The Local Geometry of Orthogonal Dictionary Learning using L1 Minimization

Qiuwei Li*, Zhihui Zhu†, Michael Wakin* and Gongguo Tang*
*Department of Electrical Engineering, Colorado School of Mines. {qiuli,gtang,mwakin}@mines.edu
†Mathematical Institute for Data Science, Johns Hopkins University. zzhu29@jhu.edu

***Abstract*—Feature learning that extracts concise and generalizable representations for data is one of the central problems in machine learning and signal processing. Sparse dictionary learning, also known as sparse coding, distinguishes from other feature learning techniques in sparsity exploitation, allowing the formulation of nonconvex optimizations that simultaneously uncover a structured dictionary and sparse representations. Despite the popularity of dictionary learning in applications, the landscapes of these optimizations that enable effective learning largely remain a mystery. This work characterizes the local optimization geometry for a simplified version of sparse coding where the L1 norm of the sparse coefficient matrix is minimized subject to orthogonal dictionary constraints. In particular, we show that the ground-truth dictionary and coefficient matrix are locally identifiable under the assumption that the coefficient matrix is sufficiently sparse and the number of training data columns is sufficiently large.**

I. Introudction

The name of "dictionary" in the representation theory first appeared in the pioneering work [1] by Stephane Mallat and Zhifeng Zhang in 1993, who introduced the concept of dictionary in contrast to the more traditional concept of transformation such as the discrete Fourier transform, discrete cosine transform, and discrete wavelet transform, etc. In comparing with the concept of dictionary learning, we call these transformations as pre-constructed dictionaries.

We say a signal $\mathbf{y} \in \mathbb{R}^n$ has a sparse representation over a given dictionary $\mathbf{D} \in \mathbb{R}^{n\times n}$ if we can write $\mathbf{y} = \mathbf{D}\mathbf{x}$ with $\|\mathbf{x}\|_0 \ll n$. Here $\|\cdot\|_0$ denotes the ℓ_0 "norm", which counts the number of nonzero entries. In the quest for a proper dictionary, one way is to choose from those pre-constructed dictionaries. While pre-constructed dictionaries usually lead to fast transforms, they are typically limited in their sparsification power. Furthermore, most of these pre-constructed dictionaries are restricted to certain types of signals, and lack adaptivity to practical training datasets [2]. These have led many to adopt a data-driven approach for obtaining dictionaries that overcome these limitations. The resulting dictionary learning method, also known as sparse coding, starts by building a training dataset $\mathbf{Y} = [\mathbf{y}_1 \ \cdots \ \mathbf{y}_L] \in \mathbb{R}^{n\times L}$ consisting of L raw data samples, and then learns a dictionary $\mathbf{D} \in \mathbb{R}^{n\times n}$ with $n \ll L$ that can concisely represent the training dataset, i.e., $\mathbf{Y} \approx \mathbf{D}\mathbf{X}$ and the representation matrix $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_L] \in \mathbb{R}^{n\times L}$ is as sparse as possible. See Figure 1 for a visual illustration of this process.

Notation We denote $[n]$ as the collection of integers from 1 to n. The symbols $\mathbf{I}$ and $\mathbf{0}$ are reserved for the identity matrix and zero matrix/vector, respectively. Denote $\mathcal{O}_n$ as the set of all orthogonal matrices in $\mathbb{R}^{n\times n}$. The gradient of a smooth function $f(\mathbf{X})$ with a matrix variable $\mathbf{X} \in \mathbb{R}^{m\times n}$ is an $m \times n$ matrix, whose (i,j)th entry is $[\nabla f(\mathbf{X})]_{i,j} = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}(i,j)}$ for $i \in [m]$, $j \in [n]$, where $\mathbf{X}(i,j)$ denotes (i,j)th entry of $\mathbf{X}$. For a nonsmooth function $f(\mathbf{X})$ with a matrix variable $\mathbf{X}$, we denote $\partial f(\mathbf{X})$ as the Clarke subdifferential [3], which is a set composed of $m \times n$ matrices, whose (i,j)th position is the partial subdifferential of f with respect to the (i,j)th entry of $\mathbf{X}$. For functions of multi-block variables $f(\mathbf{X},\mathbf{Y})$, we use $\nabla_{\mathbf{X}} f(\mathbf{X},\mathbf{Y})$ to denote the partial gradient of f with respect to $\mathbf{X}$ when f is smooth function of $\mathbf{X}$ and use $\partial_{\mathbf{X}} f(\mathbf{X},\mathbf{Y})$ to denote the partial subdifferential of f with respect to $\mathbf{X}$ when f is a nonsmooth function of $\mathbf{X}$. Similar notations apply to $\nabla_{\mathbf{Y}} f(\mathbf{X},\mathbf{Y})$ and $\partial_{\mathbf{Y}} f(\mathbf{X},\mathbf{Y})$.

