_\ell)^\ast, \alpha)\|_2^2 \]
\[ + \frac{1}{4 \theta^4} \|((\zeta_\ell)^\ast, \alpha)\|_2^2 - \frac{2}{4 \theta^4} \|\zeta_\ell - \zeta\|_2 \]
\[ = \frac{1}{6 \theta^4} \|((\zeta_\ell)^\ast, \alpha)\|_2^2 + \frac{1}{12 \theta^4} \|\zeta_\ell - \zeta\|_2 \]
\[ \geq \frac{1}{6 \theta^4} \|((\zeta_\ell)^\ast, \alpha)\|_2^2 - c_1 n \theta^4 \gamma \]
\[ \geq \frac{1}{6 \theta^4} \|((\zeta_\ell)^\ast, \alpha)\|_2^2 - c_1 n \theta^4 \gamma \]
\[ \geq \frac{1}{6 \theta^4} \|((\zeta_\ell)^\ast, \alpha)\|_2^2, \]  
(1.78)
where the last inequality is true since \( \theta^3 \ll \gamma. \)

I.5 Proof of Theorem B.1

By collecting result from above, we are ready to prove the acclaimed geometric result in Theorem B.1. It guarantees that for every \( a \) near \( S_\tau \), either one of the following in true
\[ \lambda_{\min} (\text{Hess} [\varphi_\rho](a)) \leq -c_1 n \theta \lambda, \]  
(1.79)
\[ \langle \sigma(0)^\ast s(0)[a_0], -\text{grad} [\varphi_\rho](a) \rangle \geq c_2 n \theta (\log^{-2} \theta^{-1}) \lambda^2, \]  
(1.80)
\[ \text{Hess} [\varphi_\rho](a) \succ c_3 n \theta \cdot P_{a^\perp}, \]  
(1.81)
all local minimizer \( a \) satisfies for some \( a_\ast \in \{ \pm \ell^* s(\ell) \mid \ell \in [\pm p_0] \}, \)
\[ \| a - a_\ast \|_2 \leq c_4 \sqrt{c_\mu} \max \{ \mu, p_0^{-1} \}, \]  
(1.82)
and whenever \( \frac{1}{2} \leq d_\ast (a, S_\tau) \leq \gamma \), coefficient of \( a \) and its gradient \( g, \alpha \), written as \( \zeta \), satisfies
\[ \langle (\alpha)^\ast, \alpha \rangle \geq \frac{c_1}{\theta^3} \| \alpha \|_2^2 \]  
(1.83)
To connect the geometric results introduced in Lemma I.1, Lemma I.3, Lemma I.5 and Lemma I.7, we are only required to prove the required signal condition claimed in Theorem B.1 is necessary from Definition E.1. In particular, when the subspace dimension \( |\tau| \leq 4p_0 \theta \). On top of that, we are also required to show the chosen smooth parameter \( \delta \) in the pseudo-Huber penalty \( p(x) = \sqrt{x^2 + \delta^2} \approx \|x\| \) sufficiently well, hence results of Corollary I.2, Corollary I.4, Corollary I.6 and Corollary I.8 also holds.
Proof. Firstly we will show when largest solution subspace dimension $k = 4p_0\theta$, the signal condition of Definition E.1 will be satisfied. Recall that the signal condition of Theorem B.1 requests

$$\frac{2}{p_0 \log^2 p_0} \leq \theta \leq \frac{c}{(p_0 \sqrt{\mu} + \sqrt{p_0}) \log^2 p_0},$$  \hspace{1cm} (I.84)

since $p = 3p_0 - 2$, this implies the lower bounds for sparsity $\theta$ as:

$$\theta \geq \frac{1}{2p_0 (\frac{1}{2} \log p_0)^2} \geq \frac{1}{p \log^2 \theta - 1};$$  \hspace{1cm} (I.85)

the upper bound of $\theta$ via $\theta \sqrt{p_0} \log^2 p_0 \leq c$:

$$\theta \leq \frac{9c}{\sqrt{p_0} (3 \log p_0)^2} \leq \frac{16c}{\sqrt{p} \log^2 \theta - 1}, \quad \theta \leq \frac{4c^2}{k \log^4 p_0} \leq \frac{36c^2}{k (3 \log p_0)^2} \leq \frac{36c^2}{k \log^2 \theta - 1};$$  \hspace{1cm} (I.86)

and the upper bound for coherence $\mu$ as:

$$\mu \max \left\{ k^2, (p\theta)^2 \right\} \log^2 \theta - 1 \leq \mu \max \left\{ 16(p_0\theta)^2, 9(p_0\theta)^2 \right\} \log^2 \theta - 1 \leq 16 \left( \sqrt{\mu}p_0\theta \right)^2 \log^2 p_0 \leq 16c.$$  \hspace{1cm} (I.87)

Therefore Definition E.1 holds if $\max \left\{ 16c, 36c^2 \right\} \leq c\mu/4$ via (I.85)-(I.87).

Furthermore, we know from lemma assumption all interested $a$ are near subspace $S_r$ by

$$d_a(a, S_r) \leq \frac{c}{\sqrt{p_0} \log^2 \theta - 1} \cdot \min \left\{ \frac{1}{\sqrt{\theta}}, \frac{1}{\sqrt{\mu}}, \frac{1}{(p_0\theta)^{3/2}} \right\} \leq \frac{c}{\log^2 \theta - 1} \min \left\{ \frac{2}{\sqrt{k}}, \frac{1}{\sqrt{p_0\mu}}, \frac{4}{kp_0\sqrt{\mu}k} \right\} \leq \gamma$$  \hspace{1cm} (I.88)

where $\gamma$ is defined in Definition E.3 of widened subspace $\mathcal{R}(S_r, \gamma(c_{\mu}))$.

Lastly, the pseudo-Huber function $\rho(x) = \sqrt{x^2 + \delta^2}$ is an $\ell^1$ smoothed sparse surrogate defined in Definition H.2, by observing that it is convex, smooth, even, whose second order derivative (according to Table 1) $\nabla^2 \rho(x) = \frac{\delta^2}{(x^2 + \delta^2)^{3/2}}$ is monotone decreasing in $|x|$. More importantly

$$\sup_{x \in \mathbb{R}} |\rho(x) - |x|| = |\rho(0) - |0|| = \delta.$$  \hspace{1cm} (I.89)

Hence, by choosing $\delta \leq \frac{c' \rho^3}{p^3 \log^2 p}$, for some sufficiently small constant $c'$ and letting $\lambda = 0.2\sqrt{k} = 0.1/\sqrt{p_0}\theta$ in $\varphi_{p\theta}$. We obtain the geometrical results in Corollary I.2 when $|\beta_{(1)}| \geq \frac{1}{5} |\beta_{(0)}|$, Corollary I.4 when $\frac{4}{5} |\beta_{(0)}| \geq |\beta_{(1)}| \geq \frac{\lambda}{4 \log^2 \theta - 1}$ and Corollary I.6 when $\frac{\lambda}{4 \log^2 \theta - 1} \geq |\beta_{(1)}|$, and the retraction result in Corollary I.8.

**J Analysis of algorithm — minimization within widened subspace**

In this section, we prove convergence of the first part of our algorithm—minimization of $\varphi_{p\theta}$ near $S_r$. We begin by proving the initialization method guarantees that $a^{(0)}$ is near $S_r$, in the sense that

$$d_a(a^{(0)}, S_r) \leq \gamma,$$  \hspace{1cm} (J.1)

where the distance $d_a$ is defined in (B.15). We then demonstrate that small-stepping curvilinear search converges to a desired local minimum of $\varphi_{p\theta}$ 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 $S_r$ and (ii) the geometric properties of $\varphi_{p\theta}$ near $S_r$.  

50
J.1 Initialization near subspace

The following lemma shows that the initialization $a^{(0)} = P_{SP^{-1}} [\nabla \varphi_{\ell i}(a^{(-1)})]$, where

$$a^{(-1)} = P_{SP^{-1}} \left[ \sum_{\ell \in \tau} x_{\ell}^0 t_{p_0}^* s_r[a_0] \right], \quad (J.2)$$

and is very close to the subspace $S_{\tau}$:

**Lemma J.1** (Initialization from a piece of data). Let $x \in \mathbb{R}^{2p_0 - 1}$ indexed by $[\pm p_0]$, with $x_1 \sim i.i.d. \mathcal{B}(\theta)$. Define $\tilde{y} = x * a_0$, and $a^{(0)}$ as

$$a^{(0)} = -P_{SP^{-1}} \nabla \varphi_{\ell i} \left( P_{SP^{-1}} \left[ 0^{p_0 - 1}; [y_0^1; \cdots; y_{p_0 - 1}^1]; 0^{p_0 - 1} \right] \right), \quad (J.3)$$

with $\lambda = 0.2 / \sqrt{p \theta}$ in $\varphi$. Set $\tau = \text{supp}(x)$. Suppose that $(a_0, \theta, k)$ satisfies the sparsity-coherence condition SCC($c_\mu$) and $a_0$ satisfies $a_0 \leq \sqrt{\min(p_0 \theta)}$, as in (J.4). Then there exists some constant $c, \tau > 0$ such that if $p_0 \theta > 100c$ and $c_\mu \leq \tau$, then with probability at least $1 - 1/c$, we have

$$d_{\alpha} \left( a^{(0)}, S_{\tau} \right) \leq \frac{c_\mu}{4 \log^2 \theta^{-1}} \min \left\{ \frac{1}{\sqrt{|\tau|}}, \frac{1}{\sqrt{\mu p}}, \frac{1}{\mu p \sqrt{\theta}} \right\} \cdot \tau \quad (J.4)$$

**Proof.**

1. **(Distance to $S_{\tau}$ from $a^{(0)}$)** Let $\eta = \|e^*_{p_0} (a_0 * x)\|_2^2 = \|e^*_{p_0} a_{0} \chi\|_2$, and $\gamma = \gamma(c_\mu)$, as in (J.4). Expand the expression of $a^{(0)}$ from (J.3) we have

$$a^{(0)} = P_{SP^{-1}} t^* \tilde{C}_y \tilde{S}_\lambda \left[ \tilde{C}_y t_{p_0} P_{SP^{-1}} t^*_{p_0} (a_0 * x) \right] = P_{SP^{-1}} t^* C_{a_0} \chi \left[ \frac{\eta}{\gamma} C_{a_0} t_{p_0} t^*_{p_0} C_{a_0} x \right] \quad (J.5)$$

To relate $a^{(0)}$ to its coefficient, introduce the truncated autocorrelation matrix $\tilde{M} = C_{a_0} t_{p_0} t^*_{p_0} C_{a_0}$, define $\tilde{\alpha}, \tilde{\beta}$ as

$$\tilde{\beta} = \frac{1}{\eta} \tilde{M} x, \quad \tilde{\alpha} = \chi \left[ \frac{1}{\eta} \tilde{M} x \right] = \chi(\tilde{\beta}) \quad (J.6)$$

and note that $\tilde{M}$ is bounded entrywise as

$$|\tilde{M}_{ij}| \leq \begin{cases} 1 & i = j \in [-p_0 + 1, p_0 - 1] \\ \mu & i \neq j \in [-p_0 + 1, p_0 - 1], |i - j| < p_0 \\ 0 & \text{otherwise} \end{cases} \quad (J.7)$$

From (J.5), we can write $a^{(0)} = P_{SP^{-1}} t^* C_{a_0} \tilde{\alpha}$, meaning that the normalized version of $\tilde{\alpha}$ is a valid coefficient vector for $a^{(0)}$. Let $\tau^c = [\pm 2p_0] \setminus \tau$. The distance $d_{\alpha}$ to subspace $S_{\tau}$ (B.15) is upper bounded as

$$d_{\alpha} \left( a^{(0)}, S_{\tau} \right) \leq \frac{\|\hat{\alpha}_{\tau^c}\|_2}{\|t^* C_{a_0} \hat{\alpha}\|_2} \leq \frac{\|\hat{\alpha}_{\tau^c}\|_2}{\|t^* C_{a_0} \hat{\alpha}_{\tau^c}\|_2} - \frac{\|t^* C_{a_0} \hat{\alpha}_{\tau^c}\|_2}{\|\hat{\alpha}_{\tau^c}\|_2}$$

where the last inequality is derived with Lemma E.4. Therefore, it is sufficient to show

$$\left( 1 + \gamma \sqrt{1 + \mu p} \right) \|\hat{\alpha}_{\tau^c}\|_2 \leq \gamma \sqrt{1 - \mu |\tau|} \|\hat{\alpha}_{\tau^c}\|_2 \quad (J.8)$$

to complete the proof that $d_{\alpha} \left( a^{(0)}, S_{\tau} \right) \leq \gamma$. 

51
2. (Bound $\eta$) Condition on the following two events

$$\mathcal{E}_\tau := \{ \tau < 4p_0 \theta \}, \quad \mathcal{E}_{||x||_2} := \left\{ \sqrt{p_0 \theta} \leq ||x||_2 \leq \sqrt{3p_0 \theta} \right\} \tag{J.9}$$

and utilize $\mu$ bound from Lemma E.5 such that $\mu |\tau| < 0.1$. An upper bound on $\eta$ can be obtained using properties of $\tilde{M}$ of (J.7):

$$\eta = \left\{ \|t_{p_0}^* C_{a_0} x\|_2 \leq \|t^* C_{a_0} x\|_2 \leq \sqrt{1+\mu |\tau|} \|x\|_2 \leq 2\sqrt{p_0 \theta} \right\} \tag{J.10}$$

To lower bound $\eta$, use $\eta^2 = g^* P_{\tau} \bar{M} P_{\tau} g$ where $g$ is the standard Gaussian vector. Observe the submatrix of $\bar{M}$ is diagonal:

$$\left\{ \begin{array}{ll}
\bar{M}_{ii} = \|t_{p_0}^* s_i[a_0]\|_2^2 & \in [0, 1] \\
\operatorname{tr} (\bar{M}) = \sum_{i \in \{\pi \leq p_0\}} \|t_{p_0}^* s_i[a_0]\|_2^2 = \|a_0\|_2^2 + \sum_{i=1}^{p_0-1} \left( \|t_{p_0}^* s_i[a_0]\|_2^2 + \|t_{p_0}^* s_{i-p_0}[a_0]\|_2^2 \right) = p_0 . 
\end{array} \right. \tag{J.11}$$

Write $x = g \circ w$ where $w$ and $g$ are Bernoulli and Gaussian vector respectively with supp$(w) = \tau$, then the trace of $P_{\tau} \bar{M} P_{\tau}$ can be written as sum of independent r.v.s as:

$$\operatorname{tr} \left( P_{\tau} \bar{M} P_{\tau} \right) = \sum_{i \in \{\pi \leq p_0\}} w_i \|t_{p_0}^* s_i[a_0]\|_2^2 .$$

Bernstein inequality Lemma N.4 and (J.11) gives

$$\mathbb{P} \left[ \operatorname{tr} \left( P_{\tau} \bar{M} P_{\tau} \right) < \frac{3p_0 \theta}{4} \right] \leq \mathbb{P} \left[ \operatorname{tr} \left( P_{\tau} \bar{M} P_{\tau} \right) - p_0 \theta \leq -\frac{p_0 \theta}{4} \right]$$

$$\leq 2 \exp \left( \frac{-\left(\frac{p_0 \theta}{4}\right)^2}{2p_0 \theta + p_0 \theta / 2} \right) \leq 2 \exp \left( \frac{-p_0 \theta}{40} \right), \tag{J.12}$$

thus condition on $\omega$ satisfies $\operatorname{tr} \left( P_{\tau} \bar{M} P_{\tau} \right) \geq \frac{3p_0 \theta}{4}$ and $\mathcal{E}_\tau$, expectation $\eta^2$ has lower bound

$$\mathbb{E}_{g|w} \eta^2 = \mathbb{E}_{g|w} \left[ g^* P_{\tau} \bar{M} P_{\tau} g \right] = \operatorname{tr} \left( P_{\tau} \bar{M} P_{\tau} \right) \geq \frac{3p_0 \theta}{4}$$

then apply Bernstein inequality again by first writing svd of $P_{\tau} \bar{M} P_{\tau} = U \Sigma U^*$ with $\Sigma$ being rank $|\tau| < 4p_0 \theta$ and square orthobasis $U$. Let $g' = U^* g$, then $g'$ is standard i.i.d. Gaussian vector, provides alternative expression $\eta^2 < \sum_{i=1}^{4p_0 \theta} \sigma_i^2$ where $\sigma_i \leq 1 + \mu |\tau| \leq 1.1$. We obtain probability of $\eta^2$ to be small as

$$\mathbb{P}_{g|w} \left[ \eta^2 < \frac{p_0 \theta}{2} \right] \leq \mathbb{P}_{g|w} \left[ \eta^2 - \mathbb{E}_{g|w} \eta^2 < -\frac{p_0 \theta}{4} \right]$$

$$\leq 2 \exp \left( \frac{-\left(\frac{p_0 \theta}{4}\right)^2}{2(1 + \mu |\tau|)(12p_0 \theta + p_0 \theta / 2)} \right) \leq 2 \exp \left( \frac{-p_0 \theta}{440} \right) \tag{J.13}$$

by applying moment bounds $(\sigma^2, R) = (12p_0 \theta (1 + \mu |\tau|), 2(1 + \mu |\tau|))$. We thereby define event

$$\mathcal{E}_\eta = \left\{ \sqrt{p_0 / 2} \leq \eta \leq 2\sqrt{p_0 \theta} \right\} , \tag{J.14}$$

which holds w.h.p. based on (J.9), (J.12) and (J.13).
3. (Bound $\tilde{\alpha}$) Condition on $\mathcal{E}_\eta \cap \mathcal{E}_{||x||_2} \cap \mathcal{E}_\tau$. Use definition $\tilde{\beta} = \frac{1}{\eta} \tilde{M} x$ from (J.6), and properties of $\tilde{M}$ from (J.7) we can obtain:

$$
\begin{align*}
\|\tilde{\beta}_r\|_2 &\leq \frac{1}{\eta} \|\epsilon_r \tilde{M}_r\|_2 \|x\|_2 \leq \frac{\mu \sqrt{p_0 |\tau|}}{\sqrt{p_0\theta} / 2} \cdot \sqrt{3p_0 \theta} \leq 3\mu \sqrt{p_0 |\tau|}.
\end{align*}
$$

