Journal of Machine Learning Research 21 (2020) 1-52 Submitted 2/19; Revised 2/20; Published 4/20

Unique Sharp Local Minimum in ℓ_1-minimization Complete Dictionary Learning

Yu Wang WANG.YU@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Siqi Wu * SIQI@STAT.BERKELEY.EDU
Citadel Securities
131 South Dearborn
Chicago, IL 60603, USA

Bin Yu BINYU@BERKELEY.EDU
Department of Statistics and EECS
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Editor: Sanjiv Kumar

Abstract

We study the problem of globally recovering a dictionary from a set of signals via ℓ_1-minimization. We assume that the signals are generated as *i.i.d.* random linear combinations of the K atoms from a complete reference dictionary $\boldsymbol{D}^* \in \mathbb{R}^{K\times K}$, where the linear combination coefficients are from either a Bernoulli type model or exact sparse model. First, we obtain a necessary and sufficient norm condition for the reference dictionary $\boldsymbol{D}^*$ to be a sharp local minimum of the expected ℓ_1 objective function. Our result substantially extends that of Wu and Yu (2018) and allows the combination coefficient to be non-negative. Secondly, we obtain an explicit bound on the region within which the objective value of the reference dictionary is minimal. Thirdly, we show that the reference dictionary is the unique sharp local minimum, thus establishing the first known global property of ℓ_1-minimization dictionary learning. Motivated by the theoretical results, we introduce a perturbation based test to determine whether a dictionary is a sharp local minimum of the objective function. In addition, we also propose a new dictionary learning algorithm based on Block Coordinate Descent, called DL-BCD, which is guaranteed to decrease the obective function monotonically. Simulation studies show that DL-BCD has competitive performance in terms of recovery rate compared to other state-of-the-art dictionary learning algorithms when the reference dictionary is generated from random Gaussian matrices.

Keywords: dictionary learning, ℓ_1-minimization, local and global identifiability, non-convex optimization, sharp local minimum

*. The views expressed in this paper reflect those of Siqi Wu and should not be interpreted to represent the views of Citadel Securities or its personnel.

1. Introduction

Dictionary learning is a class of unsupervised learning algorithms that learn a data-driven representation from signals such as images, speech, and video. It has been widely used in many applications ranging from image imputation to texture synthesis (Rubinstein et al., 2010; Mairal et al., 2009a; Peyré, 2009). Compared to pre-defined dictionaries, data-driven dictionaries can extract meaningful and interpretable patterns from scientific data (Olshausen and Field, 1996, 1997) and exhibit enhanced performance in blind source separation, image denoising and matrix completion. See, e.g., Zibulevsky and Pearlmutter (2001); Kreutz-delgado et al. (2003); Lesage et al. (2005); Elad and Aharon (2006); Aharon et al. (2006); Mairal et al. (2009b); Qiu et al. (2014) and the references therein. Dictionary learning is also closely related to Non-negative Matrix Factorization (NMF) (Lee and Seung, 2001) which has broad applications in biology (Brunet et al., 2004; Wu et al., 2016).

Despite many successful applications, dictionary learning formulations and algorithms are generally hard to analyze due to their non-convex nature. With different initial inputs, a dictionary learning algorithm typically outputs different dictionaries as a result of this non-convexity. For those who use the dictionary as a basis for downstream analyses, the choice of the dictionary may significantly impact the final conclusions. Therefore, it is natural to ask the following questions: can dictionary learning algorithms recover the "ground-truth" dictionary if there is one? Among the many outputs from a dictionary learning algorithm, which one should be selected for further analysis?

To answer the above questions, we need to understand the theoretical properties of dictionary learning under generative models. In a number of recent works, the signals are generated as linear combinations of the columns of a reference dictionary (Gribonval and Schnass, 2010; Geng et al., 2014; Gribonval et al., 2015). Specifically, denoting by $\boldsymbol{D}^* \in \mathbb{R}^{d\times K}$ the reference dictionary and $\boldsymbol{x}^{(i)} \in \mathbb{R}^d, i = 1, \dots, n$ the signal vectors, we have:

$$\boldsymbol{x}^{(i)} \approx \boldsymbol{D}^* \boldsymbol{\alpha}^{(i)}, \tag{1}$$

where $\boldsymbol{\alpha}^{(i)} \in \mathbb{R}^K$ denotes the sparse coefficient vector. If $K = d$ and $\boldsymbol{D}^*$ is full rank, the dictionary is called *complete*. If the matrix has more columns than rows, i.e., $K > d$, the dictionary is *overcomplete*. Under the model (1), for any reasonable dictionary learning objective function, the reference dictionary D^* ought to be equal or close to a local minimum. This wellposedness requirement, also known as *local identifiability* of dictionary learning, turns out to be nontrivial. For a complete dictionary and noiseless signals, Gribonval and Schnass (2010) studies the following ℓ_1-minimization formulation:

$$\begin{aligned} \text{minimize}_{\boldsymbol{D},\{\boldsymbol{\beta}^{(i)}\}_{i=1}^n} & \sum_{i=1}^n \|\boldsymbol{\beta}^{(i)}\|_1. \\ \text{subject to } & \|\boldsymbol{D}_j\|_2 \le 1, j = 1, \dots, K, \\ & \boldsymbol{x}^{(i)} = \boldsymbol{D}\boldsymbol{\beta}^{(i)},\ i = 1, \dots, n. \end{aligned} \tag{2}$$

They proved a sufficient condition for local identifiability under the Bernoulli-Gaussian model. A more refined analysis by Wu and Yu (2018) gave a sufficient and almost necessary condition. The sufficient local identifiability condition in Gribonval and Schnass (2010) was extended to the over-complete case (Geng et al., 2014) and the noisy case (Gribonval et al., 2015).

As most of dictionary learning formulations are nonconvex, local identifiability alone does not guarantee that the output dictionary is the reference dictionary — the initial dictionary must also be quite close to the reference dictionary. There are only limited results on how to choose an appropriate initialization. For example, Arora et al. (2015) showed that their initialization algorithm guarantees that the output dictionary is within a small neighborhood of the reference dictionary when certain μ-incoherence condition is met. In practice, initialization is usually done by using a random matrix or randomly selecting a sample of signals (Mairal et al., 2014). These algorithms are typically run for multiple times and the dictionary that achieves the smallest objective value is selected.

The difficulty of initialization is a major challenge of establishing the recovery guarantee that under some generative models, the output dictionary of an algorithm is indeed the reference dictionary. This motivates the study of *global identifiability.* There are two versions of global identifiability. For the first version, we say that the reference dictionary $\boldsymbol{D}^*$ is globally identifiable with respect to an objective function $L(\cdot)$ if $\boldsymbol{D}^*$ is a global minimum of L. The second and stricter version, requires all local minima of L are the same as $\boldsymbol{D}^*$ up to column sign changes and permutation. If the second version of global identifiability holds, all local minima are global minimum. Thus any algorithm capable of converging to a local minimum will also recover the reference dictionary. For some matrix decomposition tasks such as low rank PCA (Srebro and Jaakkola, 2003) and matrix completion (Ge et al., 2016), despite the fact that the objective function is non-convex, the stricter version of global identifiability holds under certain conditions. For dictionary learning, several papers proposed new algorithms with theoretical recovery guarantees that ensure the output is close or equal to the reference dictionary. For the complete and noiseless case, Spielman et al. (2013) proposed a linear programming based algorithm that provably recovers the reference dictionary when the coefficient vectors are generated from a Bernoulli Gaussian model and contain at most $O(\sqrt{K})$ nonzero elements. Sun et al. (2017a,b) improved the sparsity tolerance to $O(K)$ using a Riemannian trust region method. For over-complete dictionaries, Arora et al. (2014b) proposed an algorithm which performs an overlapping clustering followed by an averaging algorithm or a K-SVD type algorithm. Additionally, there is another line of research that focuses on the analysis of alternating minimization algorithms, including Agarwal et al. (2013, 2014); Arora et al. (2014a, 2015); Chatterji and Bartlett (2017). Barak et al. (2014) proposed an algorithm based on sum-of-square semi-definite programming hierarchy and proved its desirable theoretical performance with relaxed assumptions on coefficient sparsity under a series of moment assumptions.

1.1. Our contributions

Despite numerous studies of global recovery in dictionary learning, there are no global identifiability results for the ℓ_1-minimization problem. As we illustrate in Section 3, the reference dictionary may not be the global minimum even for a simple data generation model. This motivates us to consider a different condition to distinguish the reference dictionary from other local minima. We show that the reference dictionary is the unique “sharp” local minimum (see Definition 1) of the ℓ_1 objective function when certain conditions are met – in other words, there are no other sharp local minima than the reference dictionary.

Based on this new characterization and the observation that a sharp local minimum is more resilient to small perturbations, we propose a method to empirically test the sharpness of the objective function at a given dictionary. Furthermore, we also design a new algorithm to solve the ℓ_1-minimization problem using Block Coordinate Descent (DL-BCD) and the re-weighting scheme inspired by Candes et al. (2008). Our simulations demonstrate that the proposed method compares favorably with other state-of-the-art algorithms in terms of recovery rate if the reference dictionary is generated from random Gaussian matrices.

Our work differs from other recent studies in two main aspects. Firstly, instead of proposing new dictionary learning formulations, we study the global property of the existing ℓ_1-minimization problem that is often considered difficult in previous studies (Mairal et al., 2009b; Wu and Yu, 2018). While there are many dictionary learning algorithms that do not rely on the ℓ_1-type penalty, formulations with ℓ_1 penalties remain as the most frequently used method in many applications due to their good practical performance and the availability of efficient algorithms (Mairal et al., 2009b,a). The theoretical understanding of ℓ_1-minimization is therefore of interest to a wider audience than other dictionary learning methods. Secondly, our data generation models are novel and cover several important cases not studied by prior works, e.g., non-negative linear coefficients. Even though there is a line of research that focuses on non-negative dictionary learning in the literature (Aharon et al., 2005; Hoyer, 2002; Arora et al., 2014a), the reference dictionary and the corresponding coefficients therein are both non-negative. In comparison, we allow the dictionary to have arbitrary values but only constrain the reference coefficients to be non-negative. This non-negative coefficient case is difficult to analyze and does not satisfy the recovery conditions in previous studies, for instance Barak et al. (2014); Sun et al. (2017a,b).

The rest of this paper is organized as follows. Section 2 introduces notations and basic assumptions. Section 3 presents main theorems and discusses their implications. Section 4 proposes the sharpness test and the block coordinate descent algorithm for dictionary learning (DL-BCD). Simulation results are provided in Section 5. We conclude our results and discuss possible extensions in Section 6.