II. Problem Formulation

A. Data model and learning algorithm

For theoretical analysis, we assume the dataset $\mathbf{Y} \in \mathbb{R}^{n\times L}$ is generated by a "ground-truth" dictionary $\mathbf{D}_\sharp$ and a "ground-truth" sparse coefficient matrix $\mathbf{X}_\sharp$:

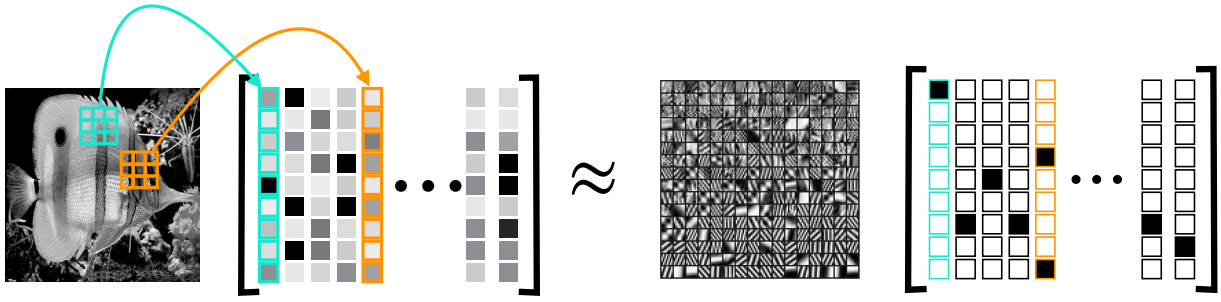
$$\mathbf{Y} = \mathbf{D}_\sharp \mathbf{X}_\sharp. \tag{1}$$

This work considers the following orthogonal dictionary learning problem: *Given the training dataset $\mathbf{Y} \in \mathbb{R}^{n\times L}$ from the data generating model (1), can we recover the ground-truth dictionary $\mathbf{D}_\sharp \in \mathbb{R}^{n\times n}$ and the ground-truth sparse coefficient matrix $\mathbf{X}_\sharp \in \mathbb{R}^{n\times L}$ by solving certain optimization problems?* Orthogonal dictionary learning provides an important special case of general dictionary learning that is more amenable to analysis.

Using the fact that ℓ_1 norm can promote the sparsity of the solutions, we consider the following optimization problem

$$\begin{aligned} \underset{\mathbf{D}\in\mathbb{R}^{n\times n},\mathbf{X}\in\mathbb{R}^{n\times L}}{\text{minimize}} \ & \|\mathbf{X}\|_1, \\ \text{subject to} \ & \mathbf{Y} = \mathbf{D}\mathbf{X}, \\ & \mathbf{D}^\top\mathbf{D} = \mathbf{I}, \end{aligned} \tag{2}$$

where $\|\mathbf{X}\|_1 := \sum_{i,j} |\mathbf{X}(i,j)|$ denotes the ℓ_1 norm of $\mathbf{X}$. Since $\mathbf{D}^\top\mathbf{D} = \mathbf{I}$ implies that $\mathbf{X} = \mathbf{D}^\top\mathbf{Y}$, the optimization



Raw data | Training data from patches $\mathbf{Y} \in \mathbb{R}^{n\times L}$ | Learned dictionary $\mathbf{D} \in \mathbb{R}^{n\times n}$ | Sparse representation $\mathbf{X} \in \mathbb{R}^{n\times L}$

Figure 1. A visual illustration of dictionary learning process.

(2) is equivalent to

$$\begin{aligned} &\underset{\mathbf{D}\in\mathbb{R}^{n\times n}}{\text{minimize}} \ \|\mathbf{D}^\top\mathbf{Y}\|_1, \\ &\text{subject to} \ \ \mathbf{D}^\top\mathbf{D} = \mathbf{I}. \end{aligned} \tag{3}$$

III. Main Results

Using the Karush-Kuhn-Tucker (KKT) conditions of the optimization problems (2) and (3), this work studies under what conditions the ground-truth dictionary and coefficients $\{\mathbf{D}_\sharp, \mathbf{X}_\sharp\}$ are locally identifiable by solving these two optimization problems.