(J.15)

Use definition $\|\tilde{\alpha}\|_2 = \|\chi|\tilde{\beta}|\|_2$, condition on event

$$
\mathcal{E}_\chi := \left\{ \sigma, \chi|\tilde{\beta}|_i \geq n\theta \sigma_p \lambda \|\tilde{\beta}_i\| - \frac{\epsilon_n \theta}{p}, \quad \forall i \in \tau \right\},
$$

also from Definition E.1 we have $\mu (p\theta)^{1/2} |\tau|^{3/2} < \frac{c_\mu}{\log \theta}$ and from lemma assumption $\lambda = \frac{1}{\sqrt{\log \theta}}$, provides bounds of $\|\tilde{\alpha}\|_2$ via triangle inequality as:

$$
\begin{align*}
\|\tilde{\alpha}_r\|_2 &\leq 4n\theta^2 |\tau| \cdot \|\tilde{\beta}_r\|_2 + \frac{c_n \theta}{p} \cdot \sqrt{2p_0} \leq 3c_\mu n\theta \left( \frac{\sqrt{\theta}}{\log \theta} + \frac{c_\mu}{p} \right)
\|
\tilde{\alpha}_r\|_2 &\geq n\theta \left( \|\tilde{\beta}_r\|_2 - \nu_2 \lambda \sqrt{|\tau|} - \frac{\epsilon_n \theta}{p} \sqrt{|\tau|} \right) \geq n\theta \left( 0.45 - \sqrt{\frac{\epsilon_n \theta}{p}} - \frac{1}{c} - c_\mu \right) \geq 0.2n\theta,
\end{align*}
$$

(J.16)

since both $\theta |\tau|, \mu p \theta |\tau| < c_\mu$, we have

$$
\begin{align*}
\sqrt{1 + \mu p} \|\tilde{\alpha}_r\|_2 &\leq 3c_\mu n\theta \sqrt{1 + \mu p} \left( \sqrt{\theta} + p^{-1} \right) \leq 6c_\mu n\theta
\|
\|\tilde{\alpha}_r\|_2 &\leq \frac{6c_\mu n\theta \cdot \min \left\{ \frac{1}{\sqrt{|\tau|}}, \frac{1}{\sqrt{p}}, \frac{1}{\mu p \sqrt{\theta} |\tau|} \right\} \leq 24\sqrt{c_\mu} n\theta \gamma,
\end{align*}
$$

which satisfies (J.8), since $\mu |\tau| < c_\mu < \frac{1}{1000}$,

$$
(1 + \gamma \sqrt{1 + \mu p}) \|\tilde{\alpha}_r\|_2 \leq \left( 24\sqrt{c_\mu} + 6c_\mu \right) n\theta \gamma \leq 0.1n\theta \gamma
\|
\leq \gamma \sqrt{1 - \mu |\tau|} \|\tilde{\alpha}_r\|_2.
$$

(J.17)

Finally, given $p_0 \theta > 1000c$, this result holds with probability at least

$$
1 - \frac{p \mathbb{P} \left[ \mathcal{E}_\eta \right]}{\mathbb{E} \left[ \mathcal{E}_{||x||_2} \right]} - \frac{p \mathbb{P} \left[ \mathcal{E}_\eta \right]}{\mathbb{E} \left[ \mathcal{E}_{\eta} \right]} - \frac{p \mathbb{P} \left[ \mathcal{E}_\chi \right]}{\mathbb{E} \left[ \mathcal{E}_\chi \right]} \geq 1 - 2 \frac{2}{p_0 \theta} - \frac{1 - 4 \exp \left( \frac{-p_0 \theta}{440} \right) \geq 1 - \frac{1}{c}
$$

(J.18)

### J.2 Minimization near subspace (Proof of Theorem C.1)

Before we start the proof of theorem, writing $g = \text{grad}[\varphi_p](a)$ and $H = \text{Hess}[\varphi_p](a)$, we will first restate the results of Theorem B.1 in simplified terms. The theorem shows that for any $a \in S^{p-1}$ whose distance to subspace $d_o(a, S_\tau) \leq \gamma$, then at least one of the the following statement hold:

$$
\|g\|_2 \geq \eta_g \tag{J.19}
$$

$$
\lambda_{\min}(H) \leq -\eta_v \tag{J.20}
$$
Furthermore, \( \varphi_\rho \) is retractive near \( S_\tau \): wherever \( d_\alpha(a, S_\tau) \geq \frac{\rho}{2} \), writing \( \alpha(a), \alpha(g) \) to be the coefficient of \( a, g \), we have

\[
(\alpha(a), \alpha(g)) \geq \eta_r \|\alpha(g)\|_2.
\]  

Also, the the gradient, Hessian and the third order derivative are all bounded as follows:

**Remark J.2.** With high probability, for every \( a \) whose \( d_\alpha(a, S_\tau) < \gamma \), its max \( \{\|g\|_2, \|H\|_2, \|\nabla H\|_2\} \leq \bar{\eta} = \text{poly}(n, p) \).

We state **Remark J.2** without explicit proof since its derivation is similar to the proof in Theorem B.1.

We prove that if the negative curvature direction \(-v\) is chosen to be the least eigenvector with \( v^*Hv < -\eta_v \) and \( v^*g \) (if not, let \( v = 0 \)), then the iterates

\[
a^{(k+1)} = P_{S_\rho-1} \left[ a^{(k)} - t g^{(k)} - t^2 v^{(k)} \right]
\]  

converges toward the minimizer \( \bar{a} \) 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 \( a_0 \) being \( \mu \)-truncated shift coherent implies that \( a_0 \) is at most \( 2\mu \)-shift coherent. Hence we utilize the usual incoherence condition in the proof.

**Proof.** Notice that when \( a \) is in the region near some signed shift \( \bar{a} \) of \( a_0 \), 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 \( k \), \( a^{(k)} \) is in this region, then

\[
\|a^{(k)} - \bar{a}\|_2 \leq (1 - t\eta_c)^{-(k-k)} \|a^{(k)} - \bar{a}\|_2
\]  

AMS9 (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 \( a^{(k)} \) remain close to the subspace \( S_\tau \) and that after \( \bar{k} = \text{poly}(n, p) \) iterations they indeed remain in the strongly convex region around some \( \bar{a} \).

1. (Existence of Armijo steplength). First, we show there exists a nontrivial step size \( t \) at every iteration, in the sense that for all \( a \in S_{\rho-1} \), there exists \( T > 0 \) such that for all \( t \in (0, T) \), the Armijo step condition (C.11) is satisfied. Note that since \( \varphi_\rho \) is a smooth function, \( a \rightarrow \varphi_\rho \circ P_{S_{\rho-1}}(a) \) admits a version of Taylor’s theorem (see also AMS9 (Section 7.1.3)): for any \( \xi \perp a \), writing \( a^+ = P_{S_{\rho-1}}(a + \xi) \),

\[
|\varphi_\rho(a^+) - (\varphi_\rho(a) + (\text{grad}\varphi_\rho)(a), \xi) + \frac{1}{2} \xi^*\text{Hess}\varphi_\rho(a)\xi| \leq \bar{\eta} \|\xi\|_2^3,
\]  

using \( \|\nabla H\|_2 \leq \bar{\eta} \). Now, let \( \xi = -tg - t^2v \) as in the iterates (C.10). Suppose the Armijo step condition (C.11) does not hold, so

\[
\varphi_\rho(a^+) > \varphi_\rho(a) - \frac{1}{2} \left( t \|g\|^2_2 + \frac{1}{2} t^4 \eta_c \|v\|_2^3 \right).
\]  

Since \( g^*v \geq 0 \) and \( v^*Hv \leq -\eta_v \|v\|^2_2 \) or \( v = 0 \), using \( \|a + b\|^3_2 \leq 4\|a\|^3_2 + 4\|b\|^3_2 \) (Hölder’s inequality) and \( \|H\|_2 < \bar{\eta} \), we can derive

\[
\langle g, -tg - t^2v \rangle + \frac{1}{2}(tg + t^2v)^*H(tg + t^2v)
\]  

\[
+ e \left( t \|g\|^2_2 + \frac{1}{2} t^4 \eta_c \|v\|^3_2 \right)
\]  

\[
\Rightarrow - \frac{1}{2} t \|g\|^2_2 + \frac{1}{2} t^2 g^*Hg + t^3 v^*Hg
\]  

\[
- \frac{1}{2} t \eta_c \|v\|^2_2 + 4\bar{\eta}t^3 \|g\|^3_2 + 4\bar{\eta}t^6 \|v\|^3_2 > 0
\]  

\[
\Rightarrow - \frac{1}{2} t \|g\|^2_2 + t^2 \left( \frac{1}{2} \eta \|g\|^2_2 + t\bar{\eta} \|v\|_2 \|g\|_2 + 4\bar{\eta}t^3 \|g\|^3_2 \right)
\]  

\[
- \frac{1}{2} t \eta_c \|v\|^2_2 + 4\bar{\eta}t^6 \|v\|^3_2 > 0.
\]  

(J.26)
If
\[ t < T = \min \left\{ \frac{\|g\|_2}{\bar{\eta}} + \frac{2\bar{\eta}t}{\|v\|_2} + \frac{8\bar{\eta}t\|g\|_2}{\sqrt{16\|v\|_2}}, \right\} \] (J.27)
then (J.26) < 0 contradicting (J.25). Using our bounds on \( \|g\|_2, \bar{\eta}, \eta_v \) and \( \|v\|_2 \), it follows that \( T \) is lower bounded by a polynomial poly \((n^{-1}, p^{-1})\).

2. (Bounds on \( d_\alpha(g, S_\tau), d_\alpha(v, S_\tau) \)) We will show there are numerical constants \( c_g, c_v \) such that
\[ d_\alpha(g, S_\tau) \leq c_g n\theta \gamma \quad \text{and} \quad d_\alpha(v, S_\tau) \leq c_v n\theta \rho. \] (J.28)
Define
\[ \chi_{\ell^t}[\beta] = \overline{C}_{x_0} \text{prox}_{\mu^t} \overline{\alpha} \ast y, \quad \chi_\rho[\beta] = \overline{C}_{x_0} \text{prox}_\rho \overline{\alpha} \ast y, \]
then the gradient can be written as (I.58)
\[ \begin{align*}
\text{grad}[\varphi_{\ell^t}](a) &= \ell^t C a_0 (\beta^* \chi_{\ell^t}[\beta] \alpha - \chi_{\ell^t}[\beta]), \\
\text{grad}[\varphi_\rho](a) &= \ell^t C a_0 (\beta^* \chi_\rho[\beta] \alpha - \chi_\rho[\beta]).
\end{align*} \] (J.29)
Use the following inequalities:
\[ \begin{align*}
\frac{1}{2} n \theta &\leq \|\beta^* \chi_{\ell^t}[\beta]\| \leq \frac{3}{2} n \theta, \\
\|\chi_{\ell^t}[\beta]\|_2 &\leq \frac{1}{50} n \theta \gamma, \\
\|I - \alpha \beta^*\|_2 &\leq 4 \sqrt{\rho}, \\
\|\chi_{\ell^t}[\beta] - \chi_\rho[\beta]\|_2 &\leq n \theta^2,
\end{align*} \]
where the first and second bounds of \( \chi_{\ell^t}[\beta] \) based on event (I.59); the third by observing \( \|\alpha\|_2 \leq 2 \) and \( \|\beta\|_2 \leq 2 + c_\mu \sqrt{\rho} \); the last from (H.21) of Lemma H.6 when \( \delta \) is sufficiently small. Hence, by definition of \( d_\alpha(\cdot, S_\tau) \) (B.15) and knowing \( \alpha \) is close to subspace \( \|\alpha - \tau\|_2 \leq \gamma \), via triangle inequality, we get
\[ \begin{align*}
d_\alpha(g, S_\tau) &\leq d_\alpha(\text{grad}[\varphi_{\ell^t}](a), S_\tau) + d_\alpha(\text{grad}[\varphi_\rho](a) - \text{grad}[\varphi_{\ell^t}](a), S_\tau) \\
&\leq \|\beta^* \chi_{\ell^t}[\beta] \alpha - \chi_{\ell^t}[\beta]\|_2 + \|\chi_{\ell^t}[\beta] - \chi_\rho[\beta]\|_2 \\
&\leq \frac{3}{2} n \theta \gamma + \frac{1}{50} n \theta \gamma + 4 \sqrt{\rho} n \theta^4 \\
&\leq 3 n \theta \gamma.
\end{align*} \] (J.31)
To bound the \( d_\alpha \) norm of least eigenvector \( v \), note that \( \beta^* \chi_\rho[\beta] > 0 \), we can conclude
\[ v^* \nabla^2 \varphi_\rho(a) = v^* P_{a^\perp} \nabla^2 \varphi_\rho(a) P_{a^\perp} v + \beta^* \chi_\rho[\beta] = v^* H v < -\eta_v, \]
expand \( \nabla^2 \varphi_\rho(a) \) as in (H.8), and since \( v \) is the eigenvector of smallest eigenvalue \( \lambda_{\min} < -\eta_v \),
\[ P_{a^\perp} \nabla^2 \varphi_\rho(a) P_{a^\perp} = (I - aa^*) \ell^t C a_0 \overline{C}_{x_0} \nabla \text{prox}_\rho \overline{\alpha} \ast y \overline{C}_{x_0} C_{a_0}^* v = \lambda_{\min} v, \] (J.32)
and hence there exists \( \alpha(v) \) satisfies \( v = \ell^t C a_0 \alpha(v) \) and
\[ \alpha(v) = \lambda_{\min}^{-1} \left[ \overline{C}_{x_0} \nabla \text{prox}_\rho \overline{\alpha} \ast y \overline{C}_{x_0} C_{a_0}^* v - \left( \beta^* \overline{C}_{x_0} \nabla \text{prox}_\rho \overline{\alpha} \ast y \overline{C}_{x_0} C_{a_0}^* v \right) \alpha \right]. \]
Now since \( \nabla \text{prox}_\rho \overline{\alpha} \ast y \) is a diagonal matrix with entries in \([0, 1]\),
\[ d_\alpha(v, S_\tau) \leq \|\alpha(v)\|_2 \leq \|\alpha\|_2 \|\alpha\|_2 (1 + \|\alpha\|_2^2) < c_v n \theta \rho, \] (J.33)
where we use upper bound of $\|x_0\|_1 < cn\theta$ from Lemma D.2 and $|\lambda_{\min}| > \eta_\nu > cn\theta\lambda$ from Corollary I.2.

3. (Iterates stay within widened subspace). Suppose (J.22) holds. We will show that whenever

$$t \leq T' = \frac{1}{10n\theta},$$

(J.34)

then setting $a^+ = P_{S_{\gamma}^{-1}} \left[ a - tg - t^2v \right]$, we have

$$|d_{\alpha}(a^+, S_{\tau}) - d_{\alpha}(a, S_{\tau})| \leq \frac{\gamma}{2},$$

(J.35)

and whenever $d_{\alpha}(a, S_{\tau}) \in \left[ \frac{\gamma}{2}, \gamma \right]$

$$d_{\alpha}^2(a^+, S_{\tau}) \leq d_{\alpha}^2(a, S_{\tau}) - t \cdot c'n\theta\gamma^2.$$  

(J.36)

If both (J.35) and (J.36) hold, then all iterates $a^{(k)}$ will stay near the subspace: $d_{\alpha}(a^{(k)}, S_{\tau}) < \gamma$.