2. Preliminaries

For a vector $\boldsymbol{w} \in \mathbb{R}^m$, denote its j-th element by w_j. For an arbitrary matrix $A \in \mathbb{R}^{m\times n}$, let $A[k,], A_j, A_{k,j}$ denote its k-th row, j-th column, and the (k,j)-th element respectively. Denote by $A[k,-j] \in \mathbb{R}^{n-1}$ the k-th row of A without its j-th entry. Let $\mathbb{I} \in \mathbb{R}^{K\times K}$ denote the identity matrix of size K and for $k \in \{1,\ldots,K\}$, $\mathbb{I}_k$ denotes $\mathbb{I}$'s k-th column, whose k-th entry is one and zero elsewhere. $\mathbf{1} \in \mathbb{R}^{K\times 1}$ denotes a column vector whose elements are all ones. For a positive semi-definite square matrix $X \in \mathbb{R}^{K\times K}$, $X^{1/2}$ denotes its positive semi-definite square root. We use $\|\cdot\|$ to denote vector norms and $|||\cdot|||$ to denote matrix (semi-)norms. In particular, $|||\cdot|||_F$ denotes the Frobenius norm, whereas $|||\cdot|||_2$ denotes the spectral norm. For any two real functions $w(t), q(t) : \mathbb{R} \to \mathbb{R}$, we denote $w(t) = \Theta(q(t))$ if there exist constants $c_1, c_2 > 0$ such that for any $t \in \mathbb{R}$, $c_1 < \frac{w(t)}{q(t)} < c_2$. If $q(t) > 0$ and $\lim_{q(t)\to 0} \frac{w(t)}{q(t)} = 0$, then we write $w(t) = o(q(t))$. Define the indicator and the sign functions as

$$\mathbf{1}(x=0) = \begin{cases} 1 & x = 0 \\ 0 & x \neq 0 \end{cases}, \ \text{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}.$$

2.1. The ℓ_1 dictionary learning objective

In dictionary learning, a dictionary is represented by a matrix $\boldsymbol{D} \in \mathbb{R}^{d \times K}$. We call a column of the dictionary matrix an atom of the dictionary. In this paper, we consider complete dictionaries, that is, the dictionary matrix is square ($K = d$) and invertible. Note that for the noiseless case, an undercomplete dictionary ($K < d$) can always be reduced to a complete dictionary by removing certain rows. A complete or undercomplete dictionary matrix is typically used in applications such as Independent Component Analysis (Comon, 1994) and Non-negative Matrix Factorization (Lee and Seung, 2001; Brunet et al., 2004; Wu et al., 2016).

For a complete dictionary $\boldsymbol{D}$, define L as the ℓ_1 objective function:

$$L(\boldsymbol{D}) = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\beta}^{(i)}\|_1. \quad \text{where } \boldsymbol{\beta}^{(i)} = \boldsymbol{D}^{-1}\boldsymbol{x}^{(i)} \ \forall \ i \in 1, \ldots, n. \tag{3}$$

The ℓ_1-minimization formulation (2) is equivalent to the following optimization problem (Wu and Yu, 2018):

$$\text{minimize}_{\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)} L(\boldsymbol{D}), \tag{4}$$

where $\mathbb{B}(\mathbb{R}^K)$ is the set of all feasible dictionaries:

$$\mathbb{B}(\mathbb{R}^K) \triangleq \left\{ \boldsymbol{D} \in \mathbb{R}^{K \times K} \middle| \|\boldsymbol{D}_1\|_2 = \ldots = \|\boldsymbol{D}_K\|_2 = 1, \ \text{rank}(\boldsymbol{D}) = K \right\}.$$

2.2. Generative models

Let $\boldsymbol{D}^* \in \mathbb{B}(\mathbb{R}^K)$ be the reference dictionary of interest. We assume that the signal vector $\boldsymbol{x} \in \mathbb{R}^K$ is generated from a linear model without noise: $\boldsymbol{x} = \boldsymbol{D}^*\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \in \mathbb{R}^K$ is a random reference coefficient vector. Below, we will introduce two classes of generative models for $\boldsymbol{\alpha}$: Bernoulli type models and exact sparse models.

- *Bernoulli type model* $\mathcal{B}(p_1, \ldots, p_K; f)$. Let $\boldsymbol{z} \in \mathbb{R}^K$ be a random vector whose probability density function exists and is denoted by f. Let $\boldsymbol{\xi} \in \{0,1\}^K$ be a random boolean vector. The coordinates of $\boldsymbol{\xi}$ are independent and ξ_j is a Bernoulli random variable with success probability $P(\xi_j = 1) = p_j \in (0,1)$. Define $\boldsymbol{\alpha} \in \mathbb{R}^K$ such that $\alpha_j = \xi_j z_j$ for all j. We say that $\boldsymbol{\alpha}$ is generated from the Bernoulli type model $\mathcal{B}(p_1, \ldots, p_K; f)$.

- *Exact sparse model* $\mathcal{S}(s; f)$. Let $\boldsymbol{z} \in \mathbb{R}^K$ be a random vector whose probability density function exists and is denoted by f. Let $\boldsymbol{S}$ be a size-s subset uniformly drawn from all size-s subsets of $1, \ldots, K$. Let $\boldsymbol{\xi} \in \{0,1\}^K$ be a random variable such that $\xi_j = 1$ if $j \in \boldsymbol{S}$ otherwise 0. Define $\boldsymbol{\alpha} \in \mathbb{R}^K$ such that $\alpha_j = \xi_j z_j$ for all j. We say that $\boldsymbol{\alpha}$ is generated from the exact sparse model $\mathcal{S}(s; f)$.

These two classes can be viewed as natural extensions of Bernoulli Gaussian models and sparse Gaussian models, which have been extensively studied in dictionary learning (Gribonval and Schnass, 2010; Wu and Yu, 2018; Schnass, 2015, 2014). Denote by $\mathcal{N}(0, \mathbb{I}_{k\times k})$ the k-dimensional standard Gaussian distribution:

- *Bernoulli Gaussian model.* If $\boldsymbol{\alpha}$ is generated from the Bernoulli type model with parameters $p_j = p$ ($p > 0$) for all j and $f = \mathcal{N}(0, \mathbb{I}_{k\times k})$, we say that $\boldsymbol{\alpha}$ follows a Bernoulli Gaussian model with parameter p, or $BG(p)$.

- *Sparse Gaussian model.* If $\boldsymbol{\alpha}$ is generated from the exact sparse model with sparsity parameter s and $f = \mathcal{N}(0, \mathbb{I}_{k\times k})$, we say that $\boldsymbol{\alpha}$ follows the sparse Gaussian model with parameter s, or $SG(s)$.

Remarks: The advantage of using sparse Gaussian and Bernoulli Gaussian distributions is that they are simple and yet capable of capturing the most important characteristic of the reference coefficients: sparsity. By using sparse Gaussian and Bernoulli Gaussian distributions, Wu and Yu (2018) obtains a sufficient and almost necessary condition for local identifiability. Take sparse Gaussian distribution as an example: let the maximal collinearity μ of the reference dictionary $\boldsymbol{D}^*$ be $\mu = \max_{i\neq j} \left|\boldsymbol{D}_i^{*T}\boldsymbol{D}_j^*\right|$ and s be the sparsity of the reference coefficient vector in the sparse Gaussian model. They show that local identifiability holds when $\mu < \frac{K-s}{\sqrt{s(K-1)}}$. From the formula, we can see a trade-off between the maximal collinearity μ and the sparsity of the coefficient vector s. If the coefficient is very sparse, i.e., $s \ll K$, local identifiability holds for a wide range of μ. Otherwise, local identifiability holds for a narrower range of μ. While sparse/Bernoulli Gaussian models can be used to illustrate this trade-off, they are rather restrictive for real data. Several papers (Spielman et al., 2013; Arora et al., 2014a,b, 2015; Gribonval et al., 2015) studied more general models such as sub-Gaussian models.

Other important examples include models with z drawn from the Laplacian distribution or a non-negative distribution. In particular, the non-negativity of the coefficients breaks the popular zero expectation assumption $\mathbb{E}\boldsymbol{\alpha}_j = 0$ (Gribonval and Schnass, 2010; Gribonval et al., 2015).

- *Sparse Laplacian model.* If $\boldsymbol{\alpha}$ is generated from the exact sparse model with sparsity parameter s and density $f(z) = \frac{1}{2^K}\exp(-\|z\|_1)$, we say that $\boldsymbol{\alpha}$ follows the sparse Laplacian model with parameter s, or $SL(s)$.

- *Non-negative Sparse Gaussian model.* A random vector $\boldsymbol{\alpha}$ is said to be drawn from a non-negative sparse Gaussian model with parameter s, denoted by $|SG(s)|$, if for $j = 1, \ldots, K$, $\boldsymbol{\alpha}_j = |\boldsymbol{\alpha}'_j|$ where $\boldsymbol{\alpha}' \sim SG(s)$.

2.3. Identifiability of the reference dictionary

In this subsection, we introduce commonly used terminologies in dictionary learning with respect to the identifiability of the reference dictionary.

- *Sign-permutation ambiguity.* In most dictionary learning formulations, the order of the dictionary atoms as well as their signs do not matter. Let $P \in \mathbb{R}^{K\times K}$ be a permutation matrix and $\Lambda \in \mathbb{R}^{K\times K}$ a diagonal matrix with ± 1 diagonal entries. The matrix $\boldsymbol{D}' = \boldsymbol{D}P\Lambda$ and $\boldsymbol{D}$ essentially represent the same dictionary but $\boldsymbol{D}' \neq \boldsymbol{D}$ element-wise.

- *Local identifiability.* The reference dictionary $\boldsymbol{D}^* \in \mathbb{B}(\mathbb{R}^K)$ is *locally identifiable* with respect to L if $\boldsymbol{D}^*$ is a local minimum of L. Local identifiability is a minimal requirement for recovering the reference dictionary. It has been extensively studied under a variety of dictionary learning formulations (Gribonval and Schnass, 2010; Geng et al., 2014; Gribonval et al., 2015; Wu and Yu, 2018; Agarwal et al., 2014; Schnass, 2014).

- *Global identifiability.* The reference dictionary $\boldsymbol{D}^* \in \mathbb{B}(\mathbb{R}^K)$ is *globally identifiable* with respect to L if $\boldsymbol{D}^*$ is a global minimum of L.

Clearly, whether local or global identifiability holds depends on the objective function and the signal generation model. If the objective function is ℓ_0, i.e., $\frac{1}{n}\sum_i \|\boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}\|_0$, and the linear coefficients are generated from the Bernoulli Gaussian model, the reference dictionary is globally (and hence locally) identifiable (see Theorem 3 in Spielman et al. (2013)). However, for the ℓ_1 objective considered in this paper, global identifiability might not hold. In Section 3, we give an example where the reference dictionary is only a local minimum but not a global minimum.