a) KKT conditions for problem (2): We first derive the KKT conditions for the optimization problem (2). Towards that end, we form the Lagrangian of (2) as

$$\mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{\Omega},\mathbf{\Lambda}) = \|\mathbf{X}\|_1 - \langle \mathbf{D}\mathbf{X}-\mathbf{Y}, \mathbf{\Omega}\rangle - \left\langle \mathbf{I}-\mathbf{D}^\top\mathbf{D}, \mathbf{\Lambda}\right\rangle.$$

Then the KKT conditions for (2) is given by [4]

$$\begin{aligned} \mathbf{0} &= \nabla_{\mathbf{\Omega}}\mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{\Omega},\mathbf{\Lambda}), \\ \mathbf{0} &= \nabla_{\mathbf{\Lambda}}\mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{\Omega},\mathbf{\Lambda}), \\ \mathbf{0} &= \nabla_{\mathbf{D}}\mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{\Omega},\mathbf{\Lambda}), \\ \mathbf{0} &\in \partial_{\mathbf{X}}\mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{\Omega},\mathbf{\Lambda}), \end{aligned} \tag{4}$$

where we used the partial subdifferential in the last line since $\mathcal{L}$ is a nonsmooth function of the variable $\mathbf{X}$. Plugging the form of $\mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{\Omega},\mathbf{\Lambda})$ to the KKT conditions (4), we obtain the following equivalent conditions

$$\begin{aligned} \mathbf{D}\mathbf{X} &= \mathbf{Y}, \\ \mathbf{D}^\top\mathbf{D} &= \mathbf{I}, \\ \mathbf{\Omega}\mathbf{X}^\top - \mathbf{D}(\mathbf{\Lambda}+\mathbf{\Lambda}^\top) &= \mathbf{0}, \\ \mathbf{D}^\top\mathbf{\Omega} &\in \operatorname{sign}(\mathbf{X}), \end{aligned} \tag{5}$$

where $\operatorname{sign}(\mathbf{X})$ is an $n \times L$ matrix with entries given by

$$[\operatorname{sign}(\mathbf{X})]_{i,j} = \begin{cases} \operatorname{sign}(x_{i,j}), & x_{i,j} \neq 0 \\ [-1,1], & x_{i,j} = 0. \end{cases}$$

Combine the last two equations in (6) to get

$$\mathbf{\Lambda}+\mathbf{\Lambda}^\top \in \operatorname{sign}(\mathbf{X})\mathbf{X}^\top \iff \mathbf{0} \in \operatorname{sign}(\mathbf{X})\mathbf{X}^\top - \mathbf{X}\operatorname{sign}(\mathbf{X})^\top.$$

Then we obtain a simplified version of KKT conditions

$$\begin{aligned} \mathbf{D}\mathbf{X} &= \mathbf{Y}, \\ \mathbf{D}^\top\mathbf{D} &= \mathbf{I}, \\ \mathbf{0} &\in \operatorname{sign}(\mathbf{X})\mathbf{X}^\top - \mathbf{X}\operatorname{sign}(\mathbf{X})^\top. \end{aligned} \tag{6}$$

b) KKT conditions of problem (3): We also derive the KKT conditions for the optimization problem (3). Towards that end, we form the Lagrangian of (3)

$$\mathcal{L}(\mathbf{D},\mathbf{\Lambda}) = \|\mathbf{D}^\top\mathbf{Y}\|_1 - \left\langle \mathbf{D}^\top\mathbf{D}-\mathbf{I}, \mathbf{\Lambda}\right\rangle. \tag{7}$$

Then the KKT conditions of (3) is given by [4]

$$\begin{aligned} \mathbf{0} &= \nabla_{\mathbf{\Omega}}\mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{\Lambda}), \\ \mathbf{0} &\in \partial_{\mathbf{D}}\mathcal{L}(\mathbf{D},\mathbf{X},\mathbf{\Lambda}). \end{aligned} \tag{8}$$

Plugging the form (7) of $\mathcal{L}(\mathbf{D},\mathbf{\Lambda})$ to the KKT conditions (8), we can simplify (3) as follows:

$$\begin{aligned} \mathbf{D}^\top\mathbf{D} &= \mathbf{I}, \\ \mathbf{D}(\mathbf{\Lambda}+\mathbf{\Lambda}^\top) &\in \mathbf{Y}\operatorname{sign}(\mathbf{Y}^\top\mathbf{D}), \end{aligned} \tag{9}$$

where in the second line we use the subdifferential chain rule $\partial_{\mathbf{D}}\|\mathbf{D}^\top\mathbf{Y}\|_1 = \mathbf{Y}\operatorname{sign}(\mathbf{Y}^\top\mathbf{D})$. Now multiplying $\mathbf{D}^\top$ on both sides of the second line of (9) and taking a transpose, we get

$$\mathbf{\Lambda}+\mathbf{\Lambda}^\top \in \operatorname{sign}(\mathbf{D}^\top\mathbf{Y})(\mathbf{D}^\top\mathbf{Y})^\top,$$

which is equivalent to

$$\mathbf{0} \in \operatorname{sign}(\mathbf{D}^\top\mathbf{Y})(\mathbf{D}^\top\mathbf{Y})^\top - (\mathbf{D}^\top\mathbf{Y})\operatorname{sign}(\mathbf{D}^\top\mathbf{Y})^\top. \tag{10}$$