To derive (J.35), since both $g \perp a$ and $v \perp a$ we have $\|a - tg - t^2v\|_2^2 = \|a\|_2^2 + \|tg + t^2v\|_2^2 > 1$, and since $d_{\alpha}(\cdot, S_{\tau})$ is a seminorm Lemma E.2:

$$d_{\alpha}(a^+, S_{\tau}) = d_{\alpha}(P_{S_{\gamma}^{-1}} \left[ a - tg - t^2v \right], S_{\tau}) \leq d_{\alpha}(a - tg - t^2v, S_{\tau})$$

$$\leq d_{\alpha}(a, S_{\tau}) + t d_{\alpha}(g, S_{\tau}) + t^2 d_{\alpha}(v, S_{\tau})$$

(J.37)

suggests (J.35) holds via (J.28) and let $n > C\rho^2\theta^{-2}$, we have

$$t d_{\alpha}(g, S_{\tau}) + t^2 d_{\alpha}(v, S_{\tau}) \leq \frac{c_n\theta\gamma}{10n\theta} + \frac{c_n\theta\rho}{10n\theta^2} < \frac{\gamma}{2}$$

(J.38)

with sufficiently large $C$.

To derive (J.36), let $\alpha(a)$ to be a coefficient vector satisfying $d_{\alpha}(a, S_{\tau}) = \|\alpha(a)\|_2$, and based on (J.30) and (J.33), define

$$\alpha(g) = \beta^* \chi_{\rho}\beta \alpha(a) - \chi_{\rho}\beta$$

(J.39)

$$\alpha(v) = \lambda_{\min}^{-1} C_{x_0} \nabla \text{prox}_{\lambda_{\rho}} [a \ast y] C_{x_0} C_{\alpha}^* \ell v.$$  

(J.40)

By the retraction property and norm bounds,

$$\langle \alpha(a)_{\tau^*}, \alpha(g)_{\tau^*} \rangle \geq \frac{1}{10n\theta} \|\alpha(g)_{\tau^*}\|_2$$

(J.41)

$$\|\alpha(a)_{\tau^*}\|_2 \leq \gamma$$

(J.42)

$$\|\alpha(v)\|_2 \leq c_n n\theta\rho.$$  

(J.43)

Since $\|\alpha_{\tau^*}\|_2 > \frac{\gamma}{2}$,

$$\|a(g)_{\tau^*}\|_2 \geq \|\beta^* \chi_{\ell}\beta \alpha_{\tau^*} - \chi_{\rho}\beta\|_{\tau^*} - \|I - \alpha_{\tau^*}\|_{\tau^*} (\chi_{\rho}\beta - \chi_{\ell}\beta)\|_{\tau^*}$$

$$\geq \|\beta^* \chi_{\ell}\beta\|_{\tau^*} - \|\chi_{\rho}\beta\|_{\tau^*} - \|I - \alpha_{\tau^*}\|_{\tau^*} \|\chi_{\rho}\beta - \chi_{\ell}\beta\|_{\tau^*}$$

$$\geq \frac{1}{2} n\theta \times \frac{\gamma}{2} - \frac{1}{25} n\theta\gamma + 2n\theta^4$$

(J.44)

Finally, we can bound $d_{\alpha}(a^+, S_{\tau})$ as

$$d_{\alpha}^2(a^+, S_{\tau}) \leq d_{\alpha}^2(a - tg - t^2v, S_{\tau})$$

$$\leq \|\alpha(a) - t\alpha(g) - t^2\alpha(v)\|_{\tau^*}^2$$

$$= \|\alpha(a)_{\tau^*}\|_2^2 - 2t \langle \alpha(a)_{\tau^*}, [\alpha(g) + t\alpha(v)]_{\tau^*}\rangle + t^2 \|\alpha(g) + t\alpha(v)\|_{\tau^*}^2$$
\[
\leq \|\alpha(a)\|^2 - 2t (\alpha(a)\tau_\rho + \alpha(g))_\tau + 2t^2 \|\alpha(a)\|^2 + 2t^2 \|\alpha(g)\|^2 + 2t^2 \|\alpha(v)\|^2
\]

where the last inequality holds when \(t < \frac{0.1}{n\theta} \) with sufficiently large \(n\).

4. (Polynomial time convergence) The iterates \(a^{(k)}\) remain within a \(\gamma\) neighborhood of \(S_\tau\) for all \(k\). At any iteration \(k\), \(a^{(k)}\) 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 \(n, p, \text{poly}(n, p)\), 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 \(\|a^{(k)} - \bar{a}\|_2 \leq \varepsilon\) is bounded by \(\text{poly}(n, p) \log \varepsilon^{-1}\). In particular, the number of iterations required to achieve \(\|a^{(k)} - \bar{a}\|_2 \leq \mu + 1/p\) is bounded by a polynomial in \((n, p)\), as claimed.

**K Analysis of algorithm — local refinement**

In this section, we describe and analyze an algorithm which refines an estimate \(a^{(0)} \approx a_0\) of the kernel to exactly recover \((a_0, x_0)\). Set

\[
\lambda^{(0)} \leftarrow 5K\bar{\mu} \quad \text{and} \quad I^{(0)} \leftarrow \text{supp}(S_{\lambda} [C^*_a(y)]), \tag{K.1}
\]

where as each iteration of the algorithm consists of the following key steps:

- **Sparse Estimation using Reweighted Lasso:** Set

  \[
  x^{(k+1)} \leftarrow \arg\min_x \frac{1}{2} \|a^{(k)} \ast x - y\|^2_2 + \sum_{i \notin I^{(k)}} \lambda^{(k)} |x_i|; \tag{K.2}
  \]

- **Kernel Estimation using Least Squares:** Set

  \[
  a^{(k+1)} \leftarrow P_{S_{\rho-1}} \left[ \arg\min_a \frac{1}{2} \|a \ast x^{(k+1)} - y\|^2_2 \right]; \tag{K.3}
  \]

- **Continuation and reweighting by decreasing sparsity regularizer:** Set

  \[
  \lambda^{(k+1)} \leftarrow \frac{1}{2} \lambda^{(k)} \quad \text{and} \quad I^{(k+1)} \leftarrow \text{supp}(x^{(k+1)}). \tag{K.4}
  \]

Our analysis will show that \(a^{(k)}\) 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. Below, we will write

\[
\bar{\mu} = \max \{\mu, p^{-1}\}. \tag{K.5}
\]

Our refinement algorithm will demand an initialization satisfying

\[
\|a^{(0)} - a_0\|_2 \leq \bar{\mu}. \tag{K.6}
\]
Our goal is to show that the proposed annealing algorithm exactly solves the SaS deconvolution problem, i.e., exactly recovers \((a_0, x_0)\) up to a signed shift. We will denote the support sets of true sparse vector \(x_0\) and recovered \(x^{(k)}\) in the intermediate \(k\)-th steps as
\[
I = \text{supp}(x_0), \quad I^{(k)} = \text{supp}(x^{(k)}).
\] (K.7)
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:
\[
\kappa_I = 6 \max \{tp, \log n\},
\] (K.8)
then in the Bernoulli-Gaussian model, with high probability,
\[
\max_{\ell} |I \cap ([p] + \ell)| \leq \kappa_I.
\] (K.9)
Here, indexing and addition should be interpreted modulo \(n\). The \(\log n\) term reflects the fact that as \(n\) becomes enormous (exponential in \(p\)) eventually it becomes likely that some length-\(p\) window of \(x_0\) is densely occupied. In our main theorem statement, we preclude this possibility by putting an upper bound on \(n\) (w.r.t \(\tilde{\mu}\)). We find it useful to also track the maximum \(\ell^2\) norm of \(x_0\) over any length-\(p\) window:
\[
\|x_0\|_\square := \max \ell \|P_{[p] + \ell} x_0\|_2.
\] (K.10)
Below, we will sometimes work with the \(\square\)-induced operator norm:
\[
\|M\|_{\square \rightarrow \square} = \sup \|Mx\|_\square
\] (K.11)
For now, we note that in the Bernoulli-Gaussian model, \(\|x_0\|_\square\) is typically not large
\[
\|x_0\|_\square \leq \sqrt{\kappa_I}.
\] (K.12)

### K.1 Reweighted Lasso finds the large entries of \(x_0\)

The following lemma asserts that when \(a\) is close to \(a_0\), the reweighted Lasso finds all of the large entries of \(x_0\). Our reweighted Lasso is modified version from [CWB08], we only penalize \(x\) on entries outside of its known support subset. We write \(T\) to be the subset of true support \(I\), and define the sparsity surrogate as
\[
\sum_{i \in \bar{T}} |x_i|,
\] (K.13)
The reweighted Lasso recovers more accurate \(x\) 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 K.1** (Accuracy of reweighted Lasso estimate). Suppose that \(y = a_0 * x_0\) with \(a_0\) is \(\tilde{\mu}\)-shift coherent and \(\|x_0\|_\square \leq \sqrt{\kappa_I}\) with \(\kappa_I \geq 1\). If \(\tilde{\mu} \kappa^2 ||c|| \leq c_\mu\), then for every \(T \subseteq I\) and \(a\) satisfying \(\|a - a_0\|_2 \leq \tilde{\mu}\), the solution \(x^+\) to the optimization problem
\[
\min_x \left\{ \frac{1}{2} \|a * x - y\|_2^2 + \lambda \sum_{i \in \bar{T}} |x_i| \right\},
\] (K.14)
with
\[
\lambda > 5\kappa_I \|a - a_0\|_2,
\] (K.15)
is unique with the form
\[
x^+ = \iota_J (C^*_{\alpha, i} C_{a, i})^{-1} \iota_J^* (C^*_\alpha y - \lambda P_{\bar{T}} \sigma)
\] (K.16)
where \( \sigma = \text{sign}(x^+) \). Its support set \( J \) satisfies

\[
(T \cup I_{\geq 3\lambda}) \subseteq J \subseteq I
\]  

(K.17)

and its entrywise error is bounded as

\[
\|x^+ - x_0\|_\infty \leq 3\lambda.
\]  

(K.18)

Above, \( c_\mu > 0 \) is a positive numerical constant.

We prove Lemma K.1 below. The proof relies heavily on the fact that when \( a_0 \) is shift-incoherent and \( a \approx a_0 \), \( a \) is also shift-incoherent, an observation which is formalized in a sequence of calculations in Appendix K.4.

**Proof.** 1. (Restricted support Lasso problem). We first consider the restricted problem

\[
\min_{w \in \mathbb{R}^{|I|}} \left\{ \frac{1}{2} \| a \ast t_I w - y \|_2^2 + \lambda \sum_{i \in T^c} |(t_I w)_i| \right\}.
\]  

(K.19)

Under our assumptions, provided \( c < \frac{1}{\sigma} \), Lemma K.6 implies that

\[
t^*_I C_a^T C_a t_I = |C_a C_a|_{I, I} > 0,
\]  

(K.20)

and the restricted problem is strongly convex and its solution is unique. The KKT conditions imply that a vector \( w_* \) is the unique optimal solution to this problem if and only if

\[
t^*_I C_a C_a t_I w_* \in t^*_I C_a y - \lambda \partial \| P_{T^c} \|_1(w_*)
\]  

(K.21)

Writing \( J = \text{supp}(t_I w_*) \subseteq I \), \( C_{a, J} = C_{a, t_J} \), \( w_J = t^*_I t_I w_* \) the corresponding sub-vector containing the nonzero entries of \( w_* \) and \( \sigma_{J \setminus T} = t^*_I P_{T^c} [\text{sign}(t_I w_*)] \), the condition (K.21) is satisfied if and only if

\[
C_{a, J}^* C_{a, J} w_J = C_{a, J}^* y - \lambda \sigma_{J \setminus T},
\]  

(K.22)

\[
\|C_{a, J}^* (C_{a, J} w_J - y)\|_\infty \leq \lambda.
\]  

(K.23)

We will argue that under our assumptions, \( J \) necessarily contains all of the large entries of \( x_0 \):

\[
I_{>3\lambda} = \{ \ell \in I | |x_{0\ell}| > 3\lambda \} \subseteq J.
\]  

(K.24)

We show this by contradiction – namely, if some large entry of \( x_0 \) is not in \( J \), then the dual condition (K.23) is violated, contradicting the optimality of \( w_* \). To this end, note that by Corollary K.7, \( C_{a, J}^* C_{a, J} \) has full rank. From (K.22),

\[
w_J = [C_{a, J}^* C_{a, J}]^{-1} [C_{a, J}^* y - \lambda \sigma_{J \setminus T}].
\]  

(K.25)

Write \( x_{0,J} = t^*_J x_0 \) and \((x_0)_{I \setminus J} = P_{I \setminus J} x_0 \). We can further notice that

\[
C_{a, J} w_J - y = \left( C_{a, J} [C_{a, J}^* C_{a, J}]^{-1} C_{a, J}^* - I \right) C_{a, J}^* \left( x_{0,J} - C_{a, J} \sigma_{J \setminus T} \right)
\]

\[
= \left( C_{a, J} [C_{a, J}^* C_{a, J}]^{-1} C_{a, J}^* - I \right) C_{a, J}^* x_{0,J}
\]

\[
+ \left( C_{a, J} [C_{a, J}^* C_{a, J}]^{-1} C_{a, J}^* - I \right) C_{a, J} (x_0)_{I \setminus J}
\]

\[- \lambda C_{a, J} [C_{a, J}^* C_{a, J}]^{-1} \sigma_{J \setminus T},
\]

\[
= \left( C_{a, J} [C_{a, J}^* C_{a, J}]^{-1} C_{a, J}^* - I \right) C_{a, J} x_{0,J}
\]

\[
+ \left( C_{a, J} [C_{a, J}^* C_{a, J}]^{-1} C_{a, J}^* - I \right) C_{a, J} (x_0)_{I \setminus J}
\]

\[- \lambda C_{a, J} [C_{a, J}^* C_{a, J}]^{-1} \sigma_{J \setminus T},
\]  

(K.26)
where in the final line we have used that
\[
(C_{a,J} [C_{a,J}^* C_{a,J}]^{-1} C_{a,J} - I) C_{a,J} = 0.
\] (K.27)

Suppose that \( J \) is a strict subset of \( I \) (otherwise, if \( J = I \), we are done). Take any \( i \in I \setminus J \) such that \( |x_{0i}| = \| (x_0)_{I \setminus J} \|_\infty \), and let \( \xi = \text{sign}(x_{0i}) \). Using (K.26), Corollary K.7 and Lemma K.8, and simplify the induced norms \( \| \cdot \|_\infty \) and \( \| \cdot \|_2 \) as \( \| \cdot \|_\infty \) and \( \| \cdot \|_2 \), we have
\[
-\xi s_i[a]^*(C_{a,J} w_J - y) = \xi s_i[a]^* \left( I - C_{a,J} [C_{a,J}^* C_{a,J}]^{-1} C_{a,J}^* \right) s_i[a_0] x_{0i} + \xi s_i[a]^* \left( I - C_{a,J} [C_{a,J}^* C_{a,J}]^{-1} C_{a,J}^* \right) C_{a_0}(x_0)_{I \setminus (J \cup \{i\})} + \xi s_i[a]^* \left( I - C_{a,J} [C_{a,J}^* C_{a,J}]^{-1} C_{a,J}^* \right) C_{a_0-a_J} x_{0,J} + \xi \lambda s_i[a]^* C_{a,J} [C_{a,J}^* C_{a,J}]^{-1} \sigma_J \setminus T
\] (K.28)
\[
geq \left( s_i[a], s_i[a_0] \right) - \| s_i[a]^* C_{a,J} \|_1 \left( \| C_{a,J}^* C_{a,J}^{-1} \|_\infty \| C_{a,J}^* s_i[a_0] \|_\infty \right) \| (x_0)_{I \setminus J} \|_\infty
+ \left( \| s_i[a]^* C_{a_0-iJ} \|_1 \right) + \| s_i[a]^* C_{a,J} \|_2 \left( \| C_{a,J}^* C_{a,J}^{-1} \|_\infty \| C_{a,J} C_{a_0-iJ} \|_\infty \right) \| (x_0)_{I \setminus J} \|_\infty
- \lambda \| s_i[a]^* C_{a,J} \|_1 \left( \| C_{a,J}^* C_{a,J}^{-1} \|_\infty \| \sigma_j \setminus T \|_\infty \right) \| (x_0)_{I \setminus J} \|_\infty
\] (K.29)
\[
geq \left( 1 - \| a - a_0 \|_2 \right) - C_1 \kappa I \tilde{\mu} \times I \times \tilde{\mu} \left( (x_0)_{I \setminus J} \right) \| (x_0)_{I \setminus J} \|_\infty
- C_2 \left( \kappa I \tilde{\mu} + \kappa I \tilde{\mu} \times I \times \kappa I \tilde{\mu} \right) \| (x_0)_{I \setminus J} \|_\infty
- \left( 2 \kappa I \tilde{\mu} a - a_0 \|_2 + C_3 \kappa I \tilde{\mu} \times I \times \kappa I \| a - a_0 \|_2 \right) \| (x_0)_{I \setminus J} \|_\infty
- \lambda C_4 \kappa I \tilde{\mu}
\] (K.30)
\[
geq \left( 1 - C_1 \kappa I \tilde{\mu} - C_2 \left( \kappa I \tilde{\mu} \right)^2 \right) \| (x_0)_{I \setminus J} \|_\infty
- 2 \kappa I \| a - a_0 \|_2 - \left( C_3 \kappa I \tilde{\mu} \times I \times \kappa I \| a - a_0 \|_2 \right) \| (x_0)_{I \setminus J} \|_\infty
- \lambda C_4 \kappa I \tilde{\mu}
\] (K.31)
\[
geq \frac{1}{2} \| (x_0)_{I \setminus J} \|_\infty - \lambda/2,
\] (K.32)
where the last line holds provided \( \tilde{\mu} \kappa^2 \leq \epsilon_\mu \) to be a sufficiently small numerical constants. If \( \| (x_0)_{I \setminus J} \|_\infty > 3 \lambda \), this is strictly larger than \( \lambda \), implying that \( |a|^* C_{a,J} w_J - y | > \lambda \), and contradicting the KKT conditions for the restricted problem. Hence, under our assumptions
\[
\| (x_0)_{I \setminus J} \|_\infty \leq 3 \lambda.
\] (K.33)