In this paper, we consider a variant of global identifiability: instead of the global minimum, we require the reference dictionary $\boldsymbol{D}^*$ to be the *unique sharp local minimum* of the dictionary learning objective function. In other words, no dictionary other than $\boldsymbol{D}^*$ is a sharp local minimum. Other dictionaries can still be local minima but cannot be *sharp* at the same time. This property allows us to globally distinguish the reference dictionary from other spurious local minima and can be used as a criterion to select the best dictionaries from a set of algorithm outputs. Sharp local minimum, as per Definition 1, is a common concept in the field of optimization (Dhara and Dutta, 2011; Polyak, 1979). However, to the best of our knowledge, we are the first to connect dictionary learning theory with sharp local minimum and use it to distinguish the reference dictionary from other spurious local minima.

Definition 1 (Sharp local minimum) *Let $L(\boldsymbol{D}) : \mathbb{B}(\mathbb{R}^K) \to \mathbb{R}$ be a dictionary learning objective function. A dictionary $\boldsymbol{D}^0 \in \mathbb{B}(\mathbb{R}^K)$ is a sharp local minimum of $L(\cdot)$ with sharpness ϵ (Polyak, 1979) if there exists $\delta > 0$ such that for any $\boldsymbol{D} \in \{\boldsymbol{D} : |||\boldsymbol{D} - \boldsymbol{D}^0|||_F < \delta\}$:*

$$L(\boldsymbol{D}) - L(\boldsymbol{D}^0) \geq \epsilon |||\boldsymbol{D} - \boldsymbol{D}^0|||_F + o(|||\boldsymbol{D}^0 - \boldsymbol{D}|||_F).$$

Remarks: The definition here can be viewed as a matrix analog of the sharp minimum in the one dimensional case. For a function $f : \mathbb{R} \to \mathbb{R}$, v^0 is a sharp local minimum of f is $f(v) - f(v^0) \geq \epsilon|v - v^0| + o(|v - v^0|)$. Note that the definition of sharp local minimum is different from the definition of strict local minimum, which means there are no other local minima in its neighborhood. A sharp local minimum is always a strict local minimum but

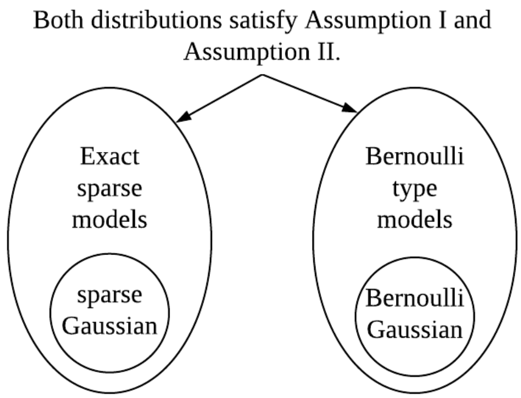

Figure 1: Both exact sparse models and Bernoulli type models satisfy Assumption I and II. Sparse Gaussian distribution is a special case of exact sparse models, while Bernoulli Gaussian distribution is a special case of Bernoulli type models.

not vice versa. For example, consider ℓ_q functions $|x|^q$ for $q > 0$. When $q \leq 1$, $x = 0$ is a strict local minimum as well as a sharp local minimum of ℓ_q. When $q > 1$, $x = 0$ is still a strict local minimum but not a sharp local minimum. This definition is also different from the sharp local minimum concepts that are commonly used in the study of artificial neural networks and stochastic gradient descent (Hochreiter and Schmidhuber, 1997).

2.4. Technical assumptions

In this subsection, we introduce two important technical assumptions that will be used in our theoretical analysis. All the models introduced in Section 2.2 satisfy these assumptions. Their relationship is depicted in Fig. 1.

We need additional notations before introducing the assumptions. For any $\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)$, define $M(\boldsymbol{D}) = \boldsymbol{D}^T\boldsymbol{D}$ as the collinearity matrix of $\boldsymbol{D}$. For example, if the dictionary is an orthogonal matrix, $M(\boldsymbol{D}) = \mathbb{I}$ is the identity matrix. If all the atoms in the dictionary are collinear with constant $\mu > 0$, then $M(\boldsymbol{D}) = \mu\mathbf{1}\mathbf{1}^T + (1-\mu)\mathbb{I}$ is a matrix whose off-diagonal elements are all μ's. When the context is clear, we use M instead of $M(\boldsymbol{D})$ for notation ease. Denote by M^* the collinearity matrix for the reference dictionary $\boldsymbol{D}^*$. Also, define the matrix $B(\boldsymbol{\alpha}, M) \in \mathbb{R}^{K\times K}$ as

$$(B(\boldsymbol{\alpha}, M))_{k,j} \triangleq \mathbb{E}\boldsymbol{\alpha}_j \text{sign}(\boldsymbol{\alpha}_k) - M_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j| \quad \text{for } k, j = 1, \ldots, K.$$

Here the expectation is with respect to the random coefficient vector $\boldsymbol{\alpha}$. By the definition of $B(\boldsymbol{\alpha}, M^*)$, the quantity is the difference between two matrices

$$B(\boldsymbol{\alpha}, M^*) = B_1(\boldsymbol{\alpha}) - B_2(\boldsymbol{\alpha}, M^*),$$

where $(B_1(\boldsymbol{\alpha}))_{k,j} = \mathbb{E}\boldsymbol{\alpha}_j \text{sign}(\boldsymbol{\alpha}_k)$ and $B_2(\boldsymbol{\alpha}, M^*)_{k,j} = M^*_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j|$. Roughly speaking, the first matrix measures the "correlation" between different coordinates of the coefficients while the second matrix measures the collinearity of the atoms in the reference dictionary. For instance, when the coordinates of $\boldsymbol{\alpha}$ are independent and mean zero, $B_1(\boldsymbol{\alpha}) = 0$. When all atoms in the dictionary are orthogonal, i.e., $M^* = \mathbb{I}$, $B_2(\boldsymbol{\alpha}, M^*) = 0$. In that extreme case, $B(\boldsymbol{\alpha}, M) = 0$.

For any random vector $\boldsymbol{\alpha}$, define the semi-norm $|||\cdot|||_{\boldsymbol{\alpha}}$ induced by $\boldsymbol{\alpha}$ as:

$$|||A|||_{\boldsymbol{\alpha}} \triangleq \sum_{k=1}^{K} \mathbb{E}|\sum_{j=1}^{K} A_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0).$$

Note that the subscript $\boldsymbol{\alpha}$ in $|||\cdot|||_{\boldsymbol{\alpha}}$ is used to indicate the dependence on the distribution of $\boldsymbol{\alpha}$. $|||\cdot|||_{\boldsymbol{\alpha}}$ is a semi-norm but not a norm because $|||A|||_{\boldsymbol{\alpha}} = 0$ does not imply $A = 0$. Actually, for any nonzero diagonal matrix $A \neq 0$, $|||A|||_{\boldsymbol{\alpha}} = 0$ because $\sum_{k=1}^{K} \mathbb{E}|A_{k,k}\boldsymbol{\alpha}_k|\mathbf{1}(\boldsymbol{\alpha}_k = 0) = 0$. Note that the reason why we define B and $|||\cdot|||_{\boldsymbol{\alpha}}$ this way is because these quantities appear naturally in the first order optimality condition of ℓ_1-minimization. Hopefully, the motivation of defining these definitions will become clear later.

Assumption I (Regular data-dependent norm) $|||\cdot|||_{\boldsymbol{\alpha}}$ is $c_{\boldsymbol{\alpha}}$-regular: There exists a number $c_{\boldsymbol{\alpha}} > 0$, dependent on the distribution of $\boldsymbol{\alpha}$, such that for any matrix $A \in H^K$, where $H^K = \{A \in \mathbb{R}^{K \times K} \mid A_{i,i} = 0 \text{ for all } 1 \leq i \leq K\}$, $|||A|||_{\boldsymbol{\alpha}}$ is bounded below by A's Frobenious norm: $|||A|||_{\boldsymbol{\alpha}} \geq c_{\boldsymbol{\alpha}} |||A|||_F$.

Note that similar to $|||\cdot|||_{\boldsymbol{\alpha}}$, we use the subscript $\boldsymbol{\alpha}$ in $c_{\boldsymbol{\alpha}}$ to indicate that the quantity $c_{\boldsymbol{\alpha}}$ depends on the distribution of $\boldsymbol{\alpha}$. Assumption I has several implications. First, it ensures that the coefficient vector $\boldsymbol{\alpha}$ does not lie in a linear subspace of $\mathbb{R}^K$. Otherwise, we can make rows of A orthogonal to $\boldsymbol{\alpha}$ and show that $|||\cdot|||_{\boldsymbol{\alpha}}$ is not regular. Second, it also guarantees that the coefficient vector $\boldsymbol{\alpha}$ must have some level of sparsity. To see why this is the case, suppose there exists some coordinate k' such that the coefficient $\boldsymbol{\alpha}_{k'} \neq 0$ almost surely. We can then construct A such that all of its elements are zero except the k'-th row. Thus, $|||A|||_{\boldsymbol{\alpha}} = \mathbb{E}|\sum_{j=1}^{K} A_{k',j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_{k'} = 0) = 0$, but $|||A|||_F > 0$. Third, if we define the dual (semi-)norm of $|||\cdot|||_{\boldsymbol{\alpha}}$ in the subspace H^K as

$$|||X|||_{\boldsymbol{\alpha}}^* = \sup_{A \neq 0, A \in H^K} \frac{\operatorname{tr}(X^T A)}{|||A|||_{\boldsymbol{\alpha}}}, \quad \text{for } X \in \mathbb{R}^{K \times K},$$

the regularity of $|||\cdot|||_{\boldsymbol{\alpha}}$ implies that the corresponding dual semi-norm is bounded above by the Frobenius norm. To see this, simply note that $|||X|||_{\boldsymbol{\alpha}}^* \leq \frac{1}{c_{\boldsymbol{\alpha}}} |||X|||_F$ with the above definition. Assumption I is crucial for the study of the local identifiability property. As can be seen later in Theorems 4 and 5, regularity of $|||\cdot|||_{\boldsymbol{\alpha}}$ is indispensable in determining the sharpness of the local minimum corresponding to the reference dictionary $\boldsymbol{D}^*$ as well as the bounding region.