Then we get an equivalent form of the KKT conditions of (3)

$$\begin{aligned} \mathbf{D}^\top\mathbf{D} &= \mathbf{I}, \\ \mathbf{0} &\in \operatorname{sign}(\mathbf{D}^\top\mathbf{Y})(\mathbf{D}^\top\mathbf{Y})^\top - (\mathbf{D}^\top\mathbf{Y})\operatorname{sign}(\mathbf{D}^\top\mathbf{Y})^\top. \end{aligned} \tag{11}$$

A. Local identifiability

In this part, we consider under what conditions, the ground-truth dictionary and coefficient matrix $\{\mathbf{D}_\sharp, \mathbf{X}_\sharp\}$ can be locally identifiable by solving the optimizations (2) and (3). The following Bernouli-Gaussian model is commonly used in the sparse recovery and dictionary learning problems [5]–[10].

Therefore, we assume the same model on the sparse coefficient matrix $\mathbf{X}_\sharp$.

Assumption 1. *Suppose each element of $\mathbf{X}_\sharp$ is iid generated from a Bernouli-Gaussian distribution of parameter θ, denoted by $\mathbf{X}_\sharp \sim_{iid} \mathrm{BG}(\theta)$, i.e.,*

$$\mathbf{X}_\sharp = \mathbf{B} \odot \mathbf{G}, \qquad \mathbf{B} \sim_{iid} \mathrm{Ber}(\theta), \qquad \mathbf{G} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where the parameter θ controls the sparsity level.

Under Assumption 1 with parameter θ, this work aims to build the relationship of the parameter θ and the probability of local identifiability of the ground-truth dictionary and coefficient $\{\mathbf{D}_\sharp, \mathbf{X}_\sharp\}$.

Theorem 1. *Suppose the ground-truth sparse coefficient matrix $\mathbf{X}_\sharp \in \mathbb{R}^{n\times L}$ satisfies Assumption 1, i.e., $\mathbf{X}_\sharp \sim_{iid} \mathrm{BG}(\theta)$ for some parameter $\theta \in (0,1)$ and the ground-truth dictionary is any orthogonal matrix in $\mathcal{O}_n$. The dataset $\mathbf{Y} \in \mathbb{R}^{n\times L}$ is generated by (1). Then the ground-truth dictionary and coefficient matrix $\{\mathbf{D}_\sharp, \mathbf{X}_\sharp\}$ satisfy the KKT conditions (6) of problem (2) with probability at least $1-8n^2 \exp\left(-\frac{\frac{1}{2\pi}\theta(1-\theta)^2 L}{2\theta+(1-\theta)\sqrt{\frac{2}{\pi}}}\right)$.*

Theorem 2. *Suppose the ground-truth sparse coefficient matrix $\mathbf{X}_\sharp \in \mathbb{R}^{n\times L}$ satisfies Assumption 1, i.e., $\mathbf{X}_\sharp \sim_{iid} \mathrm{BG}(\theta)$ for some parameter $\theta \in (0,1)$ and the ground-truth dictionary is any orthogonal matrix in $\mathcal{O}_n$. The dataset $\mathbf{Y} \in \mathbb{R}^{n\times L}$ is generated by (1). Then the ground-truth dictionary and coefficient matrix $\{\mathbf{D}_\sharp, \mathbf{X}_\sharp\}$ satisfy the KKT conditions (11) of problem (3) with probability at least $1-8n^2 \exp\left(-\frac{\frac{1}{2\pi}\theta(1-\theta)^2 L}{2\theta+(1-\theta)\sqrt{\frac{2}{\pi}}}\right)$.*