2. (Solution of Full Lasso problem) We next argue that the solution of the restricted support Lasso problem, \( w_J \), when extended to \( \mathbb{R}^n \) as \( x^* = I_J w_J \), is the unique optimal solution to the full Lasso problem
\[
\min_x \varphi_{lasso}(x) = \frac{1}{2} \| a + x - y \|_2^2 + \lambda \sum_{i \in T^x} | x_i |.
\] (K.34)
To prove that \( x^+ \) is the unique optimal solution, it suffices to show that for every \( i \in I' \),
\[
|s_i[a]^*(a \ast x^+ - y)| \leq \lambda. \tag{K.35}
\]

Indeed, suppose that this inequality is in force. Write \( \varepsilon = \lambda - \max_{i \in I'} |s_i[a]^*(a \ast x^+ - y)| \), and notice that from the KKT conditions for the restricted problem,
\[
0 \in P_{I'} \partial \varphi_{\text{lasso}}(x) \tag{K.36}
\]
Combining with (K.35), we have that for every vector \( \zeta \) with \( \text{supp}(\zeta) \subseteq I' \) and \( \|\zeta\|_\infty \leq 1 \), then \( \varepsilon \zeta \in \partial \varphi_{\text{lasso}}(x^+) \). Let \( x' \) be any vector with \( x'_J \neq 0 \) and set \( \zeta = P_{I'} \text{sign}(x') \), then from the subgradient inequality,
\[
\varphi_{\text{lasso}}(x') \geq \varphi_{\text{lasso}}(x^+) + \langle \varepsilon \zeta, x' - x^+ \rangle \geq \varphi_{\text{lasso}}(x^+) + \varepsilon \|x'_J\|_1,
\]
which is strictly larger than \( \varphi_{\text{lasso}}(x^+) \). Hence, when (K.35) holds, any optimal solution \( \bar{x} \) to the full Lasso problem must satisfy \( \text{supp}(\bar{x}) \subseteq I \). By strong convexity of the restricted problem, the solution to (K.34) is unique and equal to \( x^+ \).

We finish by showing (K.35). Using the same expansion as above, we obtain
\[
|s_i[a]^*(C_{a,j}w_J - y)| \leq |s_i[a]^*(I - C_{a,j}[C_{a,j}^*C_{a,j}]^{-1}C_{a,j}^*)C_{a_0 - a,J}(x_0)_I J| + |s_i[a]^*(I - C_{a,j}[C_{a,j}^*C_{a,j}]^{-1}C_{a,j}^*)C_{a_0 - a,J}x_0 J| + \lambda |s_i[a]^*C_{a,j}[C_{a,j}^*C_{a,j}]^{-1} \sigma_{J,T}| \leq \left( \|s_i[a]^*C_{a_0 - a,J}\|_1 + \|s_i[a]^*C_{a,J}\|_1 \left\|\left[C_{a,j}^*C_{a,j}\right]^{-1}\right\|_\infty \right) \|C_{a,j}^*C_{a_0 - a,J}\|_\infty \|x_0\|_\infty + \lambda \|s_i[a]^*C_{a,J}\|_1 \left\|\left[C_{a,j}^*C_{a,j}\right]^{-1}\right\|_\infty \|\sigma_{J,T}\|_\infty \leq C_1 (\tilde{\mu}_K \times 1 \times \tilde{\mu}_K) \times 2\lambda + (2\sqrt{K_I} \|a - a_0\|_2 + C_2 \sqrt{K_I} \tilde{\mu} \times 1 \times \tilde{\mu}_K \|a - a_0\|_2) \times \sqrt{K_I} \leq \left( (C_1 + C_3) \tilde{\mu}_K + C_1 (\tilde{\mu}_K^2) \lambda + (2 + C_2 \tilde{\mu}_K) \tilde{\mu}_K \|a - a_0\|_2 \right)
\]
(3) (Entrywise difference to \( x_0 \)) Finally we will be controlling \( \|x^+_J - (x_0)_J\|_\infty \). Indeed, from Corollary K.7, Lemma K.8,
\[
\|x^+_J - (x_0)_J\|_\infty \leq \left\|\left[C_{a,j}^*C_{a,j}\right]^{-1}C_{a,j}^*C_{a_0 - a,J}x_0 - \lambda \left[C_{a,j}^*C_{a,j}\right]^{-1} \sigma_{J,T} - (x_0)_J\right\|_\infty \leq \left\|\left[C_{a,j}^*C_{a,j}\right]^{-1}C_{a,j}^*C_{a_0 - a,J}(x_0)_J\right\|_\infty + \lambda \left\|\left[C_{a,j}^*C_{a,j}\right]^{-1} \sigma_{J,T}\right\|_\infty + \left\|\left[C_{a,j}^*C_{a,j}\right]^{-1}C_{a_0 - a,J}(x_0)_I J\right\|_\infty \leq 2 \|C_{a,j}^*C_{a_0 - a,J}\|_\infty \|x_0\|_\infty + \lambda \|\sigma_{J,T}\|_\infty \leq \lambda \|
\]
To this end, since we know the solution of least square problem

\[
2\lambda + 2 \| C_{a,j} C_{a,I \setminus j} \|_\infty \| (x_0)_{I \setminus j} \|_\infty \\
\leq 2\sqrt{2\kappa_I} \| a - a_0 \|_2 \| x_0 \|_\square + 2\lambda + 2 \times 3\mu \times 2\kappa_{I,I \setminus j} \times 3\lambda \\
\leq 3\kappa_I \| a - a_0 \|_2 + 2\lambda + 36\lambda \mu \kappa_I \\
\leq 3\lambda,
\]

(K.43)

establishing the claim.

\[\square\]

### K.2 Least squares solution \( a^{(k)} \) contracts

In this section, given \( x \) to be the solution to the reweighted Lasso from \( a \), we will show the solution of the least squares problem

\[
a^+ \leftarrow \arg\min_{a' \in \mathbb{R}^p} \frac{1}{2} \| a' * x - y \|_2^2
\]

is closer to \( a_0 \) compared to \( a \). Observe that in Lemma K.1, the solution of (K.16)

\[
x = \iota_j (C_{a,j}^* C_{a,j})^{-1} \iota_j^* (C_{a,j}^* C_{a,j} x_0 - \lambda P_{J \setminus T} \sigma),
\]

by assuming \( C_{a,j}^* C_{a,j} \approx I, a \approx a_0 \) and \( J \setminus T \approx \emptyset \), is a good approximation to the true sparse map \( x_0 \)

\[
x \approx I (x_0 - 0) = x_0;
\]

furthermore, its difference to the true sparse map \( \| x_0 - x \|_2 \) is proportional to \( \| a_0 - a \|_2 \) as

\[
x - x_0 \approx P_I (C_{a,j}^* C_{a,j} x_0 - C_{a,j}^* C_{a,j} x_0) \approx P_I (C_{a,j}^* C_{a,j} x_0 I (a_0 - a)).
\]

(K.47)

To this end, since we know the solution of least square problem \( a^+ \) is simply

\[
a^+ = (\iota^* C_x^* C_x \iota)^{-1} (\iota^* C_x^* C_x \iota a_0),
\]

(G.48)

this implies the difference between the new \( a^+ \) and \( a_0 \), has the relationship with \( a - a_0 \) roughly

\[
a^+ - a_0 = (\iota^* C_x^* C_x \iota)^{-1} (\iota^* C_x^* C_x \iota a_0 - \iota^* C_x C_x \iota a_0)
\approx (n\theta)^{-1} \iota^* C_x^* C_x \iota a_0 (x_0 - x)
\approx (n\theta)^{-1} \iota^* C_x^* C_x \iota P_I C_{a,j}^* C_{a,j} x_0 I (a - a_0).
\]

(K.49)

To make this point precise, we introduce the following lemma:

**Lemma K.2 (Approximation of least square estimate).** Given \( a_0 \in \mathbb{R}^n \) to be \( \mu \)-shift coherent and \( x_0 \sim \mathcal{B}(\theta) \in \mathbb{R}^n \). There exists some constants \( C, C' \) such that if \( \lambda < c' \mu \kappa_I, \mu \kappa_I^2 \leq c_\mu \) and \( n > C p^2 \log p \), then with probability at least \( 1 - c/n \), for every \( a \) satisfying \( \| a - a_0 \|_2 \leq \mu \) and \( x \) of the form

\[
x = \iota_j (C_{a,j}^* C_{a,j})^{-1} \iota_j^* (C_{a,j}^* y - \lambda P_{J \setminus T} \sigma)
\]

(K.50)

where the set \( J, T \) satisfies \( J \geq 6\lambda \subseteq T \subseteq J \subseteq I \), we have

\[
\frac{1}{n\theta} \| \iota^* C_x^* C_x \iota a_0 - \iota^* C_x^* C_x \iota P_I C_{a,j}^* C_{a,j} x_0 I (a - a_0) \|_2 \\
\leq C' \lambda (\hat{\lambda} + \tilde{\mu} \kappa_I) + \frac{1}{32} \| a - a_0 \|_2
\]

(K.51)

with \( \hat{\lambda} = \lambda + \frac{\log n}{\sqrt{n\theta}} \).
Proof. We will begin with listing the conditions we use for both $x$ and $x_0$. First, we know from Lemma K.1 and our assumptions on the set $T$, then $x$ approximates $x_0$ in the sense that
\[
\|x - x_0\|_\infty \leq 3\lambda
\]  
(K.52)
\[
\|(x_0)_{T \setminus J}\|_\infty \leq 3\lambda
\]  
(K.53)
\[
\|(x_0)_{T \setminus J}\|_2 \leq 6\lambda.
\]  
(K.54)
Write $x_0 = g \circ \omega$ with $g$ iid standard normal, $\omega$ iid Bernoulli and $g$ and $\omega$ independent. From (K.53) we know $|I \setminus J| = \{i \mid |g_i| \leq 3\lambda, \omega_i \neq 0\}$. Since $P[\omega_i \neq 0] = \theta$ and $P[|g_i| \leq 3\lambda] \leq 3\lambda$, Lemma D.1 implies that with probability at least $1 - 2/\lambda$:
\[
|I \setminus J| \leq 3\lambda n \theta + 6\sqrt{\lambda \theta \log n} \leq 3\lambda n \theta
\]  
(K.55)
\[
|I \setminus T| \leq 6\lambda n \theta + 12\sqrt{\lambda \theta \log n} \leq 6\lambda n \theta,
\]  
(K.56)
and
\[
|(I \setminus J) \cap s_t[I]| \leq 3\lambda n \theta^2 + 6\sqrt{\lambda \theta^2 \log n} \leq 3\lambda n \theta^2;
\]  
(K.57)
together with base on properties of Bernoulli-Gaussian vector $x_0$ from Appendix D and we conclude with probability at least $1 - c/n$, all the following events hold:
\[
\frac{1}{2} n \theta \leq |I| \leq 2n \theta,
\]  
(K.58)
\[
\max_{t \neq 0} |I \cap s_t[I]| \leq 2n \theta^2
\]  
(K.59)
\[
\max_{t \neq 0} |(I \setminus J) \cap s_t[I]| \leq \tilde{\lambda} n \theta^2,
\]  
(K.60)
\[
\|x_0\|_{\square}^2 \leq \kappa_f,
\]  
(K.61)
\[
\|\tilde{a}_0 + x_0\|_{\square}^2 \leq \kappa_f,
\]  
(K.62)
\[
\|x_0\|_2 \leq 2n \theta,
\]  
(K.63)
\[
\|x_0\|_1 \leq 2n \theta,
\]  
(K.64)
\[
\max_{t \neq 0} \|P_{I \cap s_t[I]} x_0\|_2^2 \leq 2n \theta^2,
\]  
(K.65)
\[
\max_{t \neq 0} \|P_{I \cap s_t[I \setminus J]} x_0\|_1 \leq 12\lambda n \theta^2,
\]  
(K.66)
\[
\|C_{x_0, t}\|_2^2 \leq 3n \theta,
\]  
(K.67)
provided by $n \geq C \theta^{-2} \log p$ for sufficiently large constant $C$.

1. (Approximate $C_x$ with $C_{x_0}$) Since
\[
\lambda^* C_{x-x_0, t a_0} = \lambda^* C_{x_0} C_{x-x_0, t a_0} + \lambda^* C_{x-x_0} C_{x-x_0, t a_0}
\]  
(K.68)
where
\[
\|\lambda^* C_{x-x_0} C_{x-x_0, t a_0}\|_2 \leq \|a_0\|_2 \|x - x_0\|_2^2 + \|C_{a_0, t}\|_2 \sqrt{2p} \max_{t \neq 0} |s_t[x - x_0], x - x_0| \leq \|x - x_0\|_2^2 \times |I| + \sqrt{2p} \max_{t \neq 0} |I \cap s_t[I]| \leq C_1 \left(\lambda^2 n \theta + \sqrt{2p} (\lambda^2 n \theta^2) \right) \leq 2C_1 \lambda^2 n \theta,
\]  
(K.69)
we have that

\[ \| \iota^* C^* \mathcal{C}_{x-x_0} \iota a_0 - \iota^* C^* \mathcal{C}_{x-x_0} \iota a_0 \|_2 \leq 2C_1 \lambda^2 n \theta. \]  

(K.70)

2. (Extract the \(a_0 - a\) term) Observe that

\[
\begin{align*}
\iota^* C^* \mathcal{C}_{x-x_0} \iota a_0 &= \iota^* C^* \mathcal{C}_{a_0} (x-x_0) \\
&= \iota^* C^* \mathcal{C}_{a_0} \left( t_j (C^*_{aJ} C_{aJ})^{-1} t_j^* (C^*_{a0} C_{a0} x_0 - \lambda P_{J\setminus T} \sigma) - t_j (C^*_{aJ} C_{aJ})^{-1} (C^*_{a0} C_{a0} x_0 - P_{J_1 \setminus J} x_0) \right) \\
&= \iota^* C^* \mathcal{C}_{a_0} t_j (C^*_{aJ} C_{aJ})^{-1} (C^*_{a0} C_{a0} - C_{aJ} (x_0)) \\
&\quad + \iota^* C^* \mathcal{C}_{a_0} (C^*_{aJ} C_{aJ})^{-1} (C^*_{a0} C_{a0} - C_{aJ} (x_0)) \\
&\quad - \lambda \iota^* C^* \mathcal{C}_{a_0} (C^*_{aJ} C_{aJ})^{-1} t_j^* P_{J\setminus T} \sigma, \\
\end{align*}
\]

(K.71)

where, the second term in (K.71) is bounded as

\[
\begin{align*}
\| \iota^* C^* \mathcal{C}_{a_0} t_j (C^*_{aJ} C_{aJ})^{-1} (C^*_{a0} C_{a0} - C_{aJ} (x_0)) \|_2 \\
&\leq \| C_{a_0} \|_2 \| t_j \|_2 \| (C^*_{aJ} C_{aJ})^{-1} \|_2 \\
&\quad \times \| (x_0)_{J_1 \setminus J} \|_2 \\
&\leq C_2 \left( \sqrt{n \theta} \times 3 \bar{\mu} \kappa I \times \lambda \sqrt{\lambda n \theta} \right) \\
&\leq 3C_2 \bar{\mu} \kappa I \lambda n \theta. \\
\end{align*}
\]