Assumption II (Probabilistic linear independence) For any fixed constants $c_1, \ldots, c_K \in \mathbb{R}$, the following statement holds almost surely

$$\sum_{l=1}^{K} c_l \boldsymbol{\alpha}_l = 0 \Longrightarrow c_l \boldsymbol{\alpha}_l = 0 \ \forall \ l = 1, \ldots, K,$$

or equivalently, for any fixed $c_1, \dots, c_K$,

$$P\left(\sum_{l=1}^{K} c_l \boldsymbol{\alpha}_l = 0, \sum_{l=1}^{K} c_l^2 \boldsymbol{\alpha}_l^2 > 0\right) = 0.$$

Assumption II controls the sparsity of any coefficient vector $\boldsymbol{\beta}$ under a general dictionary $\boldsymbol{D}$. For the noiseless signal $\boldsymbol{x} = \boldsymbol{D}^* \boldsymbol{\alpha}$, its j-th coefficient under a dictionary $\boldsymbol{D}$ can be written as a linear combination of reference coefficients $\boldsymbol{\alpha}_l$: $\boldsymbol{\beta}_j = \boldsymbol{D}^{-1}[j,]\boldsymbol{D}^*\boldsymbol{\alpha} = \sum_{l=1}^{K} c_l \boldsymbol{\alpha}_l$ where $c_l = \boldsymbol{D}^{-1}[j,]\boldsymbol{D}_l^*$ for $l = 1, \dots, K$. Thus, Assumption II implies that under any general dictionary, the resulting coefficient $\boldsymbol{\beta}_j$ is zero if and only if for each l, either the reference coefficient is zero ($\boldsymbol{\alpha}_l = 0$) or the corresponding constant is zero ($c_l = 0$). In other words, elements in the reference coefficient vector cannot "cancel" with each other unless all the elements are zeros. This assumption seems very similar to the *linear independence* property of random variables (Rodgers et al., 1984): Random variables $\psi_1, \dots, \psi_K$ are linearly independent if $c_1\psi_1 + \ldots + c_K\psi_K = 0$ a.s. implies $c_1 = c_2 = \cdots = c_K = 0$. It is worth pointing out that Assumption II is a weaker assumption than linear independence. Many distributions of interest, such as Bernoulli Gaussian distributions, are not linearly independent but satisfy Assumption II (Proposition 3). This assumption is essential when we study the uniqueness of the sharp local minimum in Theorem 8.

In the following propositions, we show that both Bernoulli type models and exact sparse models satisfy Assumption I and II.

Proposition 2 *The norm $|||\cdot|||_{\boldsymbol{\alpha}}$ induced by exact sparse models or Bernoulli type models satisfy Assumption I. The regularity constant has explicit form when the coefficient is from $SG(s)$ or $BG(p)$:*

- *If $\boldsymbol{\alpha}$ is from $SG(s)$, the norm $|||\cdot|||_{\boldsymbol{\alpha}}$ is c_s-regular, where $c_s \geq \frac{s(K-s)}{K(K-1)}\sqrt{\frac{2}{\pi}}$.*
- *If $\boldsymbol{\alpha}$ is from $BG(p)$, the norm $|||\cdot|||_{\boldsymbol{\alpha}}$ is c_p-regular, where $c_p \geq p(1-p)\sqrt{\frac{2}{\pi}}$.*

Proposition 3 *If the coefficient vector is generated from a Bernoulli type model or an exact sparse model, Assumption II holds.*

Remarks: Although those assumptions are quite general, certain distributions considered in other papers do not satisfy our assumptions. A key requirement in Bernoulli type or exact sparse models is that the probability density function of the base random variable z must exist. For instance, the Bernoulli Randemacher model (Spielman et al., 2013) does not satisfy Assumption II. To see this, take the following Bernoulli Randemacher model for $K = 2$ as an example: suppose $\boldsymbol{\xi} \in \{0,1\}^2$ where $P(\xi_1 = 1) = p_1$, $P(\xi_2 = 1) = p_2$. The base random vector $z \in \{-1,1\}^2$ with $P(z_1 = 1) = P(z_2 = 1) = 1/2$. If we take $c_1 = 1$ and $c_2 = -1$, $P(c_1\boldsymbol{\alpha}_1 + c_2\boldsymbol{\alpha}_2 = 0, c_1\boldsymbol{\alpha}_1 \neq 0, c_2\boldsymbol{\alpha}_2 \neq 0) = P(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2 = 0, \xi_1 \neq 0, \xi_2 \neq 0) = P(\xi_1 = 1, \xi_2 = 1, z_1 = z_2) = p_1 \cdot p_2/2 > 0$. Therefore, Assumption II does not apply in this case.

3. Main Theoretical Results

Similar to Wu and Yu (2018), we first study the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{D}}{\text{minimize}} \quad & \mathbb{E}L(\boldsymbol{D}) = \mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 \\ \text{subject to} \quad & \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \end{aligned} \tag{5}$$

Here, the notation $\mathbb{E}$ is the expectation with respect to $\boldsymbol{x} = \boldsymbol{D}^*\boldsymbol{\alpha}$ under a probabilistic model for $\boldsymbol{\alpha}$. Therefore, this optimization problem is equivalent to the case when we have infinite number of samples. As we shall see, the analysis of this population level problem gives us significant insights into the identifiability properties of dictionary learning. We also consider the finite sample case (4) in Theorem 9.

3.1. Local identifiability

In this subsection, we will establish a sufficient and necessary condition for the reference dictionary to be a sharp local minimum.

Theorem 4 (Local identifiability) *Suppose* $|||\cdot|||_{\boldsymbol{\alpha}}$ *is* $c_{\boldsymbol{\alpha}}$*-regular (see Assumption I) and the* ℓ_1 *norm of the reference coefficient vector* $\boldsymbol{\alpha}$ *has bounded first order moment:* $\mathbb{E}\|\boldsymbol{\alpha}\|_1 < \infty$*.* $\boldsymbol{D}^*$ *is a sharp local minimum of Formulation* (5) *with sharpness at least* $\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}^*|||_2^2}(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*)$ *if and only if*

$$|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* < 1. \tag{6}$$

If $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* > 1$*,* $\boldsymbol{D}^*$ *is not a local minimum.*

Remarks: Wu and Yu (2018) studied the local identifiability problem when the coefficient vector $\boldsymbol{\alpha}$ is from Bernoulli Gaussian or sparse Gaussian distributions. They gave a sufficient and almost necessary condition that ensures the reference dictionary to be a local minimum. Theorem 4 substantially extends their result in two aspects:

- The reference coefficient distribution can be exact sparse models and Bernoulli type models, which is more general than sparse/Bernoulli Gaussian models.
- In addition to showing that the reference dictionary $\boldsymbol{D}^*$ is a local minimum, we show that $\boldsymbol{D}^*$ is actually a sharp local minimum with an explicit bound on the sharpness.

To prove Theorem 4, we need to calculate how the objective function changes along any direction in the neighborhood of the reference dictionary. The major challenge of this calculation is that the objective function is neither convex nor smooth, which prevents us from using sub-gradient or gradient to characterize its local structure. Instead, we obtain a novel sandwich-type inequality of the ℓ_1 objective function (Lemma 19). With the help of this inequality, we are able to carry out a more fine-grained analysis of the ℓ_1-minimization objective. The detailed proof of Theorem 4 can be found in the Appendix 10.

Theorem 4 gives the condition under which the reference dictionary is a sharp local minimum. The below Theorem 5 gives an explicit bound of the size of the region. To the best of authors' knowledge, this is the first result about the region where local identifiability holds for ℓ_1-minimization.

Theorem 5 *Under notations in Theorem 4, if* $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* < 1$*, for any* $\boldsymbol{D}$ *in the set*

$$S = \left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| |||\boldsymbol{D}|||_2 \leq 2|||\boldsymbol{D}^*|||_2, |||\boldsymbol{D} - \boldsymbol{D}^*|||_F \leq \frac{(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*) \cdot c_{\boldsymbol{\alpha}}}{8\sqrt{2}|||\boldsymbol{D}^*|||_2^2 \max_j \mathbb{E}|\boldsymbol{\alpha}_j|} \right\},$$

we have $\mathbb{E}L(\boldsymbol{D}) \geq \mathbb{E}L(\boldsymbol{D}^*)$.

Remarks: First of all, note that the set S we study here is different from what Agarwal et al. (2014) called the "basin of attraction". The basin of attraction of an iterative algorithm is the set of initialization dictionaries under which the algorithm converges to the reference dictionary $\boldsymbol{D}^*$. For an iterative algorithm that decreases its objective function at each step, its basin of attraction must be a subset of the region within which $\boldsymbol{D}^*$ has the minimal objective value. Secondly, Theorem 5 only tells us that $\boldsymbol{D}^*$ admits the smallest objective function value within the set S. It does not, however, indicates that $\boldsymbol{D}^*$ is the only local minimum within S.

For certain generative models, the conditions in Theorem 4 and 5 can be made more explicit to compare with other local identifiability results. In what follows, we will study two examples to gain a better understanding of those conditions. These examples demonstrate the trade-off between coefficient sparsity, collinearity of atoms in the reference dictionary and signal dimension K. For simplicity, we set the reference dictionary to be the constant collinearity dictionary with coherence $\mu > 0$: $\boldsymbol{D}^*(\mu) = ((1-\mu)\mathbb{I} + \mu \mathbf{1}\mathbf{1}^T)^{1/2}$ where $\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{K\times K}$ is a square matrix whose elements are all ones. This simple dictionary class was used to illustrate the local identifiability conditions in Gribonval and Schnass (2010) and Wu and Yu (2018). The coherence parameter μ controls the collinearity between dictionary atoms. By studying this class of reference dictionaries, we can significantly simplify the conditions and demonstrate how the coherence μ affects dictionary identifiability.

Corollary 6 *Suppose the reference dictionary* $\boldsymbol{D}^*$ *is a constant collinearity dictionary with coherence* $\mu > 0$*:* $\boldsymbol{D}^*(\mu) = ((1-\mu)\mathbb{I} + \mu \mathbf{1}\mathbf{1}^T)^{1/2}$ *, and the reference coefficient vector* $\boldsymbol{\alpha}$ *is from* $SG(s)$*. If and only if*

$$\mu\sqrt{s} < \frac{K-s}{K-1},$$

$\boldsymbol{D}^*$ *is a sharp local minimum with sharpness at least* $\frac{s}{\sqrt{\pi}(1+\mu(K-1))K}\left(\frac{K-s}{K-1} - \mu\sqrt{s}\right)$*. For any*

$$\boldsymbol{D} \in S = \left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| |||\boldsymbol{D}|||_2 \leq 2\sqrt{1+\mu(K-1)}, |||\boldsymbol{D} - \boldsymbol{D}^*|||_F \leq \frac{\frac{K-s}{K-1} - \mu\sqrt{s}}{8\sqrt{2}(1+\mu(K-1))} \right\},$$

we have $\mathbb{E}L(\boldsymbol{D}) \geq \mathbb{E}L(\boldsymbol{D}^*)$.