B. Related work

- [2] considers using ℓ_0 norm minimization to recover the sparse coefficient matrix. It shows that $\mathbf{D}_\#$ is globally optimal if $\|\mathbf{X}_\#(:,j)\|_0 \le n/2$ for $j \in [n]$. However, solving the ℓ_0 norm minimization is an NP-hard problem.
- [9] considers the same computational method and covers the overcomplete dictionary case, but it requires a high sample complexity $L = \Omega(n^4\theta)$ for the local identifiability of the ground-truth dictionary and coefficient matrix. In the meanwhile, to achieve a probability with $1-\epsilon$, our work requires $L = \Omega\left(\frac{\log(8n^2/\epsilon)}{\theta(1-\theta)^2}\right)$.
- [11] studies the same cost function and shows that $\mathbf{D}_\#$ is locally optimal with a probability $1-\epsilon$ under the sample complexity $L = \Omega\left(n\log(n)/(\theta\epsilon^2)\right)$.
- Another line of work [6], [8], [12] recover the columns of the ground-truth dictionary one by one from the unit spheres. By characterizing the favorable geometry of the corresponding optimization problem where every critical point is either a ground-truth atom or a strict saddle, this line of work ensure the recovery of the ground-truth dictionary and sparse coefficient. This geometric based analysis has been applied to other important problems in signal processing and machine learning [13]–[22].

IV. Proof

Both proofs of Theorems 1 and 2 rely on the following technical lemma.

Lemma 1. *Suppose the ground-truth sparse coefficient matrix $\mathbf{X}_\sharp \in \mathbb{R}^{n\times L}$ satisfies Assumption 1, i.e., $\mathbf{X}_\sharp \sim_{iid} \mathrm{BG}(\theta)$ for some parameter $\theta \in (0,1)$. Then the following condition*

$$\mathbf{0} \in \mathrm{sign}(\mathbf{X}_\sharp)\mathbf{X}_\sharp^\top - \mathbf{X}_\sharp \,\mathrm{sign}(\mathbf{X}_\sharp)^\top \tag{12}$$

holds with probability at least $1-8n^2 \exp\left(-\frac{\frac{1}{2\pi}\theta(1-\theta)^2 L}{2\theta+(1-\theta)\sqrt{\frac{2}{\pi}}}\right)$.

The formal proof is provided in the appendix. We can observe that the probability of (12) scales positively as the number of columns L. The intuitions is as follows. Consider two sparse rows $\mathbf{X}_\sharp(i,:)$ and $\mathbf{X}_\sharp(j,:)$. One sufficient condition of (12) is that any two sparse rows $\mathbf{X}_\sharp(i,:)$ and $\mathbf{X}_\sharp(j,:)$ have disjoint supports. See Figure 2 for an example. The probability of joint support with a single position is the probability of the event that both entry is nonzero, which is of probability θ^2. Thus the probability for a disjoint single support is $(1-\theta^2)$. Then the probability of disjointedness of the whole support is $(1-\theta^2)^L$. Finally, the probability of two vector having common support is $1-(1-\theta^2)^L$ is an increasing function with respect to L.

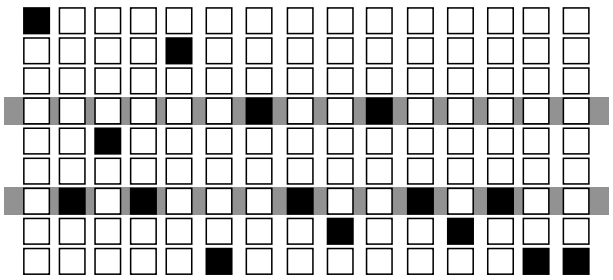


Figure 2. Intuition for (12). The two graw-shaded rows denote two disjoint sparse rows $\mathbf{X}_\sharp(i,:)$ and $\mathbf{X}_\sharp(j,:)$ of $\mathbf{X}_\sharp \in \mathbb{R}^{n\times L}$. The empty square means a zero entry and the black square means a non-zero entry.

A. Proof of Theorem 1

To prove Theorem 1, we show that the KKT conditions (6) hold with high probability for the ground-truth pair $\{\mathbf{D}_\sharp, \mathbf{X}_\sharp\}$.

First of all, by the data-generating model (1) and the assumption that the ground-truth dictionary $\mathbf{D}_\sharp \in \mathcal{O}_n$, we immediately have that the ground-truth pair $\{\mathbf{D}_\sharp, \mathbf{X}_\sharp\}$ satisfies the first and second lines of the KKT conditions (6). Therefore, it remains to show the last line of the KKT conditions (6), which is

$$\mathbf{0} \in \mathrm{sign}(\mathbf{X}_\sharp)\mathbf{X}_\sharp^\top - \mathbf{X}_\sharp \,\mathrm{sign}(\mathbf{X}_\sharp)^\top.$$

Then the proof is completed by using Lemma 1.

B. Proof of Theorem 2

To prove Theorem 2, we show that the KKT conditions (11) hold with high probability for the ground-truth pair $\{\mathbf{D}_\sharp, \mathbf{X}_\sharp\}$.