(K.72)

the third term in (K.71) is bounded as

\[
\begin{align*}
\| \iota^* C^* \mathcal{C}_{a_0} t_j (C^*_{aJ} C_{aJ})^{-1} (C^*_{a0} C_{a0} - C_{aJ} (x_0)) \|_2 \\
&\leq \| t_j \|_2 \| (x_0)_{J_1 \setminus J} \|_2 \\
&\quad + \| C_{a_0} \|_2 \sqrt{2p} \max_{j \neq 0} \| P_{I \setminus J} x_0 \|_1 \times \| (x_0)_{J_1 \setminus J} \|_\infty \\
&\leq C_3 \left( \sqrt{\lambda^2 + \lambda n \theta} + \sqrt{\bar{\mu}^2 \times \lambda n \theta^2} \times \lambda \right) \\
&\leq 2C_3 \bar{\mu} \kappa I \lambda n \theta, \\
\end{align*}
\]

(K.73)

and finally, write \( \Delta = (C^*_{aJ} C_{aJ})^{-1} - I \), then the forth term in (K.71) is bounded as

\[
\begin{align*}
\lambda \| \iota^* C^* \mathcal{C}_{a_0} t_j (C^*_{aJ} C_{aJ})^{-1} t_j^* P_{J\setminus T} \sigma \|_2 \\
&= \lambda \| \iota^* C^* \mathcal{C}_{a_0} (P_{I \setminus J} + e_0 e_0^* C^*_{aJ} t_j (I + \Delta) t_j^* P_{J\setminus T} \sigma \|_2 \\
&\leq \lambda \| C^*_{a_0} t_j \|_2 \sqrt{2p} \max_{j \neq 0} \| P_{I \setminus J} x_0 \|_1 + \lambda \| a_0 \|_2 \| P_{J\setminus T} x_0 \|_1 \\
&\quad + \lambda \| C^*_{a_0} t_j \|_2 \sqrt{2p} \| P_{I \setminus J} x_0 \|_1 \| \Delta \|_{\infty \rightarrow \infty} \\
&\quad + \lambda \| a_0 \|_2 \| x_0 \|_2 \| \Delta \|_2 \sqrt{|J \setminus T|} \\
&\leq C_4 \lambda \left( \sqrt{\bar{\mu}^2 \times \lambda n \theta^2} + \lambda \bar{\mu} \kappa I \right) \\
&\quad + \sqrt{\bar{\mu}^2 \times n \theta^2} \times \bar{\mu} \kappa I + \sqrt{\bar{\mu} \kappa I} \sqrt{\lambda n \theta} \\
&\leq 2C_4 \left( \lambda + \bar{\mu} \kappa I \right) \lambda n \theta. \\
\end{align*}
\]

(K.74)
Therefore, combining (K.72)-(K.74) we obtain
\[
\| \iota^* C_{x_0}^* C_{x_0} t a_0 - \iota^* C_{x_0}^* C_{a_0} (C_{a,j}^* C_{a,j})^{-1} C_{a,j} C_{a_0-a} x_0 \|_2 \\
\leq C_5 \left( \lambda + \mu \kappa_j \right) \lambda n \theta.
\]  
(K.75)

3. (Extract the set \( J \)) Lastly, we will further simplify the term with \( a - a_0 \) in (K.75) by extracting the set \( J \):
\[
\iota^* C_{x_0}^* C_{a_0} (C_{a,j}^* C_{a,j})^{-1} C_{a,j} C_{a_0-a} x_0 \\
= \iota^* C_{x_0}^* C_{a_0} (I + \Delta) C_{a_0}^{\ast} (a_0 - a_0) J C_{x_0} t (a_0 - a) \\
= \iota^* C_{x_0}^* C_{a_0} P_I C_{a_0} C_{x_0} t (a_0 - a) \\
+ \iota^* C_{x_0}^* C_{a_0} (C_{a,j}^* C_{a,j})^{-1} C_{a-a_0} J C_{x_0} t (a_0 - a) \\
- \iota^* C_{x_0}^* C_{a_0} P_I C_{a_0} C_{x_0} t (a_0 - a),
\]  
(K.76)
where, the latter terms in (K.76) are bounded as
\[
\| \iota^* C_{x_0}^* C_{a_0} (I + \Delta) C_{a_0}^{\ast} C_{a_0-a} x_0 \|_2 \\
\leq \| C_{x_0} t \|_2 \| C_{a_0} t \|_2 \| \Delta \|_2 \leq C_6 \mu \kappa_j n \theta \\
\| \iota^* C_{x_0}^* C_{a_0} (C_{a,j}^* C_{a,j})^{-1} C_{a-a_0} x_0 \|_2 \\
\leq \| C_{x_0} t \|_2 \| C_{a_0} t \|_2 \| (C_{a,j}^* C_{a,j})^{-1} \|_2 \| C_{a_0-a} t \|_2 \leq C_7 \mu \kappa_j n \theta \\
\| P_I \setminus C_{a_0}^{\ast} C_{x_0} t \|_2 \\
\leq C_8 \lambda n \theta \times \kappa_j \leq C_8 \left( \lambda \kappa_j + \frac{\log n}{\sqrt{n \theta}} \right) n \theta,
\]  
(K.77)
whence we conclude, that since \( c_\mu \kappa_j^2 \leq c_\mu \) and \( \lambda \kappa_j \leq 5 c_\mu \), as long as \( c_\mu < \frac{1}{10} \left( \frac{1}{c_\mu} + \frac{1}{c_\sigma} + \frac{1}{c_\tau} \right) \) and \( n > 10^6 C_8^2 \theta^{-2} \kappa_j^2 \log^2 n \), we gain:
\[
\| \iota^* C_{x_0}^* C_{a_0} (C_{a,j}^* C_{a,j})^{-1} C_{a,j} C_{a_0-a} x_0 \|_2 \\
- \iota^* C_{x_0}^* C_{a_0} P_I C_{a_0} C_{x_0} t (a_0 - a) \|_2 \\
\leq \left( \frac{3}{100} + \frac{1}{1000} \right) n \theta \| a_0 - a \|_2 \\
\leq \frac{1}{32} n \theta \| a_0 - a \|_2.
\]  
(K.78)
The claimed result therefore is followed by combining (K.70), (K.75) and (K.78).

The next thing is to show the operator
\[
(n \theta)^{-1} (\iota^* C_{x_0}^* C_{a_0} P_I C_{a_0} C_{x_0} t)
\]  
(K.79)
contracts \( a \) toward \( a_0 \). We first will show that
\[
(n \theta)^{-1} (\iota^* C_{x_0}^* C_{a_0} P_I C_{a_0} C_{x_0} t) \approx a_0 a_0^* 
\]  
(K.80)
by seeing \( \iota^* C_{x_0}^* P_I C_{x_0} t \approx (n \theta) e_0 e_0^* \) via sparsity of \( x_0 \). Finally since the local perturbation on sphere is close to a quadratic function in \( \ell^2 \)-norm of difference, we have
\[
|\langle a_0, a - a_0 \rangle| \leq \frac{1}{2} \| a - a_0 \|_2^2.
\]  
(81)
Again, we introduce the following lemma to solidify our claim:
Lemma K.3 (Contraction of $a$ to $a_0$). Given $a_0 \in \mathbb{R}^m$ to be $\bar{\mu}$-shift coherent and $x_0 \sim \text{BG}(\theta) \in \mathbb{R}^n$. There exists some constants $C, C', c, c', c_\mu$ such that if $\lambda < c' \mu \kappa_1, \mu \kappa_1^2 \leq c_\mu$ and $n > C \theta^{-2} p^2 \log p$, then with probability at least $1 - c/n$, for every $\|a - a_0\|_2 \leq \bar{\mu}$,

$$\| \iota^* C_{x_0}^* C_{a_0} P_{I} C_{a_0}^* C_{x_0} \iota (a_0 - a) \|_2 \leq \frac{1}{32}\|a - a_0\|_2 n \theta. \quad (K.82)$$

Proof. Since $\mathbb{E} \langle P_{s_i} [x_0], s_j [x_0] \rangle = 0$ for all $i \neq j$ and set $I$, we calculate

$$\mathbb{E} [\iota^* C_{x_0}^* C_{a_0} P_{I} C_{a_0}^* C_{x_0} \iota] = \sum_{i \in [\pm p]} \mathbb{E} [e_i^* C_{x_0}^* P_{I} C_{x_0} e_i] e_i e_i^*$$

$$= \mathbb{E} \|x_0\|^2 e_0 e_0^* + \sum_{i \in [\pm p] \setminus 0} \mathbb{E} \|P_{s_i} [x_0]\|_2^2 e_i e_i^*$$

$$= n \theta e_0 e_0^* + n \theta^2 P_{[\pm p] \setminus 0}$$

$$= n \theta^2 I + n \theta (1 - \theta) e_0 e_0^*. \quad (K.83)$$

whence

$$\mathbb{E} [\iota^* C_{x_0}^* C_{a_0} P_{I} C_{a_0}^* C_{x_0} \iota] = \iota^* C_{x_0}^* C_{a_0} \mathbb{E} [C_{x_0}^* P_{I} C_{x_0}] C_{a_0} \iota$$

$$= n \theta^2 \iota^* C_{a_0} C_{a_0} \iota + n \theta (1 - \theta) a_0 a_0^*,$$

implying the expectation is a contraction mapping for $a_0 - a$ when $c_\mu < \frac{1}{200}$:

$$\| \mathbb{E} [\iota^* C_{x_0}^* C_{a_0} P_{I} C_{a_0}^* C_{x_0} \iota] (a_0 - a) \|_2$$

$$\leq n \theta^2 \| \iota^* C_{a_0} C_{a_0} \iota \|_2 \|a_0 - a\|_2 + n \theta \|a_0\|_2 \langle a_0, a_0 - a \rangle$$

$$\leq n \theta^2 \times 2 \bar{\mu} p \times \|a_0 - a\|_2 + \frac{1}{2} n \theta \|a_0 - a\|_2^2$$

$$\leq (2c_\mu + \frac{1}{2} c_\mu) \|a_0 - a\|_2 n \theta$$

$$\leq \frac{1}{64} \|a_0 - a\|_2 n \theta. \quad (K.84)$$

For each entry of $C_{x_0}^* P_{I} C_{x_0}$, again from Appendix D we know with probability at least $1 - c/n$:

$$\| e_i^* C_{x_0}^* P_{I} C_{x_0} e_j - \mathbb{E} [e_i^* C_{x_0}^* P_{I} C_{x_0} e_j] \| \leq \begin{cases} C' \sqrt{n \theta^2 \log n} & i = j = 0 \\ C' \sqrt{n \theta^2 \log n} & \text{otherwise} \end{cases}.$$ 

Thus via Gershgorin disc theorem, when $n > 10^3 C' \theta^{-2} p^2 \log n$:

$$\lambda_{\text{max}} \left( \iota^* C_{x_0}^* P_{I} C_{x_0} \iota \right) - \mathbb{E} \left[ \iota^* C_{x_0}^* P_{I} C_{x_0} \iota \right] \leq C' p \sqrt{n \theta^2 \log n} \leq \frac{1}{64} n \theta^2. \quad (K.86)$$

Finally we combine (K.85), (K.86) and get

$$\| \iota^* C_{x_0}^* C_{a_0} P_{I} C_{a_0}^* C_{x_0} \iota (a_0 - a) \|_2 \leq \left( \frac{1}{64} n \theta + \frac{1}{64} n \theta^2 \| C_{a_0} \iota \|_2^2 \|a_0 - a\|_2 \right) \|a_0 - a\|_2$$

$$\leq \frac{1}{32} \|a_0 - a\|_2 n \theta. \quad (K.87)$$

\[\blacksquare\]

Lemma K.1-K.3 together implies the single iterate contract of alternating minimization contracts $a$ toward $a_0$. We show it with the following lemma:
Lemma K.4 (Contraction of least square estimate). Given \( a_0 \in \mathbb{R}^p \) to be \( \tilde{\mu} \)-shift coherent and \( x_0 \sim \text{BG}(\theta) \in \mathbb{R}^n \). There exists some constants \( C, C', c, c_\mu \), such that if \( \tilde{\mu} \kappa_T \leq c_\mu \) and \( n > C\theta^{-2}p^2 \log n \), then with probability at least \( 1 - c/n \), for every \( \lambda \) and \( a \) satisfying
\[
5\tilde{\mu} \kappa_T \geq \lambda \geq 5\kappa_T \| a - a_0 \|_2,
\] (K.88)
and suppose \( x^+ \) has the form of (K.16), then the solution \( a^+ \) to
\[
\min_{a' \in \mathbb{R}^p} \left\{ \| a' + x^+ - y \|_2^2 \right\}
\] (K.89)
is unique and satisfies
\[
\| P_{y^+} \left[ a^+ \right] - a_0 \|_2 \leq \frac{1}{2} \| a - a_0 \|_2.
\] (K.90)

Proof. Write \( x \) as \( x^+ \), then
\[
\lambda_p \left( \ast C \right) = \sigma_{\min}^2 \left( C_{x_0} + C_{x_0 - x_0} \right)
\]
\[
\geq \left[ \sigma_{\min}(C_{x_0}) - \| C_{x_0 - x_0} \|_2 \right]^2
\]
\[
\geq \left[ \sigma_{\min}(C_{x_0}) - 2 \sqrt{\kappa_T} \| x - x_0 \|_2 \right]^2
\]
\[
\geq \left[ \frac{2}{3} \sqrt{\theta n} - 8 \lambda \sqrt{\kappa_T \sqrt{\theta n}} \right]^2
\]
\[
\geq \frac{1}{2} \theta n,
\] (K.91)
where the fourth inequality is derived from using the upper bound of sparse convolution matrix from Remark D.6, and the last line holds by knowing \( \lambda < 5c_\mu \kappa_T^{-1} \). From (K.91) we know the least square problem of (K.89) has unique solution \( a^+ \), written as
\[
a^+ = \left( \ast C \right)^{-1} \ast C y,
\] (K.92)
whence
\[
a^+ - a_0 = \left( \ast C \right)^{-1} \left( \ast C a \right) a_0 - a_0
\]
\[
= \left( \ast C \right)^{-1} \left( \ast C a \right) a_0.
\] (K.93)

Combine Lemma K.2 and Lemma K.3, we know
\[
\left\| \ast C \right\|_2 \left( \ast C \right) a_0 - x_0 \|_2 \leq \left( C_1 \lambda \left( \lambda + \tilde{\mu} \kappa_T \right) + \frac{1}{16} \| a - a_0 \|_2 \right) n \theta
\] (K.94)
for some constant \( C_1 \). Combine (K.91), (K.93), (K.94) and since \( \lambda < \tilde{\mu} \kappa_T \), by letting \( c_\mu < \frac{1}{16} \lambda \), we gain
\[
\| a^+ - a_0 \|_2 \leq \left\| \ast C \right\|_2 \left( \ast C \right) a_0 - x_0 \|_2 \left( \lambda_p \left( \ast C \right) \right)
\]
\[
\leq 2C_1 \lambda \left( \lambda + \tilde{\mu} \kappa_T \right) + \frac{1}{8} \| a - a_0 \|_2 \| a - a_0 \|_2 \leq \frac{1}{4}
\] (K.95)

For the final bound,
\[
\left\| \frac{a^+ - a_0}{\| a^+ \|_2} \right\|_2 \leq \frac{\| a^+ - a_0 \|_2 + \| a^+ \|_2 - 1}{\| a^+ \|_2}
\]
\[
\leq \frac{2 \| a^+ - a_0 \|_2}{1 - \| a^+ - a_0 \|_2} \leq \frac{8}{3} \| a^+ - a_0 \|_2,
\]
\[ \leq C_2 \lambda \left( \lambda + \tilde{\mu} \kappa_I \right) + \frac{1}{3} \|a - a_0\|_2, \tag{K.96} \]

and since \( \lambda > \kappa_I \|a - a_0\|_2 \), finally we gain

\[ (K.96) \leq C_2 \left( \lambda \kappa_I + \frac{p \kappa_I \log n}{n \theta} + \tilde{\mu} \kappa_I^2 \right) \|a - a_0\|_2 + \frac{1}{3} \|a - a_0\|_2 \]

\[ \leq \frac{1}{2} \|a - a_0\|_2 \tag{K.97} \]

as long as \( n > 20 C_2 \theta^{-1} p \kappa_I \log n \) and \( c_\mu < \frac{1}{20 C_2} \).

\[ \square \]

**K.3 Linear convergence of alternating minimization (Proof of Theorem C.2)**

In the first two sections we have shown the iterate contract \( a \) toward \( a_0 \), 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 \( a_0 \):

**Proof.** We will prove our claim by induction on \( k \). Clearly, when \( k = 0 \), we have \( 5 \kappa_I \|a^{(0)} - a_0\|_2 \leq \lambda^{(0)} = 5\tilde{\mu} \kappa_I \) and \( I^{(0)} = \{ i : |s_i[a^{(0)}]^* \} |^* C_{a_0} x_0 > \lambda^{(0)} \} \). Then for all \( |x_j| > 6 \lambda^{(0)} \), we have

\[ \left| s_j \left[ a^{(0)} \right]^* C_{a_0} x_0 \right| \geq (1 - |\{a^{(0)} a_0\}|) |x_j| - \|P_{\{\pm 1\} \backslash \{j\}} C_{a_0}^* \left[ a^{(0)} \right] \|_2 \times \sqrt{2} \|x_0\| \]

\[ \geq (1 - 2\tilde{\mu}) 6 \lambda^{(0)} - 2\tilde{\mu} \sqrt{2} \lambda \]

\[ \geq 5 \lambda^{(0)} - 4 \lambda^{(0)} \]

\[ = \lambda^{(0)}. \tag{K.98} \]

hence \( I_{> 6 \lambda^{(0)}} \subseteq I^{(0)} \), therefore the condition of Lemma K.4 is satisfied, implies (C.32) holds for \( k = 0 \).