Three parameters play important roles for the reference dictionary to be a sharp local minimum: dictionary coherence μ, sparsity s and dimension K. Since $\mu\sqrt{s} - \frac{K-s}{K-1}$ is a monotonically increasing function with respect to μ and s, local identifiability holds when the dictionary is close to an orthogonal matrix and the coefficient vector is sufficiently sparse. Another important observation is that $\mu\sqrt{s} - \frac{K-s}{K-1}$ is monotonically decreasing as K increases. Thus, given that the number of nonzero elements per signal s is fixed, it is easier for the local identifiability condition to hold for larger K. If K tends to infinity, the condition becomes $s < \frac{1}{\sqrt{\mu}}$. Also, the set S shrinks as s or μ increases, implying that the region is smaller when the coefficients are less sparse or the dictionary has higher coherence. When $\mu = 0$, the set S becomes $\left\{\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| |||\boldsymbol{D}|||_2 \leq 2, |||\boldsymbol{D} - \boldsymbol{D}^*|||_F \leq \frac{1}{8\sqrt{2}} \frac{K-s}{K-1}\right\}$.

Next, we consider non-negative sparse Gaussian distribution in the following example. Since we do not have the explicit form of the regularity constant $c_{\boldsymbol{\alpha}}$ for non-negative sparse Gaussian distribution, we omit the corresponding results for the sharpness and the region bound.

Corollary 7 *Suppose the reference dictionary is a constant collinearity dictionary with coherence $\mu > 0$: $\boldsymbol{D}^*(\mu) = ((1-\mu)\mathbb{I} + \mu \mathbf{1}\mathbf{1}^T)^{1/2}$, and the reference coefficient vector $\boldsymbol{\alpha}$ is from non-negative sparse Gaussian distribution $|SG(s)|$. If*

$$\left|\mu - \frac{s-1}{K-1}\right| < \frac{K-s}{K-1},$$

$\boldsymbol{D}^*$ is a sharp local minimum.

Note that the condition $\frac{K-1}{K-s} \cdot \left|\mu - \frac{s-1}{K-1}\right| < 1$ is equivalent to $\frac{2s-K-1}{K-1} < \mu < 1$. When K tends to infinity, the reference dictionary is a local minimum for $\mu < 1$. Compared to the same bound from Corollary 6, $\mu < \frac{1}{\sqrt{s}}$ for large K, the bound for non-negative coefficients is less restrictive. Therefore, the non-negativity of the coefficient distribution relaxes the requirement for local identifiability.

Results for other interesting examples, such as Bernoulli Gaussian coefficients and sparse Laplacian coefficients, can be found in the Appendix 8.

3.2. Global identifiability

For ℓ_1-minimization, multiple local minima exist: as a result of sign-permutation ambiguity, if $\boldsymbol{D}$ is a local minimum, for any permutation matrix P and any diagonal matrix Λ with diagonal elements ± 1, $\boldsymbol{D}P\Lambda$ is also a local minimum. These local minima are benign in nature since they essentially refer to the same dictionary. Can there be other local minima than the benign ones? If so, how can we distinguish benign local minima from them? In this subsection, we consider the problem of global identifiability. First, we give a counter-example to show that the reference dictionary is not necessary a global minimum of the ℓ_1-minimization problem even for orthogonal dictionary and sparse coefficients.

Counter-example on global identifiability. Suppose the reference dictionary is the identity matrix $\mathbb{I} \in \mathbb{R}^{2\times 2}$. The coefficients are generated from a Bernoulli-type model $\boldsymbol{\alpha} \in \mathbb{R}^2$ such that $\boldsymbol{\alpha}_i = z_i \xi_i$ for $i = 1, 2$, where ξ_1 and ξ_2 are Bernoulli variables with success probability 0.67, and (z_1, z_2) is drawn from the below Gaussian mixture model:

$$\frac{1}{2}\mathcal{N}\left(0, \begin{pmatrix} 101 & -99 \\ -99 & 101 \end{pmatrix}\right) + \frac{1}{2}\mathcal{N}\left(0, \begin{pmatrix} 101 & 99 \\ 99 & 101 \end{pmatrix}\right).$$

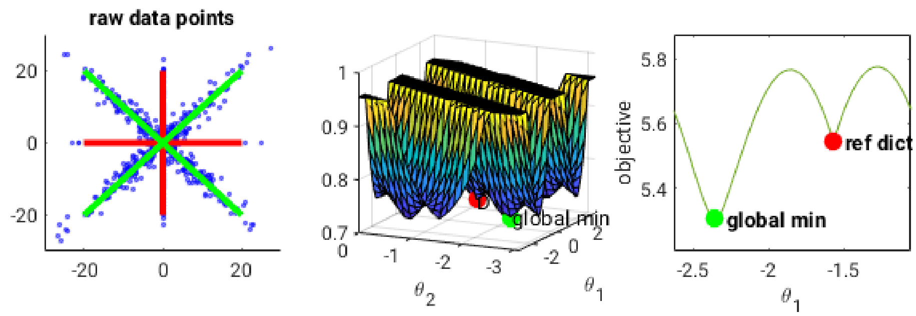

Figure 2: The empirical data (Left) and the objective surface plot (Middle). We parameterize a candidate dictionary as $\boldsymbol{D} = (a_1, a_2)$, where $a_1 = (\cos(\theta_1), \sin(\theta_1))$, $a_2 = (\cos(\theta_2), \sin(\theta_2))$. The objective of $\boldsymbol{D}$ is defined as in (3). Green dots/lines indicate global minima, whereas red dots/lines are the reference dictionary or its sign-permutation equivalents. The Right figure shows the objective curve for all orthogonal dictionaries ($\theta_1 - \theta_2 = \pi/2$). While the reference dictionary is a sharp local minimum, it is not a global minimum.

We generate 2000 samples from the model and compute the dictionary learning objective $L(\cdot)$ defined in (3) for each candidate dictionary (Fig. 2). As can be seen from the objective function surface plot, the global minimum for this data set is not the reference dictionary. Furthermore, one can show that the global minimum is not sharp, i.e., the directional derivative along certain directions at the global minimum is close to zero for finite samples and exactly zero for the population case.

The above example shows a potential drawback of directly minimizing the ℓ_1 objective compared to other objectives such as ℓ_0. For the ℓ_0 objective, under certain Bernoulli Gaussian models, the reference dictionary is a global minimum (Spielman et al., 2013) and even in our counter-example, which is not Bernoulli Gaussian, the reference dictionary can still be shown to be a global minimum. Still, the computation complexity of the ℓ_0 objective remains too high to switch from ℓ_1. To remedy this drawback of ℓ_1, we observe that in the above example, although the reference dictionary is not a global minimum, it is still a sharp local minimum and there is no other sharp local minimum. Therefore there is hope that we can combine the ℓ_1 objective and a "sharpness" test to recover the reference dictionary. Is this observation true for general cases? The answer is yes. The following theorem shows that the reference dictionary is the unique sharp local minimum of ℓ_1-minimization up to sign-permutation.

Theorem 8 (Unique sharp local minimum) *Suppose the reference coefficient vector $\boldsymbol{\alpha}$ satisfies probabilistic linear independence (see Assumption II). If $\boldsymbol{D}^*$ is a sharp local minimum of Formulation* (5)*, it is the only sharp local minimum in $\mathbb{B}(\mathbb{R}^K)$. If it is not a sharp local minimum, there is no sharp local minimum in $\mathbb{B}(\mathbb{R}^K)$.*

Note that Theorem 8 works for the population case where the sample size is infinite. For the finite sample case, we can show that the sharpness of spurious local minima is close to zero. Define $\mathcal{D}_\epsilon$ to be

$$\mathcal{D}_\epsilon = \left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| \boldsymbol{D} \text{ is a sharp local minimum of (4) with sharpness at least } \epsilon. \right\}.$$

Define $\text{eig}(\boldsymbol{D})$ to be the set of eigen-values of the matrix $\boldsymbol{D}$. For any fixed $\epsilon > 0$ and $\rho_2 > \rho_1 > 0$, define the event $A(\rho_1, \rho_2, \epsilon)$ to be

$$\begin{aligned} A(\rho_1, \rho_2, \epsilon) = \{&\text{There exists } \boldsymbol{D} \in \mathcal{D}_\epsilon \text{ s.t. } \text{eig}(\boldsymbol{D}) \subset (\rho_1, \rho_2) \text{ and} \\ &\boldsymbol{D} \neq \boldsymbol{D}^* \text{ up to sign-permutation.}\}. \end{aligned}$$

In other words, the event A represents the "bad" event that at least one of the sharp local minima in $\mathcal{D}_\epsilon$ with bounded eigenvalues is not the reference dictionary. With this notation, Theorem 8 basically shows that for the population case, the event $A(\rho_1 = 0, \rho_2 = \infty, \epsilon = 0)$ will never happen. The next theorem shows that for the finite sample case, $P(A(\rho_1, \rho_2, \epsilon))$ is upper bounded.

Theorem 9 (Finite-sample case) *Suppose n samples of $\boldsymbol{x}^{(i)}$'s are drawn i.i.d. from a model satisfying probabilistic linear independence (see Assumption II) and for any $i = 1 \dots n$, $\|\boldsymbol{x}^{(i)}\|_2 \le L < \infty$. Then for any fixed $\rho_2 > \rho_1 > 0$ and $\epsilon > 0$,*

$$P\left(A(\rho_1, \rho_2, \epsilon)\right) \le 4 \exp\left(2K \left(\ln \frac{n}{2K} + 1 \right) - n \left(\frac{\rho_1^3 \epsilon}{2L\rho_2} - \frac{1}{n} \right)^2 \right).$$

In particular, $P(A(\rho_1, \rho_2, \epsilon)) \to 0$ as $\frac{K}{n} \to 0$.

Remarks: Theorem 9 ensures that as $K/n \to 0$, with high probability *no* dictionaries other than $\boldsymbol{D}^*$ are sharp local minima within a region $\{\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \big| \text{eig}(\boldsymbol{D}) \in (\rho_1, \rho_2)\}$. However, it does not tell whether or not $\boldsymbol{D}^*$ is a sharp local minimum. For the population case, this issue is resolved in Theorem 4, which gives a sufficient and necessary condition for the reference dictionary to be a sharp local minimum.

4. Algorithms for checking sharpness and solving ℓ_1-minimization

As shown in the previous section, the reference dictionary is the unique sharp local minimum under certain conditions. Here, we will design an algorithm that uses this property as a stopping criterion for ℓ_1-minimization. If the algorithm finds a sharp local minimum, we know that it is the reference dictionary. To do so we need to answer the following practical questions:

- How to determine numerically if a given dictionary is a sharp local minimum?
- How to find a sharp local minimum and recover the reference dictionary?

In this section, we first introduce an algorithm to check if a given dictionary is a sharp local minimum. We then develop an algorithm to recover the reference dictionary. The latter algorithm is guaranteed to decrease the (truncated) ℓ_1 objective function at each iteration (Proposition 12).

4.1. Determining sharp local minimum

Despite the intuitive concept, checking whether a given dictionary is a sharp local minimum can be challenging. First of all, the dimension of the problem is very high (K^2). Secondly, if a dictionary is a sharp local minimum, the objective function is not differentiable at that point, precluding us from using gradients or Hessian matrix to solve the problem. One might also consider using sub-gradients to minimize the objective (Bagirov et al., 2013). However, because the problem is actually non-convex, sub-gradients might not be well-defined.