The first line of (11) is satisfied because $\mathbf{D}_\sharp \in \mathcal{O}_n$. By plugging the data-generating model (1) in the second line of (11), it remains to show the following condition holds with high probability

$$\mathbf{0} \in \operatorname{sign}(\mathbf{X}_\sharp)\mathbf{X}_\sharp^\top - \mathbf{X}_\sharp \operatorname{sign}(\mathbf{X}_\sharp)^\top.$$

Then the proof follows by combining Assumption 1 that $\mathbf{X}_\sharp \sim_{iid} \mathrm{BG}(\theta)$ and Lemma 1.

APPENDIX

A. Technical Tools

Lemma 2 (Moments of the Gaussian Random Variable)**.** *[23] If $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$, then it holds for all integer $m \geq 1$ that*

$$\mathbb{E}[|X|^m] \leq \sigma_X^m (m-1)!!.$$

Lemma 3 (Moment-Control Bernstein's Inequality for Random Variables)**.** *Let $X_1, \cdots, X_N$ be iid real-valued random variables. Suppose that there exist some positive numbers R and σ_X^2 such that*

$$\mathbb{E}[|X_k|^m] \leq \frac{m!}{2}\sigma_X^2 R^{m-2}, \quad \text{for all integers } m \geq 2.$$

Let $S \doteq \frac{1}{N}\sum_{k=1}^N X_k$, then for all $t > 0$, it holds that

$$\mathbb{P}\left[|S - \mathbb{E}[S]| \geq t\right] \leq 2\exp\left(-\frac{Nt^2}{2\sigma_X^2 + 2Rt}\right).$$

B. Proof of Lemma 1

Proof. To simplify notations in the analysis, we denote

$$\mathbf{X}_\sharp = \begin{bmatrix} -\mathbf{x}_1^\top - \\ \vdots \\ -\mathbf{x}_n^\top - \end{bmatrix}$$

with $\mathbf{x}_k^\top$ for $k \in [n]$ as the rows of $\mathbf{X}_\sharp$. Then the condition (12) is equivalent to

$$\begin{aligned} &\mathbf{0} \in \operatorname{sign}(\mathbf{X}_\sharp)\mathbf{X}_\sharp^\top - \mathbf{X}_\sharp \operatorname{sign}(\mathbf{X}_\sharp)^\top, \\ \iff &0 \in \operatorname{sign}(\mathbf{x}_i)^\top \mathbf{x}_j - \operatorname{sign}(\mathbf{x}_j)^\top \mathbf{x}_i, \forall i, j, \\ \iff &0 \in \sum_{k\in I_j} \operatorname{sign}(\mathbf{x}_i(k))\mathbf{x}_j(k) - \sum_{k\in I_i} \operatorname{sign}(\mathbf{x}_j(k))\mathbf{x}_i(k), \forall i, j, \end{aligned} \tag{13}$$

where $I_i := \{k : \mathbf{x}_i(k) \neq 0\}$.

Define

$$\begin{aligned} I_{i,j}^1 &:= \{k : k \in I_i, k \notin I_j\}, \\ I_{i,j}^2 &:= \{k : k \in I_i, k \in I_j\}, \\ I_{i,j}^3 &:= \{k : k \notin I_i, k \in I_j\}. \end{aligned} \tag{14}$$

Then (13) is equivalent to

$$\begin{aligned} &\sum_{k\in I_{i,j}^2} \left(\operatorname{sign}(\mathbf{x}_i(k))\mathbf{x}_j(k) - \operatorname{sign}(\mathbf{x}_j(k))\mathbf{x}_i(k)\right) \\ \in &\sum_{k\in I_{i,j}^1 \cup I_{i,j}^3} \left(\operatorname{sign}(\mathbf{x}_i(k))\mathbf{x}_j(k) - \operatorname{sign}(\mathbf{x}_j(k))\mathbf{x}_i(k)\right). \end{aligned} \tag{15}$$