Suppose it is true for \( 1, 2, \ldots, k - 1 \), that \( \kappa_I \|a^{(k)} - a_0\|_2 \leq \frac{1}{2} \lambda^{(k-1)} = \lambda^{(k)} \), and \( I_{> 3 \lambda^{(k-1)}} \subseteq I^{(k)} \)

and since \( I_{> 6 \lambda^{(k)}} = I_{> 3 \lambda^{(k-1)}} \subseteq I^{(k)} \), we can again apply Lemma K.4, resulting

\[ \kappa_I \|a^{(k+1)} - a_0\|_2 \leq \frac{1}{2} \kappa_I \|a^{(k)} - a_0\|_2 \leq \frac{1}{2} \lambda^{(k)} \tag{K.100} \]

as claimed.

\[ \square \]

**K.4 Supporting lemmas for refinement**

The following lemma controls the shift coherence of \( a \):

**Lemma K.5** (Coherence of \( a \) near \( a_0 \)). Suppose that \( a_0 \) is \( \tilde{\mu} \)-shift coherent, and \( \|a - a_0\|_2 \leq \tilde{\mu} \). Then

\[ \| \text{off} [C_a^* C_{a_0}] \|_\infty \leq 2\tilde{\mu} \tag{K.101} \]

\[ \| \text{off} [C_a^* C_a] \|_\infty \leq 3\tilde{\mu} \tag{K.102} \]

**Proof.** Notice that for any \( \ell \neq 0 \), \( |\langle a, s_\ell[a_0] \rangle| \leq |\langle a_0, s_\ell[a_0] \rangle| + |\langle a - a_0, s_\ell[a_0] \rangle| \leq \tilde{\mu} + \|a - a_0\|_2 \leq 2\tilde{\mu} \).

Similarly, \( |\langle a, s_\ell[a] \rangle| \leq |\langle a - a_0, s_\ell[a_0] \rangle| + |\langle a, s_\ell[a_0] \rangle| \leq \|a - a_0\|_2 + 2\tilde{\mu} \leq 3\tilde{\mu} \), as claimed.

\[ \square \]

From this we obtain the following spectral control on \( C_a^* C_a \), to simply the notations, we will write

\[ C_{a_0}^* C_{a_0} = \nu_a^2 C_{a_0}^* C_{a_0}, C_a = [C_a^* C_a]_{1,1} \tag{K.103} \]

in the latter part of this section.
Lemma K.6 (Off-diagonals of $[C_a^* C_a]_{I,I}$). Suppose that $a_0$ is $\bar{\mu}$-shift coherent and $\|a - a_0\|_2 \leq \bar{\mu}$. Then

$$\left\|C_a^* C_a - I\right\|_{I,I} \leq 9 \kappa_I \bar{\mu}. \quad (K.104)$$

We prove this lemma by noting that $C_a^* C_a = C_{r_{a,a}}$ is the convolution matrix associated with the autocorrelation $r_{a,a}$ of $a$. Since $\text{supp}(r_{a,a}) \subseteq \{-p+1, \ldots, p-1\}$ is confined to a (cyclic) stripe of width $2p - 1$, we can tightly control the norm of this matrix by dividing it into three block-diagonal submatrices with blocks of size $p \times p$. Formally:

**Proof.** Divide $I$ into $r = \lceil n/p \rceil$ subsets $I_0, \ldots, I_{r-1}$ such that for all $\ell = 0, \ldots, r-1$:

$$I_\ell = I \cap \{p\ell, p\ell + 1, \ldots, p\ell + (p-1)\} = I \cap ([p] + p\ell).$$

Notice that for each $\ell$:

$$\text{supp}((C_a^* C_a)_{I,\ell}) \subseteq I_\ell \times (I_{\ell-1} \cup I_\ell \cup I_{\ell+1}),$$

where $\ell + 1$ and $\ell - 1$ are interpreted cyclically modulo $r$.

For an arbitrary $v \in \mathbb{R}^{|I|}$, we calculate

$$\left\|C_a^* C_a - I\right\|_{I_\ell,\ell} v = \sum_{\ell=0}^{r-1} \left\|C_a^* C_a - I\right\|_{I_\ell,\ell} v$$

$$= \sum_{\ell=0}^{r-1} \left\|C_a^* C_a - I\right\|_{I_\ell,\ell} w_{I_\ell-1} \otimes I_\ell \otimes I_{\ell+1}$$

$$\leq \sum_{\ell=0}^{r-1} \left\|C_a^* C_a - I\right\|_{I_\ell,\ell} w_{I_\ell-1} \otimes I_\ell \otimes I_{\ell+1} \left\|v_{I_\ell-1} \otimes I_\ell \otimes I_{\ell+1}\right\|_2$$

$$\leq 3 \kappa_I^2 \times (3 \bar{\mu})^2 \times \sum_{\ell=0}^{r-1} \left\|v_{I_\ell-1} \otimes I_\ell \otimes I_{\ell+1}\right\|_2$$

$$\leq 3 \kappa_I^2 \times 9 \bar{\mu}^2 \times 3 \|v\|_2^2,$$

giving the claimed result.

As a consequence, we have that

**Corollary K.7** (Inverse of $[C_a^* C_a]_{I,I}$). Suppose that $a_0$ is $\mu$-shift coherent, that $\|a - a_0\|_2 \leq \bar{\mu}$ and that $\kappa_I \bar{\mu} < \frac{1}{18}$. Then for every $J \subseteq I$ and any norm $\|\cdot\|_\diamond \in \{\|\cdot\|_{\square \rightarrow \square}, \|\cdot\|_{\infty \rightarrow \infty}, \|\cdot\|_2\}$, we have

$$\left\|C_a^* C_a - I\right\|_J \leq 9 \kappa_I \bar{\mu} \quad (K.110)$$

$$\left\|C_a^* C_a \right\|_J^{-1} - I \leq 18 \kappa_I \bar{\mu} \quad (K.111)$$

$$\left\|C_a^* C_a \right\|_J^{1/2} \leq 2. \quad (K.112)$$

**Proof.** First we prove

$$\left\|C_a^* C_a - I\right\|_{J,J} \leq 9 \kappa_I \bar{\mu}, \quad (K.113)$$

$$\left\|C_a^* C_a - I\right\|_{\infty \rightarrow \infty} \leq 6 \kappa_I \bar{\mu}, \quad (K.114)$$

$$\left\|C_a^* C_a - I\right\|_{\square \rightarrow \square} \leq 6 \kappa_I \bar{\mu}. \quad (K.115)$$
Where the first claim follows from Lemma K.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 $2\kappa_I\ell$ entries, of size at most $3\bar{\mu}$. The last follows by noting that

$$\|C^*_a C_a - I\|_{J,J} \leq \max_{\ell,\ell'} \|C^*_a C_a - I\|_{J \cap \{p|\ell|+\ell\}, J \cap \{2p|\ell|+\ell\}} \leq 6\kappa_I\bar{\mu}.$$  \hfill (K.116)

Then we prove

$$\|C^*_a C_a\|_{J,J}^{-1} - I \leq 18\kappa_I\bar{\mu},$$
$$\|C^*_a C_a\|_{J,J}^{-1} - I \leq 12\kappa_I\bar{\mu},$$
$$\|C^*_a C_a\|_{J,J}^{-1} - I \leq 12\kappa_I\bar{\mu},$$

which are followed from the fact that if $\|:\|$ is a matrix norm and $\|\Delta\|_\phi < 1$, then

$$\|(I + \Delta)^{-1} - I\|_\phi \leq \frac{\|\Delta\|_\phi}{1 - \|\Delta\|_\phi}.$$  \hfill (K.117)

Finally, (K.116) follows from the triangle inequality.

Also, we need to bound the convolution of $a_0 - a$ with $\|a_0 - a\|_2$ requiring for bounds of the lasso solution:

**Lemma K.8 (Convolution of $a_0 - a$).** Suppose that $a_0$ is $\mu$-shift coherent and $\|a - a_0\|_2 \leq \bar{\mu}$, then for every $J \subseteq I$,

$$\|C^* a C_{a_0-a}\|_{J,J} \leq \sqrt{2\kappa_I} \|a - a_0\|_2 \leq \sqrt{2\kappa_I} \|a - a_0\|_2$$ \hfill (K.118)

$$\|C^* a C_{a_0-a}\|_{J,J} \leq \sqrt{2\kappa_I} \|a - a_0\|_2$$ \hfill (K.119)

**Proof.** For the first inequality, we have

$$\|C^* a C_{a_0-a}\|_{J,J} = \max_{\ell,\ell'} \|s_{\ell}^* a_{\ell} - s_{\ell} a_{\ell} - v\|_1 \leq \max_{\ell,\ell'} \|s_{\ell}^* (a_{\ell} - a) * v\|_1 \leq \|a - a_0\|_2 \leq \sqrt{2\kappa_I} \|a - a_0\|_2$$ \hfill (K.120)

The second inequality is derived by

$$\|C^* a C_{a_0-a}\|_{J,J} \leq \max_{\ell,\ell'} \|C^* a C_{a_0-a}\|_{J \cap \{p|\ell|+\ell\}, J \cap \{2p|\ell|+\ell\}} \leq \sqrt{2\kappa_I^2} \max_{i,j} \|s_i a - s_j a_{\ell} - a\|_1 \leq \sqrt{2\kappa_I} \|a - a_0\|_2,$$ \hfill (K.121)

finishing the proof.

Again, using a variant of the argument for Lemma K.6, we have the following:

**Lemma K.9 (Off-diagonal of submatrix of $C^* a C_{a_0}$).** Suppose that $a_0$ is $\mu$-shift coherent and $\|a - a_0\|_2 \leq \bar{\mu}$. For any $J \subseteq I$, if

$$\kappa_J = \max_{\ell} |J \cap \{\ell, \ell + 1, \ldots, \ell + p - 1\}|$$ \hfill (K.122)
\[
\kappa_{I \setminus J} = \max_\ell |(I \setminus J) \cap \{\ell, \ell + 1, \ldots, \ell + p - 1\}|
\]

Then
\[
\left\| [C_a^* C_{a_0}]_{J, I \setminus J} \right\|_2 \leq 6\sqrt{\kappa_{I \setminus J} \bar{\mu}}.
\]

**Proof.** Take \( r = \lceil n/p \rceil \) and for \( \ell = 0, \ldots, r - 1 \), write
\[
J_\ell = J \cap ([p] + p\ell), \quad L_\ell = (I \setminus J) \cap ([p] + p\ell),
\]

Take \( v \in \mathbb{R}^{[I \setminus J]} \) arbitrary and notice that
\[
\left\| [C_a^* C_{a_0}]_{J, I \setminus J} v \right\|_2^2 = \sum_{\ell=0}^{r-1} \left\| [C_a^* C_{a_0}]_{J_\ell, L_{\ell-1} \cup L_\ell \cup L_{\ell+1}} v_{L_{\ell-1} \cup L_\ell \cup L_{\ell+1}} \right\|_2^2
\]
\[
\leq 4\bar{\mu}^2 \times \kappa_{J} \times 3\kappa_{I \setminus J} \times \sum_{\ell=0}^{r-1} \left\| v_{L_{\ell-1} \cup L_\ell \cup L_{\ell+1}} \right\|_2^2
\]
\[
\leq 4\bar{\mu}^2 \times \kappa_{J} \times 3\kappa_{I \setminus J} \times 3\|v\|_2^2,
\]

(125)
giving the result. \( \blacksquare \)

**Lemma K.10** (Perturbation of vector over sphere). If both \( a, a_0 \) are unit vectors in inner product space, then
\[
|\langle a, a - a_0 \rangle| \leq \frac{1}{2} \|a - a_0\|_2^2.
\]

**Proof.** Via simple norm inequalities:
\[
\frac{1}{2} \|a - a_0\|_2^2 = 1 - \langle a, a_0 \rangle = 1 - \langle a, a_0 - a + a \rangle = \langle a, a - a_0 \rangle > 0
\]

(127)

**Lemma K.11** (Convolution of short and sparse). Suppose \( \delta \in \mathbb{R}^p \) and \( v \in \mathbb{R}^n \) where \( \text{supp}(v) = I \) satisfies
\[
\max_{\ell \in [n]} |I \cap ([p] + \ell)| \leq \kappa
\]

(128)

then
\[
\|\delta \ast v\|_2 \leq \sqrt{2\kappa} \|\delta\|_2 \|v\|_2
\]

(129)

**Proof.** Since every \( p \)-contiguous segment of \( I \) has at most \( \kappa \) elements, by splitting \( I = I_1 \cup I_2 \cup \ldots \cup I_\kappa \cup R \) such that each sets \( I_i \) are \( p \)-separated:
\[
I_1 = \{i_1, i_{\kappa + 1}, i_{2\kappa + 1}, \ldots\} \cap \{0, \ldots, n - p - 1\},
\]
\[
I_2 = \{i_2, i_{\kappa + 2}, i_{2\kappa + 2}, \ldots\} \cap \{0, \ldots, n - p - 1\},
\]
\[
\vdots
\]
\[
I_\kappa = \{i_\kappa, i_{2\kappa}, i_{3\kappa}, \ldots\} \cap \{0, \ldots, n - p - 1\},
\]
\[
R = I \cap \{n - p, \ldots, n - 1\}.
\]

(130)
Then the p-separating property gives \( \| \delta \ast P_i v \|_2 = \| \delta \|_2 \cdot \| P_i v \|_2 \). Hence:

\[
\| \delta \ast P_i v \|_2 = \left\| \sum_{i \in \kappa} \delta \ast P_i v + \delta \ast P_R v \right\|_2 \leq \sum_{i \in \kappa} \| \delta \ast P_i v \|_2 + \| \delta \ast P_R v \|
\]

\[
= \| \delta \|_2 \sum_{i \in \kappa} \| v_i \|_2 + \| \delta \|_2 \| P_R v \|_1
\]

\[
\leq \sqrt{\kappa} \| v_{i_1, i_2, \ldots, i_{L_v}} \|_2 \| \delta \|_2 + \sqrt{\kappa} \| v_R \|_2 \| \delta \|_2
\]

\[
\leq \sqrt{2\kappa} \| v \|_2 \| \delta \|_2 ,
\]

where the last two inequalities were coming from Cauchy-Schwartz.

\[\tag{K.132}\]

\section{Finite sample inequalities approximation}

In this section we collect several major components of proof about large sample deviation. In particular, the concentration for shift space gradient \( \chi(\beta)_i \), shift space Hessian diagonals \( \| P_{1(a)} \delta_{-i} | x_0 \|_2 \), and the set of gradients discontinuity entries \( | J_{1(b)}(a) | \).