We propose a novel algorithm to address these challenges. We decompose the problem into a series of sub-problems each of which is low-dimensional. In Proposition 10, we show that a given dictionary is a sharp local minimum in dimension K^2 if and only if certain vectors are sharp local minima for the corresponding sub-problems of dimension K. The objective function of each subproblem is strongly convex. To deal with non-existence of gradient or Hessian matrix, we design a perturbation test based on the observation that a sharp local minimum ought to be stable with respect to small perturbations. For instance, $x = 0$ is the sharp local minimum of $|x|$ but is non-sharp local minimum of x^2. If we add a linear function as a perturbation, $x = 0$ is still a local minimum of $|x| + \epsilon \cdot x$ for any ϵ such that $|\epsilon| < 1$ but not so for $x^2 + \epsilon \cdot x$. The choice of the perturbation is crucial. In Proposition 10, we show that adding a perturbation to the dictionary collinearity matrix M is sufficient. Note that perturbations to other quantities might work as well. Intuitively, a "good" perturbation should provide enough variability along any direction. Otherwise, a local minimum that is not sharp along certain directions might be mistakenly deemed as sharp.

Proposition 10 *The following three statements are equivalent:*

1) $\boldsymbol{D}$ *is a sharp local minimum of* (4)*.*

2) For any $k = 1, \ldots, K$*,* $\mathbb{I}_k$ *is the sharp local minimum of the strongly convex optimization:*

$$\mathbb{I}_k \in \mathrm{argmin}_{\boldsymbol{w}} \ \mathbb{E}|\langle \boldsymbol{\beta}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq k}^{K} \sqrt{(w_h - M_{k,h})^2 + 1 - M_{k,h}^2} \cdot \mathbb{E}|\boldsymbol{\beta}_h|. \tag{7}$$
$$\text{subject to } \boldsymbol{w} = [w_1, \ldots, w_K] \in \mathbb{R}^K, \ w_k = 1.$$

3) For a sufficiently small $\rho > 0$ *and any* $\tilde{M}$ *s.t.* $|\tilde{M}_{k,h} - M_{k,h}| \leq \rho$ *for any* $k, h = 1, \ldots, K$*,* $\mathbb{I}_k$ *is the local minimum of the convex optimization:*

$$\mathbb{I}_k \in \mathrm{argmin}_{\boldsymbol{w}} \ \mathbb{E}|\langle \boldsymbol{\beta}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq k}^{K} \sqrt{(w_h - \tilde{M}_{k,h})^2 + 1 - \tilde{M}_{k,h}^2} \cdot \mathbb{E}|\boldsymbol{\beta}_h|. \tag{8}$$
$$\text{subject to } \boldsymbol{w} = [w_1, \ldots, w_K] \in \mathbb{R}^K, \ w_k = 1.$$

for $k = 1, \ldots, K$*.*

Proposition 10 tells us that, in order to check whether a dictionary is a sharp local minimum, it is sufficient to add a perturbation to the matrix $M = \boldsymbol{D}^T\boldsymbol{D}$ and check whether the resulting dictionary is the local minimum of the perturbed objective function. Empirically, we can add a random Gaussian noise with a small enough variance ρ and minimize the objective (8). If $\mathbb{I}_k$, the k-th column vector of the identity matrix, is the local minimum for the perturbed objective, by Proposition 10 the given dictionary is guaranteed to be a sharp local minimum. We formalize this idea into Algorithm 1. We acknowledge that this algorithm might be conservative and misclassify a sharp local minimum as a non-sharp local minimum if ρ is not small enough as required in Proposition 10. There is no good rule-of-thumb in choosing ρ as it can be dependent on the data. We will explore the sensitivity of this algorithm with respect to choice of ρ in the simulation section.

Algorithm 1 Sharp local minimum test for ℓ_1-minimization dictionary learning

Require: Dictionary to be tested $\boldsymbol{D}$, samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, perturbation level $\rho \in \mathbb{R}^+$, threshold $T \in \mathbb{R}^+$.

for $i = 1, \ldots, n$ **do**
  $\boldsymbol{\beta}^{(i)} \leftarrow \boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}$.
end for
for $j = 1, \ldots, K$ **do**
  Generate $\epsilon_j \sim \mathcal{N}(0, \rho \cdot \mathbb{I}_{K\times K})$.
  $\tilde{\boldsymbol{D}}_j = \boldsymbol{D}_j + \epsilon_j$.
end for
for $k, h = 1, \ldots, K$ **do**
  $\tilde{M}_{k,h} \leftarrow \left\langle \tilde{\boldsymbol{D}}_h, \tilde{\boldsymbol{D}}_k \right\rangle$ if $k \neq h$ or 0.
end for
$r \leftarrow 0$
for $k = 1, \ldots, K$ **do**
  Solve the strongly convex optimization via BFGS:

$$\boldsymbol{w}^{(k)} \leftarrow \text{minimize}_{\boldsymbol{w}} \sum_{i=1}^{n} |\langle \boldsymbol{\beta}^{(i)}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq k}^{K} \sqrt{(w_h - \tilde{M}_{k,h})^2 + 1 - \tilde{M}_{k,h}^2} \cdot \sum_{i=1}^{n} |\boldsymbol{\beta}_h^{(i)}| \quad (9)$$
$$\text{subject to } \boldsymbol{w} = [w_1, \ldots, w_K] \in \mathbb{R}^K, \quad w_k = 1.$$

  $\mathbb{I}_k \leftarrow (0, \ldots, 0, 1, 0, \ldots, 0)$ where only the k-th element is 1.
  $r \leftarrow \max(r, \|\boldsymbol{w}^{(k)} - \mathbb{I}_k\|_2^2)$.
end for
if $r < T$ **then**
  Output $\boldsymbol{D}$ is a sharp local minimum.
else
  Output $\boldsymbol{D}$ is not a sharp local minimum.
end if

The main component of Algorithm 1 is solving the strongly convex optimization (9). To do so we use Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Witzgall and Fletcher, 1989), which is a second order method that estimates Hessian matrices using past gradient information. Each step of BFGS is of complexity $O(nK + K^2)$. If we assume the maximum iteration to be a constant, the overall complexity of Algorithm 1 is $O(nK^2 + K^3)$. Because sample size n is usually larger than the dimension K, the dominant term in the complexity is $O(nK^2)$. In the simulation section, we show that the empirical computation time is in line with the theoretical bound.

4.2. Recovering the reference dictionary

We now try to solve formulation (4). One of the most commonly used technique in solving dictionary learning is alternating minimization (Olshausen and Field, 1997; Mairal et al., 2009a), which is to update the coefficients and the dictionary in an alternating fashion until convergence. This method fails for noiseless ℓ_1-minimization: when the coefficients are fixed, the dictionary must also be fixed to satisfy all constraints. To allow dictionaries to be updated iteratively, researchers have proposed different ways to relax the constraints (Agarwal et al., 2014; Mairal et al., 2014). However, those workarounds tend to have numerical stability issues if a high precision result is desired (Mairal et al., 2014).

This motivates us to propose Algorithm 2. The algorithm uses the idea from Block Coordinate Descent (BCD). It updates each row of $\boldsymbol{D}^{-1}$ and the corresponding row in the coefficient matrix simultaneously. As we update one row of $\boldsymbol{D}^{-1}$, we also scale all the other rows of $\boldsymbol{D}^{-1}$ by appropriate constants. This is because if we only update one row of $\boldsymbol{D}^{-1}$ while keeping the others fixed, columns of the resulting dictionary will not have unit norm. The following lemma gives an admissible parameterization for updating one row of $\boldsymbol{D}^{-1}$.

Proposition 11 *For any dictionary $\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)$ and any coordinate $k \in 1, \ldots, K$, given a vector $w = [w_1, \ldots, w_K] \in \mathbb{R}^K$ such that $w_k = 1$, we can define a matrix $Q \in \mathbb{R}^{K\times K}$:*

$$Q[k,] = \begin{cases} w^T \boldsymbol{D}^{-1} & h = k \\ \sqrt{(w_h - M_{k,h})^2 + 1 - M_{k,h}^2} \cdot \boldsymbol{D}^{-1}[h,] & h \neq k \end{cases}.$$

Then $Q^{-1} \in \mathbb{B}(\mathbb{R}^K)$, which means each column of Q^{-1} is of norm 1.

With the parameterization in Proposition 11, we derive the following subproblems from ℓ_1-minimization dictionary learning: for $k = 1, ..., K$,

$$\begin{aligned} &\operatorname{argmin}_{\boldsymbol{w}} \sum_{i=1}^{n} |\langle \beta^{(i)}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq k}^{K} \sqrt{(\boldsymbol{w}_h - M_{k,h})^2 + 1 - M_{k,h}^2} \cdot \sum |\beta_h^{(i)}|. \\ &\text{subject to } \boldsymbol{w} = [w_1, \ldots, w_K] \in \mathbb{R}^K, w_k = 1. \end{aligned}$$

where $\boldsymbol{\beta}^{(i)} = \boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}$ for a dictionary $\boldsymbol{D}$. This new sub-problem is strongly convex, making it relatively easy to solve. Note that this problem is exactly the same as (7) in Proposition 10. Thus the optimization problem (7) is closely related to ℓ_1-minimization dictionary learning from two different perspectives: First, the sharpness of any solution of ℓ_1-minimization is equivalent to the sharpness of $\mathbb{I}_k$ for the optimization (7). Second, the optimization problem (7) can be viewed as a subproblem of ℓ_1-minimization under an appropriate parameterization.

A natural way to solve ℓ_1-minimization dictionary learning is to solve the above subproblems iteratively for each coordinate k. Similar ideas of learning a dictionary from a series of convex programs have been explored in other papers. For example, Spielman et al. (2013) reformulated the dictionary learning problem as a series of linear programs (LP) and construct a dictionary from the LP solutions. Nonetheless, their algorithm is not guaranteed to minimize the ℓ_1 objective at each iteration.

We propose a coordinate-descent-based dictionary learning Algorithm 2. It has a tuning parameter τ, which aims at improving the performance of ℓ_1-minimization under the high signal-to-noise ratio settings. When τ is set to be infinity, Algorithm 2 minimizes the ℓ_1 objective at each update. However, when the signal-to-noise ratio is high, ℓ_1-minimization sometimes ends up with a low quality result. This is commonly due to the fact that the ℓ_1-norm over-penalizes large coefficients, which breaks the local identifiability, i.e., the reference dictionary is no longer a local minimum. Similar ideas are used in the re-weighted ℓ_1 algorithms in the field of compressed sensing (Candes et al., 2008). The motivation of re-weighted algorithms is to reduce the bias of ℓ_1-minimization by imposing smaller penalty to large coefficients. In our algorithm, we simply truncate coefficient entries beyond the given threshold τ. The obtained problem is still strongly convex but this trick improves the numerical performance significantly.