We now further simplify (15). Using the definition of $\operatorname{sign}(\cdot)$ function $\operatorname{sign}(x) = \begin{cases} +1, & x > 0 \\ -1, & x < 0 \\ [-1, 1], & x = 0 \end{cases}$ and recalling the definitions of $I_{ij}^1, I_{ij}^2, I_{ij}^3$ in (14), we observe that the left-hand-side (LHS) of (15) is a scalar and the right-hand-side (RHS) of (15) is a sum of sets of form $[-\sigma_i, \sigma_i]$ for some positive $\sigma_i > 0$ determined by the coefficients. By symmetry of the sets, we claim that (15) is equivalent to the following condition

$$\begin{aligned} &\left|\sum_{k\in I_{i,j}^2} \left(\operatorname{sign}(\mathbf{x}_i(k))\mathbf{x}_j(k) - \operatorname{sign}(\mathbf{x}_j(k))\mathbf{x}_i(k)\right)\right| \\ \leq &\left\|\mathbf{x}_i(I_{i,j}^1 \cup I_{i,j}^3)\right\|_1 + \left\|\mathbf{x}_j(I_{i,j}^1 \cup I_{i,j}^3)\right\|_1. \end{aligned} \tag{16}$$

Therefore, the following conditions are equivalent:

$$(12) \iff (13) \iff (15) \iff (16)$$

Therefore, it suffices to show the condition (16) holds with high probability.

Upper bound the LHS of (16) First note that for any $i \neq j$,

$$\begin{aligned} &\frac{1}{L}\sum_{k\in I_{i,j}^2} \operatorname{sign}(\mathbf{x}_i(k))\mathbf{x}_j(k) \\ =&\frac{1}{L}\sum_{k=1}^L \mathbf{b}_i(k)\mathbf{b}_j(k)\operatorname{sign}(\mathbf{g}_i(k))\mathbf{g}_j(k) \\ =&\frac{1}{L}\sum_{k=1}^L \mathbf{u}(k), \end{aligned}$$

where $\mathbf{u}(k) := \mathbf{b}_i(k)\mathbf{b}_j(k)\operatorname{sign}(\mathbf{g}_i(k))\mathbf{g}_j(k)$. It is clear that

$$\mathbb{E}\left[\frac{1}{L}\sum_{k=1}^L \mathbf{u}(k)\right] = 0.$$

We now bound the moments of $\mathbf{u}(k)$ by

$$\mathbb{E}[|\mathbf{u}(k)|^m] = \theta^2 \mathbb{E}[|\mathbf{g}_j(k)|^m] \leq \theta^2 (m-1)!!,$$

where the last inequality follows from Lemma 2. Thus, by utilizing the Moment-Control Bernstein's Inequality for Random Variables in Lemma 3, we have

$$\mathbb{P}\left[\left|\frac{1}{L}\sum_{k=1}^L \mathbf{u}(k)\right| \geq t\right] \leq 2\exp\left(-\frac{Lt^2}{2\theta^2 + 2t}\right).$$

Similarly,

$$\mathbb{P}\left[\left|\frac{1}{L}\sum_{k\in I_{i,j}^2} \operatorname{sign}(\mathbf{x}_j(k))\mathbf{x}_i(k)\right| \geq t\right] \leq 2\exp\left(-\frac{Lt^2}{2\theta^2 + 2t}\right).$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{1}{L}\sum_{k\in I_{i,j}^2}(\operatorname{sign}(\mathbf{x}_i(k))\mathbf{x}_j(k)-\operatorname{sign}(\mathbf{x}_j(k))\mathbf{x}_i(k))\right|\geq 2t\right]\\
&\leq \mathbb{P}\left[\left|\frac{1}{L}\sum_{k\in I_{i,j}^2}\operatorname{sign}(\mathbf{x}_i(k))\mathbf{x}_j(k)\right|\geq t\right]\\
&\quad+\mathbb{P}\left[\left|\frac{1}{L}\sum_{k\in I_{i,j}^2}\operatorname{sign}(\mathbf{x}_j(k))\mathbf{x}_i(k)\right|\geq t\right]\\
&\leq 4\exp\left(-\frac{Lt^2}{2\theta^2+2t}\right).
\end{aligned}$$

Lower bound the RHS of (16) Similarly, for any $i\neq j$,

$$\begin{aligned}
\frac{1}{L}\|\mathbf{x}_i(I_{i,j}^1\cup I_{i,j}^3)\|_1&=\frac{1}{L}\sum_{k=1}^{L}\mathbf{b}_i(k)(1-\mathbf{b}_j(k))|\mathbf{g}_i(k)|\\
&=\frac{1}{L}\sum_{k=1}^{L}\mathbf{v}(k),
\end{aligned}$$

where $\mathbf{v}(k):=\mathbf{b}_i(k)(1-\mathbf{b}_j(k))|\mathbf{g}_i(k)|$. It is clear that

$$\mathbb{E}\left[\mathbf{v}(k)\right]=\theta(1-\theta)\sqrt{\frac{2}{\pi}}.$$

We now bound the moments of $\mathbf{v}(k)$ by

$$\mathbb{E}[|\mathbf{v}(k)|^m]=\theta(1-\theta)\mathbb{E}[|\mathbf{g}_i(k)|^m]\leq\theta(1-\theta)(m-1)!!,$$

where the last inequality follows from Lemma 2. Thus, it follows from Lemma 3 that

$$\mathbb{P}\left[\left|\frac{1}{L}\sum_{k=1}^{L}\mathbf{v}(k)-\theta(1-\theta)\sqrt{\frac{2}{\pi}}\right|\geq t\right]\leq 2\exp\left(-\frac{Lt^2}{2\theta(1-\theta)+2t}\right)$$