\subsection{Proof of Corollary F.4}

\textbf{Proof.} 1. (\( \epsilon \)-net) Write \( x \) as \( x_0 \) and \( \| \beta \|_2 = \eta \) through out this proof, firstly from Definition E.1 for every \( a \in \cup_{i \leq k} B(S_{\beta}, \gamma(c_{\mu})) \), we know \( \eta \leq 1 + c_{\mu} \sqrt{\frac{c_{\mu}}{\sqrt{\theta \log \theta^{-1}}} \leq \sqrt{\beta}} \). Define \( \epsilon = \frac{c_{\mu}^2}{2n^{3/2}p^{-1/2}} \) and consider the \( \epsilon \)-net \( N_{\epsilon} \) for sphere of radius \( \eta \). From Lemma N.5 we know for any \( c_2 < 1 \):

\[
| N_{\epsilon} | \leq \left( \frac{3\eta}{\epsilon} \right)^{2p} \leq \left( \frac{3n^{3/2}p^2}{c_2} \right)^{2p} \leq \left( \frac{3n^2}{c_2} \right)^{3p}
\]  \[\tag{L.1}\]

for each \( i \in [n] \) define such net as \( N_{\epsilon,i} \), and define an event such that all center of subsets in \( N_{\epsilon,i} \) are being well-behaved:

\[
E_{\text{Net}} := \left\{ \forall i \in [n], \quad \sigma, n^{-1} \chi(\beta)_i - \sigma, n^{-1} E\chi(\beta)_i \leq \frac{c_{\mu} \theta}{p^{3/2}} \quad \forall \beta \in N_{\epsilon,i} \right\} \]  \[\tag{L.2}\]

2. (Lipschitz constant) The Lipschitz constant \( L \) of \( \chi(\cdot)_i \), w.r.t \( \beta \) is bounded in terms of \( \epsilon \) regardless of entry \( i \):

\[
| \chi(\beta)_i - \chi(\beta')_i | \leq \left\| e_i^* \tilde{C}_x S_{\lambda} \left[ \tilde{C}_x \beta - \tilde{C}_x \beta' \right] \right\|_2
\]

\[
\leq \| x \|_2 \left\| S_{\lambda} \left[ \tilde{C}_x \beta - \tilde{C}_x \beta' \right] \right\|_2
\]

\[
\leq \| x \|_2 \left\| \sum_{j \in [n]} S_{\lambda} \left[ \tilde{C}_x \beta \right]_j - S_{\lambda} \left[ \tilde{C}_x \beta' \right]_j \right\|^2
\]

\[
\leq \| x \|_2 \cdot \| x \|_1 \cdot \| \beta - \beta' \|_2 =: L \cdot \| \beta - \beta' \|_2
\]  \[\tag{L.3}\]

Define the event that \( \chi(\beta)_i \) that has small Lipschitz constant as

\[
E_{\text{Lip}} := \left\{ L < 2n^{3/2} \theta \right\}
\]  \[\tag{L.4}\]

72
on the event $\mathcal{E}_{\text{Lip}}$, for every points in $\mathcal{R}(\mathcal{S}_T, \gamma(c_\mu))$ and $i \in [n]$, there exists some $\beta_\epsilon \in \mathcal{N}_{\epsilon,i}$ such that

$$\left| \sigma_i n^{-1} \chi[\beta]_i - \sigma_i n^{-1} \mathbb{E} \chi[\beta]_i \right| - \left( \sigma_i n^{-1} \chi[\beta]_i - \sigma_i n^{-1} \mathbb{E} \chi[\beta]_i \right) \leq 2L \epsilon \leq \frac{c_2 \theta}{p^{3/2}} \quad \text{(L.5)}$$

On event $\mathcal{E}_{\text{Lip}} \cap \mathcal{E}_{\text{Net}}$, (L.2), (L.5) implies $\chi[\beta]$ is well concentrated entrywise and anywhere in $\cup_{|\tau| \leq k} \mathcal{R}(\mathcal{S}_T, \gamma(c_\mu))$:

$$\left| \sigma_i n^{-1} \chi[\beta]_i - \sigma_i n^{-1} \mathbb{E} \chi[\beta]_i \right| \leq \frac{(c_1 + c_2) \theta}{p^{3/2}}, \quad \forall a \in \cup_{k \leq k} \mathcal{R}(\mathcal{S}_T, \gamma(c_\mu)), \, \forall i \in [n] \quad \text{(L.6)}$$

as desired, where, using Lemma D.2,

$$\mathbb{P} \left[ \mathcal{E}_{\text{Lip}} \right] \leq \mathbb{P} \left[ \|x\|_2 > 2n \theta \right] \leq 1/n; \quad \text{(L.7)}$$

and using union bound,

$$\mathbb{P} \left[ \mathcal{E}_{\text{Net}} \right] \leq \mathbb{P} \left[ \max_{a \in \mathcal{N}_{\epsilon,i}} \sigma_i n^{-1} \chi[\beta]_i - \sigma_i n^{-1} \mathbb{E} \chi[\beta]_i > \frac{c_1 \theta}{p^{3/2}} \right]$$

$$\leq n |\mathcal{N}_\epsilon| \mathbb{P} \left[ \sigma_0 n^{-1} \chi[\beta]_0 - \sigma_0 n^{-1} \mathbb{E} \chi[\beta]_0 > \frac{c_1 \theta}{p^{3/2}} \right]. \quad \text{(L.8)}$$

3. (Bound $\mathbb{P} \left[ \mathcal{E}_{\text{Net}} \right]$) Wlog write $n = t \cdot (2p)$ for some integer $t$ and $2p \geq 4p_0 - 3$ and replace $x_0$ with $x$. Observe that $Z_j(\beta)$ from (E.9) is independent of $Z_{j+2p}(\beta)$ for all $j \in [n]$ while all $Z_j$ are identical distributed. We write $\chi[\beta]_0$ as sum of iid r.v.s as

$$\chi[\beta]_0 = \sum_{j \in [n]} Z_j(\beta) = \sum_{k \in [2p]} \left( \sum_{l=0}^{n/2p-1} Z_{k+2lp}(\beta) \right)$$

wlog let $\sigma_0 = 1$ and split the independent r.v.s, write $\mathbb{E} Z_0 = \mathbb{E} Z$, bound the tail probability of $\chi[\beta]_0$ as

$$\mathbb{P} \left[ n^{-1} \chi[\beta]_0 > n^{-1} \mathbb{E} \chi[\beta]_0 + \frac{c_1 \theta}{p^{3/2}} \right] \leq 2p \cdot \mathbb{P} \left[ \frac{n}{2p} Z_{2tp}(\beta) > \frac{n}{2p} \mathbb{E} Z(\beta) + \frac{c_1 \theta}{2p^{3/2}} \right]. \quad \text{(L.9)}$$

The moments of $Z_0$ can be bounded by using $|Z_0(\beta)| \leq |x_0| |\beta_0 x_0 + s_0| \leq \beta_0 x_0^2 + |x_0| |s_0|$ where $s_0 = \sum_{l \neq 0} x_l \beta_l$, write $x = \omega \circ g \sim_{\text{i.i.d.}} \text{BG}(\theta)$. For the 2-norm we know

$$\mathbb{E} |s_0|^2 = \mathbb{E} \left| \sum_{l} x_l \beta_l \right|^2 \leq \theta \|\beta\|_2^2 \leq \theta \left( 1 + c_\mu \frac{\epsilon \mu}{\theta k^2} \right) \leq \frac{1}{2} \quad \text{(L.10)}$$

As for the $q$-norm, use the moment generating function bound, such that for all $t \geq 0$:

$$\mathbb{E} |s_0|^q \leq q! t^{-q} \mathbb{E} \exp \left[ t |s_0| \right] \leq q! t^{-q} \prod_{l} \mathbb{E}_{\omega_l, g_l} \exp \left[ t \omega_l |g_l| |\beta_l| \right]$$

$$\leq 2q! t^{-q} \prod_{l} \mathbb{E}_{\omega_l} \exp \left[ t \omega_l^2 \beta_l^2 / 2 \right]$$

$$\leq 2q! t^{-q} \prod_{l} \left( 1 - \theta + \theta \exp \left[ t^2 \beta_l^2 / 2 \right] \right) \quad \text{(L.11)}$$
notice that the entrywise twice derivative of (L.11) w.r.t. \( \beta^2 \)'s are always positive, this function is convex for all \( \beta^2 \). Constrain on the polytope \( \sum_i \beta^2_i \leq \|\beta\|_2^2 \), the maximizer of (L.11) w.r.t. \( \beta^2 \)'s occurs and a vertex point where \( \beta^2_0 = \|\beta\|_2^2 \). Thus

\[
(L.11) \leq 2q!t^{-q} \left(1 - \theta + \theta \exp \left(\frac{t^2 \|\beta\|_2^2}{2}\right)\right) \prod_{\ell \neq 0} (1 - \theta + \theta e^0) \leq 2q!t^{-q}(1 + \theta \exp[\|\beta\|_2^2 t^2 / 2]).
\]

Choose \( t = \sqrt{q / \|\beta\|_2^2} \), use \( q!! > (q/2)^{q/2} \cdot (e/q)^{q/2} \), we have

\[
E|s_0|^{q} \leq 2q!q^{-q/2} \|\beta\|_2^q (1 + \theta \exp \lceil q/2 \rceil) \leq 8 \|\beta\|_2^q \max \left\{e^{-q/2}, \theta\right\} q!!.
\] (L.12)

Apply Jensen’s inequality \( \left(\sum_{i=1}^{N} z_i\right)^q \leq N^{q-1} \sum_{i=1}^{N} z_i^q \), use Gaussian moment Lemma N.2, (L.10) and (L.12), obtain for \( q \geq 3 \),

\[
E Z(\beta)^q \leq E \left(\|\beta\|_2^q \right)^q \leq 2e \left(\|\beta\|_2^q \right)^q \leq 2e \left(\|\beta\|_2^q \right)^q \leq 2e \left(\|\beta\|_2^q \right)^q \leq \theta q!!.
\]

Thus, recall that \( \|\beta\|_2 = \eta \), use \( (\sigma^2, R) = (8\theta \eta^2, 4\eta) \), from (L.8)-(L.9), apply Bernstein inequality Lemma N.4 with \( n \geq C_\theta^5 n^{-2} \log p \), and \( c_1, c_2 \in [0, 1] \) we have

\[
\mathbb{P}\left[ \mathcal{E}^c_{\text{Net}} \right] \leq 2np |\mathcal{N}_c| \cdot \mathbb{P}\left[ \sum_{t=0}^{n/2p-1} Z_{2t p}(\beta) > \frac{n}{2p} E Z(\beta) + \frac{c_1 n \theta}{2 p^5/2} \right]
\]

\[
\leq 2np \left( \frac{3n p^2}{c_2} \right)^{3p} \exp \left( \frac{- (c_1 n \theta / 2 p^5/2)^2}{16 n \theta p^2 / 2p + 8n \theta c_1 n \theta / 2 p^5/2} \right)
\]

\[
\leq \exp \left( 4p \log \left( \frac{3n p^2}{c_2} \right) - \frac{(c_1 n \theta / 2 p^5/2)^2}{16 n \theta p^2 / p} \right)
\]

\[
\leq \exp \left( 4p \log \left( \frac{3n p^2}{c_2} \right) - \frac{c_1^2 n \theta^2}{64 n p^4} \right)
\]

\[
\leq \exp \left( \frac{-c_1^2 n \theta^2}{100 n p^4} \right) \leq \frac{1}{n}.
\] (L.13)

When \( \frac{c_1}{C_\log C} > \frac{10^5}{c_1^3 c_2^3} \). The proof of lower bound and negative \( \beta_0 \) is derived in the same manner.

\[\Box\]

### L.2 Proof of Corollary G.3

**Proof.** Write \( x \) as \( x_0 \) though out this proof. Write \( \beta, x_j + s_j = \sum_{t \in [\pm 1]} \beta_t x_{t-i} + \langle \beta, x_{[\pm 1]i} \rangle \), and the support w.r.t. some \( a \) as \( I(\beta) \). Define the random variable \( Z_{ij}(\beta) \) as

\[
\|P(I(\beta)) s_{-i}[x]\|_2^2 = \sum_{j \in [n]} x_j^2 1_{\{\langle \beta, x_{[\pm 1]i} \rangle > \lambda\}} =: \sum_{j \in [n]} Z_{ij}(\beta)
\] (L.14)

and define \( \{Z_{ij}(\beta)\}_{j \in [n]} \) that are independent r.v.s. and as an upper bounding function of \( Z_{ij}(\beta) \) as

\[
\bar{Z}_{ij}(\beta) := \begin{cases} x_j^2, & \langle \beta, x_{[\pm 1]i} \rangle > \lambda \\ 0, & \langle \beta, x_{[\pm 1]i} \rangle < \lambda/2, \\ \frac{x_j^2}{\lambda/2} (\langle \beta, x_{[\pm 1]i} \rangle - \lambda/2), & \text{otherwise} \end{cases}
\] (L.15)
Similar to proof of Corollary F.4. Let \( \| \beta \|_2 \leq \eta \leq \sqrt{\theta} \). Define \( \varepsilon = \frac{c_2^\lambda}{24np\sqrt{p\theta \log n \log \theta^{-1}}} \) for some \( c_2 > 0 \) and consider the \( \varepsilon \)-net \( \mathcal{N}_\varepsilon \) for sphere of radius \( \eta \). From Lemma N.5 we know
\[
|\mathcal{N}_\varepsilon| \leq \left( \frac{3\eta}{\varepsilon} \right)^{2p} \leq \left( \frac{72}{c_2^\lambda} np^2 \sqrt{\theta |\tau| \log n \log \theta^{-1}} \right)^{2p} \leq \left( \frac{72}{c_2^\lambda} np^2 \log n \right)^{2p},
\]
(L.16)
for each \( i \in [n] \) define such net as \( \mathcal{N}_{\varepsilon,i} \), and define an event such that all center of subsets in \( \mathcal{N}_{\varepsilon,i} \) are being well-behaved:
\[
\mathcal{E}_{\text{Net}} := \left\{ \forall i \in [n], \left| n^{-1} \sum_{j \in [n]} Z_{ij}(\beta_\varepsilon) - \mathbb{E} Z_i(\beta_\varepsilon) \right| \leq \frac{c_1^\theta}{p} \forall \beta_\varepsilon \in \mathcal{N}_{\varepsilon,i} \right\},
\]
(L.17)
Also, \( \sum_j Z_{ij}(\beta) \) is a Lipschitz function over \( \beta \) for every \( i \in [n] \) as
\[
\left| \sum_{j \in [n]} Z_{ij}(\beta) - \sum_{j \in [n]} Z_{ij}(\beta') \right| \leq \sum_{j \in [n]} \frac{x_j^2}{\lambda/2} \left| \langle \beta - \beta', x_{[-p]-i+j} \rangle \right|
\leq \sum_{j \in [n]} \frac{x_j^2}{\lambda/2} \| x_{[-p]-i+j} \|_2 \| \beta - \beta' \|_2
\leq \frac{1}{\lambda/2} \| x \|_2 \cdot \max_{j \in [n]} \| x_{[-p]+j} \|_2 \| \beta - \beta' \|_2
:= L \| \beta - \beta' \|_2,
\]
(L.18)
and define event \( \mathcal{E}_{\text{Lip}} \) such that the Lipschitz constant is bounded as
\[
\mathcal{E}_{\text{Lip}} := \left\{ L \leq 12n\theta \sqrt{p\theta \log n \log \theta^{-1} \lambda^{-1}} \right\},
\]
(L.19)
then on event \( \mathcal{E}_{\text{Lip}} \), for any points \( \beta \in \mathbb{R}(S_\tau, \gamma(c_n)) \) and \( i \in [n] \), there exists some \( \beta_\varepsilon \) in \( \mathcal{N}_{\varepsilon,i} \) with \( \| \beta - \beta_\varepsilon \|_2 \leq \varepsilon \), and thus
\[
\left| n^{-1} \sum_{j \in [n]} Z_{ij}(\beta) - \mathbb{E} Z_i(\beta) \right| - \left| n^{-1} \sum_{j \in [n]} Z_{ij}(\beta_\varepsilon) - \mathbb{E} Z_i(\beta_\varepsilon) \right| \leq 2L\varepsilon \leq \frac{c_1^\theta}{p},
\]
(L.20)
On event \( \mathcal{E}_{\text{Lip}} \cap \mathcal{E}_{\text{Net}} \), from (L.17), (L.20), we can conclude that for all \( \beta \in \mathbb{R}(S_\tau, \gamma(c_n)) \) and \( i \in [n] \) that:
\[
n^{-1} \left\| P_{l(\beta)s-1} [x_0] \right\|_2^2 - n^{-1} \mathbb{E} \left\| P_{l(\beta)s-1} [x_0] \right\|_2^2 \leq n^{-1} \sum_{j \in [n]} Z_{ij}(\beta) - \mathbb{E} Z_i(\beta)
\leq \frac{(c_1^\lambda + c_2^\lambda)^2 \theta}{p},
\]
(L.21)
as desired, where the error probability of \( \mathcal{E}_{\text{Lip}} \) is bounded using Lemma D.2 and Lemma D.3, which give
\[
P \left[ \mathcal{E}_{\text{Lip}}^c \right] \leq P \left[ \| x \|_2^2 > 2n\theta \right] + P \left[ \max_{j \in [n]} \| x_{[-p]+j} \|_2 > 3\sqrt{p\theta \log n \log \theta^{-1}} \right]
\leq 3/n,
\]
(L.22)
when \( n > 10^3 \theta^{-1} \). As for \( \mathcal{E}_{\text{Net}} \) use union bound and split the r.v.s since \( Z_j, Z_{j+2p} \) are independent for all \( j \):
\[
P \left[ \mathcal{E}_{\text{Net}}^c \right] \leq 2np \cdot |\mathcal{N}_\varepsilon| \cdot P \left[ \sum_{k} \frac{n}{2p} Z_{1,2kj}(\beta) - \frac{n}{2p} \mathbb{E} Z_i(\beta) \right] \geq \frac{c_1^\lambda \theta}{2p^2}.
\]
Now we calculate the variance and $L^q$-norm of $\sum_k Z_{1,2k_j}$ for $q \geq 3$:

\[
\left\{ \begin{array}{l}
\mathbb{E}Z_{i,j}^2 \leq \mathbb{E}x_i^4 \leq 3\theta \\
\mathbb{E}Z_{i,j}^4 \leq \mathbb{E}x_i^{2q} \leq (2q - 1)!! \leq \frac{1}{2} \cdot (3\theta) \cdot 2^{q-2}q!
\end{array} \right.
\]