The following theorem guarantees that the proposed algorithm always decreases the objective function value.

Proposition 12 (Monotonicity) *Define*

$$f(\boldsymbol{D}) = \sum_{i=1}^{n} \sum_{j=1}^{K} \min\left(\left| \boldsymbol{D}^{-1}[j,]\boldsymbol{x}^{(i)} \right|, \tau \right),$$

where τ is the threshold used in Algorithm 2. Denote by $\boldsymbol{D}^{(t,K)}$ the dictionary at the t-th iteration from Algorithm 2. $f(\boldsymbol{D}^{(t,K)})$ decreases monotonically for $t \in \mathbb{N}$: $f(\boldsymbol{D}^{(0,K)}) \geq f(\boldsymbol{D}^{(1,K)}) \geq f(\boldsymbol{D}^{(2,K)}) \ldots$

5. Numerical experiments

In this section, we evaluate the proposed algorithms with numerical simulations. We will study the empirical running time of Algorithm 1 in the first experiment and examine how the perturbation parameter ρ affects its performance in the second. In the third experiment, we study the sample size requirement for successful recovery of the reference dictionary. Finally, we will compare Algorithm 2 against other state-of-the-art dictionary learning algorithms (Parker et al., 2014a,b; Parker and Schniter, 2016). The first two less computationally intensive simulations are run on an OpenSuSE OS with Intel(R) Core(TM) i5-5200U CPU 2.20GHz with 12GB memory, while the last two simulations are conducted in a cluster with 20 cores. The source code of the DL-BCD algorithm can be found in the github repository[1].

1. https://github.com/shifwang/dl-bcd

Algorithm 2 Dictionary Learning Block Coordinate Descent (DL-BCD)

Require: Data $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, threshold τ.
  Initialize $\boldsymbol{D}^{(0,1)}$, $t \leftarrow 0$. $Q \leftarrow (\boldsymbol{D}^{(0,1)})^{-1}$.
  while Stopping criterion not satisfied **do**
    for $j = 1, \ldots, K$ **do**
      for $i = 1, \ldots, n$ **do**
        $\boldsymbol{\beta}^{(i)} \leftarrow Q\boldsymbol{x}^{(i)}$.
      end for
      for $h = 1, \ldots, K$ **do**
        $m_h \leftarrow \langle \boldsymbol{D}_h^{(t,j)}, \boldsymbol{D}_j^{(t,j)} \rangle$.
      end for
      Solve the convex optimization via BFGS:

$$\text{minimize}_{\boldsymbol{w}} \sum_{\substack{i=1..n,\\ |\beta_j^{(i)}|<\tau}} |\langle \boldsymbol{\beta}^{(i)}, \boldsymbol{w} \rangle| + \sum_{h=1, h\neq j}^{K} \sqrt{(w_h - m_h)^2 + 1 - m_h^2} \cdot \sum_{\substack{i=1..n,\\ |\beta_h^{(i)}|<\tau}} |\beta_h^{(i)}|.$$

$$\text{subject to } \boldsymbol{w} = [w_1, \ldots, w_K] \in \mathbb{R}^K,\ w_j = 1.$$

      Update j-th row of Q: $Q[j,] \leftarrow \boldsymbol{w}^T Q$.
      for $h = 1, \ldots, K$, $h \neq j$ **do**
        $Q[h,] \leftarrow Q[h,] \cdot \sqrt{(w_h - m_h)^2 + 1 - m_h^2}$.
      end for
      if $j = K$ **then**
        $\boldsymbol{D}^{(t+1,1)} \leftarrow Q^{-1}$.
      else
        $\boldsymbol{D}^{(t,j+1)} \leftarrow Q^{-1}$.
      end if
    end for
    $t \leftarrow t + 1$.
  end while

5.1. Empirical running time of Algorithm 1

We evaluate the empirical computation complexity of Algorithm 1. Let the reference dictionary be a constant collinearity dictionary with coherence $\mu = 0.5$, i.e.,

$$\boldsymbol{D}^* = (0.5\mathbb{I} + 0.5\mathbf{1}\mathbf{1}^T)^{1/2},$$

The sparse linear coefficients are generated from the Bernoulli Gaussian distribution $BG(p)$ with $p = 0.7$. This specific parameter setting ensures that the reference dictionary is not a local minimum, thus making Algorithm 1 converge slower. For a fixed dimension, the computation time scales roughly linearly with the sample size, while for fixed sample size, the computation time scales quadratically with dimension K (Fig. 3). This shows that the empirical computation complexity of Algorithm 1 is of order $O(nK^2)$, which is consistent

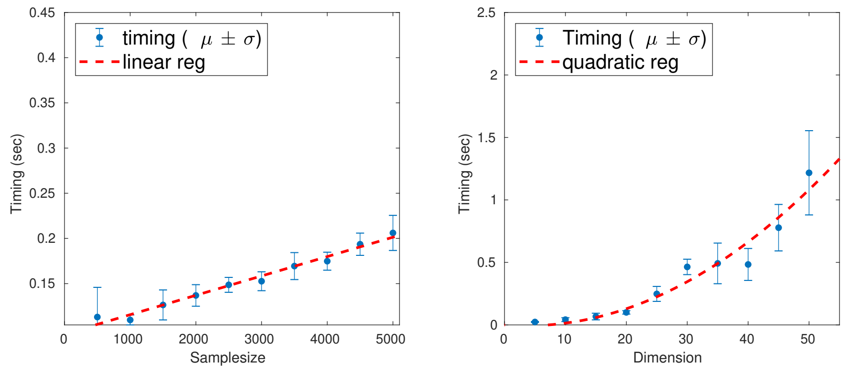

Figure 3: Computation time of Algorithm 1. Left: For $K = 20$ and $n = 500, \ldots, 5000$. Right: For $K = 5, \ldots, 50$ and $n = 400$.

with the theoretical complexity. Simulation results remain stable for different parameter settings, see Appendix 9.

5.2. Sensitivity analysis of the perturbation parameter ρ

In this experiment, we test the sensitivity of Algorithm 1 by varying the perturbation parameter ρ. We set dictionary dimension $K = 20$, sparsity parameter $s = 10$ and sample size $n = 1600$. Also, we consider constant collinearity dictionaries with coherence $\mu = \frac{1}{\sqrt{s}}(\frac{K-s}{K-1} + 0.1)$ (Fig. 4 Left) and $\mu = \frac{1}{\sqrt{s}}(\frac{K-s}{K-1} - 0.2)$ (Fig. 4 Right). For the first experiment, the reference dictionary is not a sharp local minimum of the objective function given sufficiently large sample size. Hence a small perturbation to the dictionary results in a large distance between the global minimum of the perturbed optimization and $\mathbb{I}_k$, i.e., the quantity r defined in Algorithm 1. In the second experiment, the reference dictionary is sharp, indicating the distance r in Algorithm 1 should be small after adding a perturbation. For each value of ρ between 0.05 and 0.5, we repeat the algorithm 20 times to compute the resulting distances. When ρ is small, the distance r for the non-sharp case is very big (around 1.0) whereas for the sharp case it remains small (around 10^{-12}). For the sharp case, once ρ increases beyond 0.35, r increases drastically to 10^{-3}. This experiment shows for a wide range of parameter ρ values (0.05 to 0.3), Algorithm 1 succeeds in distinguishing between the sharp and not-sharp local minima. Nonetheless, there are two caveats when using this algorithm. Firstly, the parameter ρ depends on the data generation process, which is usually not known in practice. Thus, it is still an open question about how to select ρ. Secondly, this algorithm is only useful for the noiseless case or when the noise is negligible. When the noise is significant, the reference dictionary is no longer a sharp local minimum. In that case, instead of checking the sharpness, an alternative is to check the smallest eigen-value of the Hessian matrix. This idea has not been fully explored in this paper and will be studied in future work.

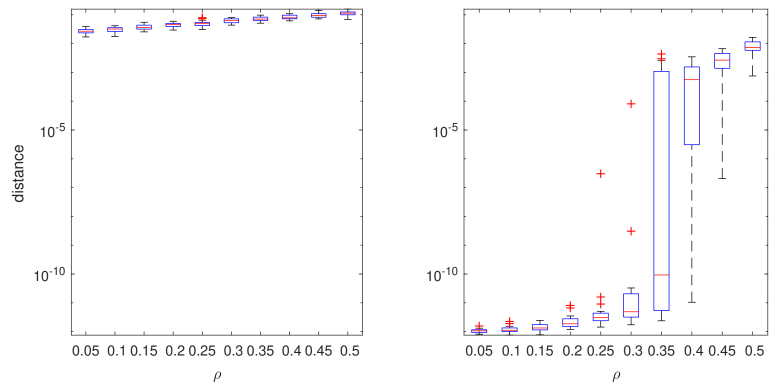

Figure 4: Sensitivity analysis of perturbation parameter ρ in Algorithm 1. Left: constant collinearity dictionary with coherence $\mu = \frac{1}{\sqrt{s}}(\frac{K-s}{K-1} + 0.1)$; Right: constant collinearity dictionary with coherence $\mu = \frac{1}{\sqrt{s}}(\frac{K-s}{K-1} - 0.2)$.

5.3. Empirical sample size requirement for local identifiability

When the reference dictionary is the constant collinearity matrix and the coefficients are sparse Gaussian, Wu and Yu (2018) shows that if the sample size n is of order $O(K \ln K)$, local identifiability holds with high probability. However, the corresponding constants that ensure local identifiability are unknown. In this subsection, we study the empirical sample size required for local identifiability with the help of Algorithm 1.

Suppose the reference dictionary has constant coherence $\mu = 0.5$ for various sizes $K = 12, 16, 20$ and the coefficients are drawn from the Sparse Gaussian distribution with sparsity $s = 5$. This specific parameter setting ensures the reference dictionary is a sharp local minimum given sufficient samples. Perturbation level is set at $\rho = 0.01$ and the threshold $T = 10^{-6}$. The experiment is repeated 20 times. Fig. 5 shows the percentage of experiments in which Algorithm 1 identifies $\boldsymbol{D}^*$ as a sharp local minimum for a variety of sample sizes n. Under this specific setting, to ensure local identifiability with 50 percent probability, the sample size n is roughly $20K$.

To further explore how dimension K affects the sample size for local identifiability, we run simulations for $K = 25, ..., 70$ and estimate the sample sizes that ensure the local identifiability with at least 50% chance. As shown in Fig. 6, the required sample size and dimension closely follow a linear relation $16.5K + 63$. It is linear, i.e., $O(K)$, instead of $O(K \ln K)$ because the sample size only ensures local identifiability with 50% chance.