Therefore,

$$\begin{aligned}
\mathbb{P}\Bigg[\Bigg|\frac{1}{L}\Big(\|\mathbf{x}_i(I_{i,j}^1\cup I_{i,j}^3)\|_1+\|\mathbf{x}_j(I_{i,j}^1\cup I_{i,j}^3)\|_1\Big)-2\theta(1-\theta)\sqrt{\frac{2}{\pi}}\Bigg|\\
\geq 2t\Bigg]\leq 4\exp\left(-\frac{Lt^2}{2\theta(1-\theta)+2t}\right).\quad(17)
\end{aligned}$$

Putting together Choosing $t=\frac{1}{2}\theta(1-\theta)\sqrt{\frac{2}{\pi}}$, we obtain

$$\begin{aligned}
&\mathbb{P}\Bigg[\Bigg|\sum_{k\in I_{i,j}^2}(\operatorname{sign}(\mathbf{x}_i(k))\mathbf{x}_j(k)-\operatorname{sign}(\mathbf{x}_j(k))\mathbf{x}_i(k))\Bigg|\\
&\qquad\geq\|\mathbf{x}_i(I_{i,j}^1\cup I_{i,j}^3)\|_1+\|\mathbf{x}_j(I_{i,j}^1\cup I_{i,j}^3)\|_1\Bigg]\\
&\qquad\leq 8\exp\left(-\frac{\frac{1}{2\pi}\theta(1-\theta)^2L}{2\theta+(1-\theta)\sqrt{\frac{2}{\pi}}}\right).
\end{aligned}$$

□

REFERENCES

[1] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[2] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

[3] F. H. Clarke, *Optimization and nonsmooth analysis*, vol. 5. Siam, 1990.

[4] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[5] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE transactions on information theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

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

[7] S. Li, H. Mansour, and M. B. Wakin, "An optimization view of music and its extension to missing data," *arXiv preprint arXiv:1806.03511*, 2018.

[8] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere i: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2016.

[9] Q. Geng and J. Wright, "On the local correctness of ℓ_1-minimization for dictionary learning," in *2014 IEEE International Symposium on Information Theory*, pp. 3180–3184, IEEE, 2014.

[10] R. Gribonval and K. Schnass, "Dictionary identification-sparse matrix-factorisation via ℓ_1-minimisation," *IEEE Transactions on Information Theory*, vol. 56, no. ARTICLE, pp. 3523–3539, 2010.

[11] S. Wu and B. Yu, "Local identifiability of ℓ_1-minimization dictionary learning: a sufficient and almost necessary condition," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6121–6176, 2017.

[12] Y. Li and Y. Bresler, "Multichannel sparse blind deconvolution on the sphere," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7943–7947, IEEE, 2019.

[13] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points?online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, pp. 797–842, 2015.

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

[15] Q. Li, Z. Zhu, and G. Tang, "Geometry of factored nuclear norm regularization," *arxiv:1704.01265*, 2017.

[16] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "The global optimization geometry of low-rank matrix optimization," *arXiv preprint arXiv:1703.01256*, 2017.

[17] Q. Li, Z. Zhu, and G. Tang, "The non-convex geometry of low-rank matrix optimization," *Information and Inference: A Journal of the IMA*, p. iay003, 2018.

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

[19] Q. Li, X. Yang, Z. Zhu, G. Tang, and M. B. Wakin, "The geometric effects of distributing constrained nonconvex optimization problems," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, IEEE, 2019.

[20] Q. Li, Z. Zhu, G. Tang, and M. B. Wakin, "The geometry of equality-constrained global consensus problems," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7928–7932, IEEE, 2019.

[21] S. Li, G. Tang, and M. B. Wakin, "The landscape of non-convex empirical risk with degenerate population risk," in *Advances in Neural Information Processing Systems*, pp. 3502–3512, 2019.

[22] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1233–1242, JMLR. org, 2017.

[23] A. Winkelbauer, "Moments and absolute moments of the normal distribution," *arXiv preprint arXiv:1209.4340*, 2012.