(L.23)

and apply Bernstein inequality with $(\sigma^2, R) = (3\theta, 2)$, then use $n \geq C p^4 \theta^{-1} \log p$ and $c_1', c_2' < 1$ to obtain

\[
2np |\mathcal{N}_i| \mathbb{P} \left[ \sum_k Z_{1,2k_j}(\beta) - \frac{n}{2p^2} \mathbb{E}Z_i \right] \geq c_1' n \theta \frac{2}{2p^2}
\]

\[
\leq \exp \left[ \log(2np) + 2p \log \left( \frac{72}{c_2' c_\lambda} n \theta^2 \log n \right) - \frac{(c_1' n \theta^2 / 2p^2)^2}{6n \theta / 2p + 4c_1' n \theta \theta s / 2p^2} \right]
\]

\[
\leq \exp \left[ 3p \log \left( \frac{72}{c_2' c_\lambda} n \theta^2 \log n \right) - \frac{c_1'^2 n \theta}{24p^3} \right]
\]

\[
\leq \exp[ -c_1'^2 n \theta / (50p^3) ] \leq 1/n,
\]

(L.24)

where the last two inequalities holds when $\frac{C}{\log c} \geq \frac{105}{c_1'^2 c_2' c_\lambda}$. The other side of inequality of (G.9) can be derived by defining $Z_{ij}$ as

\[
Z_{ij}(\beta) := \begin{cases} 
    x_i^2, & |\langle \beta, x_{[\pm p - i + j]} \rangle| > 3\lambda / 2 \\
    0, & |\langle \beta, x_{[\pm p - i + j]} \rangle| \leq \lambda \\
    x_i^2 / n \lambda^2 \left( |\langle \beta, x_{[\pm p - i + j]} \rangle| - \lambda \right), & \text{otherwise}
\end{cases}
\]

(L.25)

and define $\mathcal{E}_{\text{Net}}, \mathcal{E}_{\text{Lip}}$ similarly, such that on intersection of these events,

\[
n^{-1} \left\| P_{i(\beta)} s - i [x] \right\|_2^2 - n^{-1} \mathbb{E} \left\| P_{i(\beta)} s - i [x] \right\|_2^2 \geq n^{-1} \sum_{j \in [n]} Z_{ij}(\beta) - \mathbb{E}Z_i(\beta)
\]

\[
\geq \frac{(c_1' + c_2') \theta}{p}
\]

(L.26)

as desired.

\[\blacksquare\]

### L.3 Proof of Lemma H.5

**Proof.** 1. (Expectation upper bound) We will write $x$ as $x_0$. Similar to proof of Corollary F.4 let $\|\beta\|_2 \leq \eta \leq \sqrt{p}$. For each $i \in [n]$, define the random variable

\[
X_i(\beta) = 1_{\{ (|s_i(x)|, \beta) - \lambda \leq B \}} + 1_{\{ (|s_i(x)|, \beta) + \lambda \leq B \}},
\]

(L.27)

then number of indices for vector $x + \beta$ that are within $B$ of $\pm \lambda$ is a random variable $\sum_{i \in [n]} X_i(\beta)$. For each of the $X_i(\beta)$'s consider an upper bound $X_i(\beta)$ defined as

\[
X_i(\beta) = \begin{cases} 
    1, & (|s_i(x)|, \beta) \in [\lambda - B - M, \lambda - B] \\
    (\lambda + B + M) - |s_i(x)|, & (|s_i(x)|, \beta) \in [\lambda + B, \lambda + B + M] \\
    0, & \text{else}
\end{cases}
\]

(L.28)

where $B < M = c \lambda \theta^2 / (p \log n) \leq \lambda / 4$ for some constant $0 < c < 1$.  

76
Notice that $x \sim_{\text{i.i.d.}} B[G]$, where $g \sim_{\text{i.i.d.}} N(0, 1)$, and $I(\alpha) \subseteq [n]$ is an independent Bernoulli subset. Conditioned on $I(\alpha)$, $(x, \beta) = \langle g, P_I(\alpha) \beta \rangle \sim N(0, \|P_I(\alpha)\beta\|_2^2)$. For all realizations of $I(\alpha)$, the variance $\|P_I(\alpha)\beta\|_2^2$ is bounded by $\|P_I(\alpha)\beta\|_2^2 \leq \|\beta\|_2^2 \leq p$. 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 $N(0, \sigma^2)$ random variable, the expectation of $\sum_i X_i(\beta)$ can be upper bounded as

$$
\sum_{i \in [n]} \mathbb{E}[X_i(\beta)] \leq (2n) \cdot P[(x, \beta) \in [\lambda - B - M, \lambda + B + M]]
$$

\begin{align}
&\leq (2n) \cdot 2(B + M) \sup_{\sigma^2 \in [0, p]} \max_{t \in [\lambda - B - M, \lambda + B + M]} f_\sigma(t) \\
&\leq 4n(B + M) \sup_{\sigma^2 \in [0, p]} f_\sigma(\lambda - B - M) \\
&\leq 4n(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

$$
\sup_{\sigma^2 \in [0, p]} f_\sigma(\lambda/2) \leq f_{\lambda/2}(\lambda/2) \leq \frac{1}{\sqrt{2\pi}(\lambda/2)} \exp \left(-\frac{1}{2}\right) \leq \frac{1}{2\lambda},
$$

whence, by letting $B \leq c\lambda \theta^2 / (p \log n)$, the upper bound of expectation become:

$$
\sum_{i \in [n]} \mathbb{E}[X_i(\beta)] \leq 4n(B + M) \leq \frac{4cn\theta^2}{p \log n} =: n\mathbb{E}[X(\beta)].
$$

2. (ε-net) Define $\varepsilon = \frac{\lambda \theta\eta^2}{2p + \log n \log \log n - \log p}$. Write $\lambda = c_\lambda / \sqrt{|\tau|}$ and consider the ε-net $N_\varepsilon$ for sphere of radius $\eta \leq \sqrt{p}$. From Lemma N.5 we know

$$
|N_\varepsilon| \leq \left(\frac{3n}{\varepsilon}\right)^{2p} \leq \left(\frac{81|\tau|p^6 \log n \log \theta^{-1}}{c^4 c_\lambda^2 \theta^2}ight)^p \leq \left(\frac{2p \log n}{c \cdot c_\lambda}\right)^{13p}
$$

and define an event such that all center of subsets in $N_\varepsilon$ are being well-behaved:

$$
\mathcal{E}_{\text{Net}} := \left\{ \sum_{i \in [n]} X_i(\beta_\varepsilon) - n\mathbb{E}[X(\beta_\varepsilon)] < \frac{18cn\theta^2}{p \log n} \quad \forall \beta_\varepsilon \in N_\varepsilon \right\}
$$

3. (Lipschitz constant) Furthermore, the function $\sum_i X_i(\beta)$ is Lipchitz over $\beta$ such that

$$
\left| \sum_{i \in [n]} X_i(\beta) - \sum_{i \in [n]} X_i(\beta') \right| \leq \sum_{i \in [n]} \frac{1}{M} |\langle s_i, x \rangle, \beta - \beta'| |
$$

\begin{align}
&\leq \frac{n}{M} \max_{i \in [n]} \left\| P_\beta x \right\|_2 \left\| \beta - \beta' \right\|_2
\end{align}
where each summand has bounded variance and as desired, where the error probability of then on event thus when and apply Bernstein inequality Lemma N.4 with is independent of , for every $\beta$ in $\mathcal{H}(S, \gamma(c))$, there exists some $\beta_\epsilon$ in $\mathcal{N}_\epsilon$ with $\|\beta - \beta_\epsilon\|_2 \leq \epsilon$, thus \[
abla \left( \sum_{i \in [n]} X_i(\beta) - n\mathbb{E} X(\beta) \right) \leq 2L\epsilon \leq \frac{2cn\theta^2}{p \log n}. \] (L.34)

On event $\mathcal{E}_{\text{Net}} \cap \mathcal{E}_{\text{Lip}}$ from (L.31), (L.33) and (L.34), we can conclude that for every $\beta \in \mathcal{H}(S, \gamma(c))$ and $i \in [n],$

\[
\sum_{i \in [n]} X_i(\beta) \leq \frac{24cn\theta^2}{p \log n} \] (L.35)
as desired, where the error probability of $\mathcal{E}_{\text{Lip}}$ is bounded using Lemma D.3, which gives

\[
P(\mathcal{E}_{\text{Lip}}^c) \leq P \left[ \max_{j \in [n]} \|X_{[j]+j}\|_2 > 3\sqrt{p\theta \log n \log \theta^{-1}} \right] \leq 2/n, \] (L.36)

4. (Bound $P(\mathcal{E}_{\text{Net}}^c)$) Wlog let us assume that $2p$ divides $n$. By applying union bound and observing that $X_i(\beta)$ is independent of $X_{i+2p}(\beta)$ for any $i \in [n]$, we split $\sum_i X_i(\beta)$ into $n/2p$ independent sums of r.v.s, we have

\[
P(\mathcal{E}_{\text{Net}}^c) \leq 2p |N_\epsilon| \cdot P \left[ \sum_{j=0}^{n/2p-1} (X_{2pj}(\beta) - \mathbb{E} X(\beta)) > \frac{9cn\theta^2}{p^2 \log n} \right],
\]
where each summand has bounded variance and $L^2$-norm derived similarly as its expectation such that

\[
\mathbb{E} X_i(\beta)^q \leq 2 \cdot P(\{s_i|x|, \beta | \lambda - B - M, \lambda + B + M\} \leq 2 \cdot \frac{1}{2\lambda} \cdot 2(B + M) \leq \frac{4c\theta^2}{p \log n},
\]
and apply Bernstein inequality Lemma N.4 with $(\sigma^2, R) = (4c\theta^2 / (p \log n), 1)$, obtains

\[
P \left[ \sum_{j=0}^{n/2p-1} (X_{2pj}(\beta) - \mathbb{E} X(\beta)) > \frac{9cn\theta^2}{p^2 \log n} \right] \leq \exp \left[ \frac{-(9cn\theta^2/p^2 \log n)^2}{2c\theta^2/p^2 \log n + 2(9cn\theta^2/p^2 \log n)} \right] \leq \exp \left[ \frac{-4c\theta^2}{p^2 \log n} \right],
\]
thus when $n = C p^5 \theta^{-2} \log p$

\[
P(\mathcal{E}_{\text{Net}}^c) \leq \exp \left[ \log(2p) + 13p \log \left( \frac{2p \log n}{c \cdot c_\lambda} \right) - \frac{4c\theta^2}{p^2 \log n} \right] \leq 1/n \] (L.37)
as long as $\frac{c}{\log C} > 10^2 / (c^2 \cdot c_\lambda).$
Algorithm for experiment

In our experiment, we use Algorithm 1 (below), 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 derived from [AMS09]. Here $\text{Exp}_a : \mathbb{A}^+ \rightarrow \mathbb{S}^{p-1}$ takes a tangent vector of $a$ and produces a new point on the sphere, whereas $\text{Log}_a : \mathbb{S}^{p-1} \rightarrow \mathbb{A}^+$ takes a point $b \in \mathbb{S}^{p-1}$ and returns the tangent vector which points from $a$ to $b$.

Algorithm 1 Sâs deconvolution with Accelerated Riemannian gradient descent

Require: Observation $y$, sparsity penalty $\lambda = 0.5 / \sqrt{m_0}$, momentum parameter $\eta \in [0, 1)$.

Initialize $a^{(0)} \leftarrow -P_{\mathbb{S}^{p-1}} \nabla \varphi (P_{\mathbb{S}^{p-1}} \left[0^{p_0-1}; [y_0, \cdots, y_{p_0-1}]; 0^{p_0-1}] \right)$.

for $k = 1, 2, \ldots, K$ do

Get momentum: $w \leftarrow \text{Exp}_{a^{(k)}} (\eta \cdot \text{Log}_{a^{(k-1)}} (a^{(k)}))$.

Get negative gradient direction: $g \leftarrow -\text{grad} \varphi \cdot (w)$.

Armijo step $a^{(k+1)} \leftarrow \text{Exp}_w (tg)$, choosing $t \in (0, 1)$ s.t. $\varphi (a^{(k+1)}) - \varphi (w) < -t \|g\|^2$.

end for

Ensure: Return $a^{(K)}$.

Tools

Lemma N.1 (Tail bound for Gaussian r.v.). If $X \sim \mathcal{N}(0, \sigma^2)$, then its tail bound for $t > 0$ can be

$$\mathbb{P} \left[ X > t \right] \leq \frac{\sigma}{t \sqrt{2\pi}} \exp \left( -\frac{t^2}{2\sigma^2} \right)$$

(N.1)

Lemma N.2 (Moments of the Gaussian random variables). If $X \sim \mathcal{N}(0, \sigma^2)$, then for all integer $p \geq 1$,

$$\mathbb{E} \left[ |X|^p \right] \leq \sigma^p (p-1)!!.$$  

(N.2)

Lemma N.3 (Gaussian concentration inequality). Let $x = (x_1, \ldots, x_n)$ 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 $t > 0$,

$$\mathbb{P} \left[ |f(x) - \mathbb{E}f(x)| \geq t \right] \leq 2 \exp \left( -\frac{t^2}{2L^2} \right).$$  

(N.3)

Lemma N.4 (Moment control Bernstein inequality for scalar r.v.s). ([FR13], Theorem 7.30) Let $x_1, \ldots, x_n$ be independent real-valued random variables. Suppose that there exist some positive number $R$ and $\sigma^2$ such that

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ x_i^2 \right] \leq \sigma^2$$

and

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ |x_i|^p \right] \leq \frac{\sigma^2 R^{p-2}}{2} n! \quad \text{for all integers } p \geq 3.

Let $S \doteq \sum_{i=1}^n x_i$, then for all $t > 0$, it holds that

$$\mathbb{P} \left[ |S - \mathbb{E}[S]| \geq t \right] \leq 2 \exp \left( -\frac{t^2}{2n\sigma^2 + 2Rt} \right).$$  

(N.4)

79
Theorem N.5 (ε-net on sphere). [Ver10] Let \((X, d)\) be a metric space and let \(\varepsilon > 0\). A subset \(N_\varepsilon\) of \(X\) is called an \(\varepsilon\)-net of \(X\) if for every point \(x \in X\) there exists some point \(y \in N_\varepsilon\) so that \(d(x, y) \leq \varepsilon\). There exists an \(\varepsilon\)-net \(N_\varepsilon\) for the sphere \(\mathbb{S}^{n-1}\) of size \(|N_\varepsilon| \leq (3/\varepsilon)^n\).

Lemma N.6 (Hanson-Wright). [RV+13] Let \(x_1, \ldots, x_n\) be independent, subgaussian random variables with subgaussian norm \(\sup_{p \geq 1} p^{-1/2} (\mathbb{E} |x_i|^p)^{1/p} \leq \sigma\). Let \(A \in \mathbb{R}^{n \times n}\), then for every \(t > 0\),

\[
\mathbb{P} \left[ \|x^* A x - \mathbb{E}x^* A x\| \geq t \right] \leq 2 \exp \left(-c \min \left( \frac{t^2}{64 \sigma^4 \|A\|_F^2}, \frac{t^2}{8\sqrt{2} \sigma^2 \|A\|_2} \right) \right).
\]

N.5

Lemma N.7 (Maximum of separable convex function). Let \(f : \mathbb{R}_+ \to \mathbb{R}_+\) be a convex function of the form \(f(x) = x - s(x)\) with \(s : \mathbb{R}_+ \to \mathbb{R}_+\) satisfying

\[
\frac{s(x)}{x} \leq \frac{s(y)}{y}, \text{ for all } x, y > 0.
\]

Then for \(n \in \mathbb{N}\) and \(0 < N \leq nL\),

\[
\max_{0 \leq x \leq L, \|x\|_1 \leq N} \sum_{i=1}^{n} f(x_i) \leq N \left(1 - \frac{s(L)}{L}\right)
\]

N.6

Proof. Since the feasible set is a convex polytope; the convex function \(\sum_{i=1}^{n} f(x_i)\) is maximized at a vertex, and that its vertices consist of 0 and permutations of the vector \([L, \ldots, L, r, 0, \ldots, 0]\), where \(r = \lfloor N/L \rfloor\). If \(N - \lfloor N/L \rfloor L \leq L\). Then the function value at the maximizing vector \(x^*_s\), can be derived as:

\[
\sum_{i=1}^{n} f(x_i) = \lfloor \frac{n}{L} \rfloor f(L) + f(r) = \frac{N-r}{L} (L - s(L)) + (r - s(r))
\]

\[
= N \left(1 - \frac{s(L)}{L}\right) + r \left(\frac{s(L)}{r} - \frac{s(r)}{r}\right) \leq N \left(1 - \frac{s(L)}{L}\right)
\]

References