5.4. Comparison with other algorithms

We compare the performance of DL-BCD with other state-of-the-art algorithms, including the greedy K-SVD algorithm (Aharon et al., 2006), SPAMS for online dictionary learning (Mairal et al., 2009b,a), ER-SpUD(proj) for square dictionaries (Spielman et al., 2013), and EM-BiG-AMP algorithm (Parker et al., 2014a,b). The implementation of these algorithms is available in the MATLAB package BiG-AMP (Parker et al., 2014a,b).

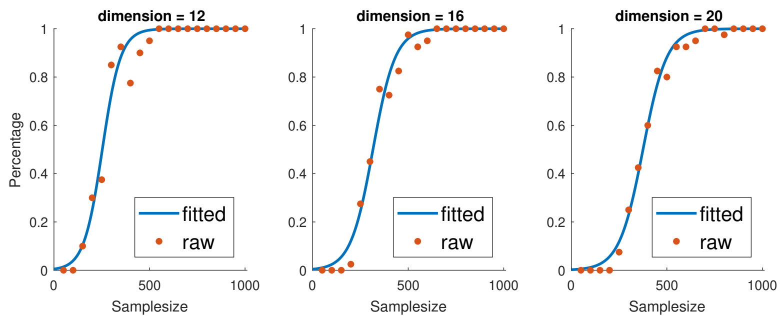

Figure 5: The percentage of experiments in which the reference dictionary is a local minimum, for different dimensions $K = 12, 16, 20$ and different sample sizes. The fitted line is obtained using a logistic regression. The sample size ensuring 50% chance is 253, 316, 375 respectively for $K = 12, 16, 20$.

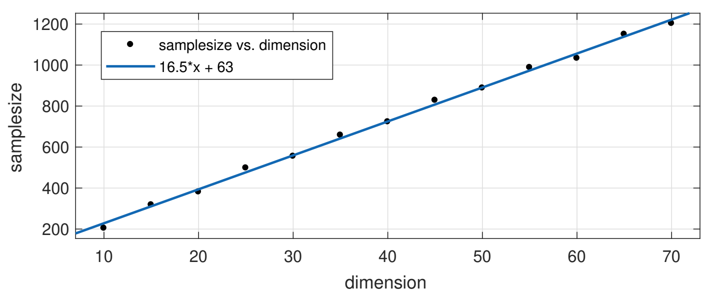

Figure 6: The estimated sample size that achieves 50 percent chance to ensure local identifiability for different K when the reference coefficient is generated from sparse Gaussian distribution and the reference dictionary has constant collinearity.

First we introduce the simulation setting. We generate $n = 100K$ samples using a noisy linear model:

$$\boldsymbol{x}^{(i)} = \boldsymbol{D}^* \boldsymbol{\alpha}^{(i)} + \epsilon^{(i)}, \quad i = 1, \dots, n.$$

The reference dictionary $\boldsymbol{D}^*$, the reference coefficients $\boldsymbol{\alpha}^{(i)}$, and the noise $\epsilon^{(i)}$ are generated as follows.

- Generation of $\boldsymbol{D}^*$: First, we randomly generate a random Gaussian matrix $X \in \mathbb{R}^{K \times K}$ where $X_{jk} \sim \mathcal{N}(0, 1)$. We then scale columns of X to get columns of the reference dictionary $\boldsymbol{D}^*_j = X_j / \|X_j\|_2$ for $j = 1, \dots, K$.

- Generation of $\boldsymbol{\alpha}^{(i)}$: We generate the reference coefficient from sparse Gaussian distribution with sparsity s: $\boldsymbol{\alpha}^{(i)} \sim SG(s)$ for $i = 1, \dots, n$.

- Generation of $\epsilon^{(i)}$: We generate $\epsilon^{(i)}$ using a Gaussian distribution with mean zero. The variance of the distribution is set such that the signal-to–noise ratio is 100:

$$\frac{\mathbb{E}\|\boldsymbol{D}^*\boldsymbol{\alpha}^{(1)}\|_2}{\mathbb{E}\|\epsilon^{(1)}\|_2} = 10^2.$$

We choose the dimension K between 2 and 20 and sparsity s between 2 and K. For each (s, K)-pair, we repeat the experiment 100 times. The accuracy of an estimated dictionary $\hat{\boldsymbol{D}}$ is quantified using the normalized mean square error (NMSE):

$$\text{NMSE}(\hat{\boldsymbol{D}}, \boldsymbol{D}^*) = \min_{J \in \mathcal{J}} \frac{\|\hat{\boldsymbol{D}}J - \boldsymbol{D}^*\|_F^2}{\|\boldsymbol{D}^*\|_F^2},$$

where $\mathcal{J} = \{\Gamma \cdot \Lambda \mid \Gamma$ is a permutation matrix and Λ is a diagonal matrix whose diagonal elements are $\pm 1.\}$ is a set of matrices introduced to resolve the permutation and scale ambiguities. We say an algorithm has a successful recovery if the NMSE of $\hat{\boldsymbol{D}}$ is smaller than the threshold 0.01. We compare different algorithms in terms of their recovery rate, defined as the proportion of simulations that an algorithm has a successful recovery.

The algorithms being tested have several important parameters. For the purpose of comparison, we choose these parameters in a way such that they are consistent with other papers (Parker et al., 2014a,b). The details of parameter settings can be found in Appendix D.

Fig. 7 shows the recovery rate for a variety of choices of dimension K and sparsity s. For each algorithm, the blue region corresponds to (s, K) configurations under which an algorithm has high recovery rate, whereas yellow region indicates low recovery rate. Our results demonstrate that DL-BCD with $\tau = 0.5$ has the best recovery performance compared to other algorithms. We tried $\tau = 0.1, 0.5, 1, 2, 10$, and ∞ but with no further fine tuning. The algorithm EM-BiG-AMP has the second best performance.

We also compare the algorithms in terms of their computation cost. We record the average computation times for $K = 20$ and $s = 10$ (Fig. 8). It can be seen that the SPAMS package is the fastest. The speed of our DL-BCD is roughly the same as that of K-SVD. ER-SpUD is the slowest among all the algorithms.

6. Conclusions and future work

In this paper, we study the theoretical properties of ℓ_1-minimization dictionary learning under complete reference dictionary and noiseless signal assumptions. First, we derive a sufficient and almost necessary condition for local identifiability of ℓ_1-minimization. Our theorems not only extend previous local identifiability results to a much wider class of coefficient distributions, but also give an explicit bound on the region within which the objective value of the reference dictionary is minimal and characterize the sharpness of a local minimum. Secondly, we show that the reference dictionary is the unique sharp local minimum for ℓ_1-minimization. Based on our theoretical results, we design an algorithm to check the sharpness of a local minimum numerically. Finally, We propose the DL-BCD algorithm and demonstrate its competitive performance over other state-of-the-art algorithms in noiseless complete dictionary learning.

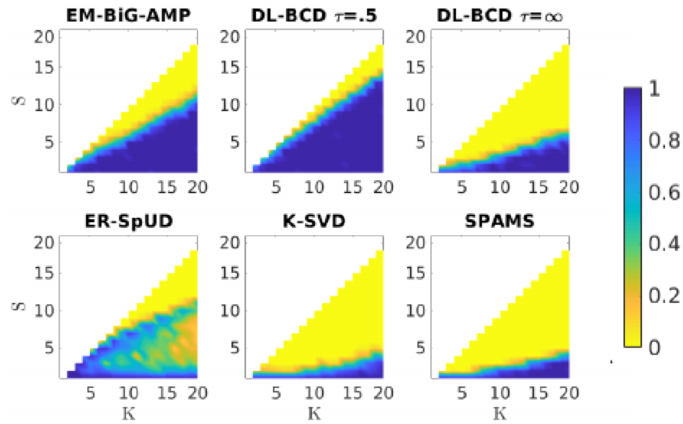

Figure 7: Recovery rate of different algorithms for $K = 2, \ldots, 20$ and $s = 2, \ldots, K$.

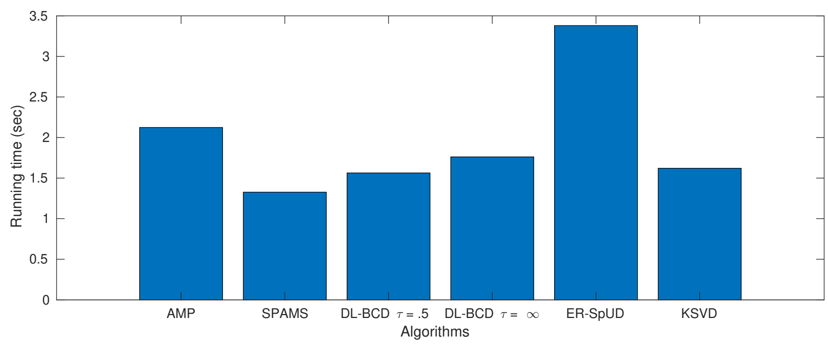

Figure 8: Average running time of different algorithms for $K = 20$ and $s = 10$.

Although we mainly focus on complete dictionaries in this paper, we believe that some of the results can be extended to the over-complete case. The challenge is that the representation of the optimization problem in the complete case (Formulation 5) will become much more complicated as the dictionary is no longer invertible. To deal with this issue, we note that some collections of the columns of the dictionary are invertible and as a result, the problem is now a double minimization $\min_{\boldsymbol{D} \in \mathcal{D}} \mathbb{E} \min_{\boldsymbol{D}' \in \mathbb{R}^{K \times K}, \boldsymbol{D}' \subset \boldsymbol{D}} \|\boldsymbol{D}'^{-1} \boldsymbol{x}\|_1$. Techniques used in compressed sensing (Chen et al., 2001; Fuchs, 2004) and prior works of overcomplete dictionary learning (Geng et al., 2014) can be useful in establishing the generalized results. Besides over-complete settings, it would also be interesting to generalize the result to the noisy case (Gribonval et al., 2015).

7. Acknowledgments

Partial supports are gratefully acknowledged from ARO grant W911NF1710005, ONR grant N00014-16-1-2664, NSF grants DMS-1613002 and IIS 1741340, and the Center for Science of Information (CSoI), a US NSF Science and Technology Center, under grant agreement

CCF-0939370. The authors would like to thank Tanya Veeravalli, Simon Walter, Raaz Dwivedi and Zihao Chen for their very helpful suggestions on the manuscript. We also thank the three referees for their thoughtful and constructive comments that substantially improves the quality of the paper.

8. Appendix A: Additional Examples

Corollary 13 *Let the reference dictionary be a constant collinearity dictionary with coherence μ. Assume that the reference coefficients are generated from the Bernoulli Gaussian model $BG(p)$. If $\frac{\mu\sqrt{p(K-1)}}{1-p} < 1$, the reference dictionary is a sharp local minimum of $\mathbb{E}L(D)$ with sharpness at least $\frac{p}{\sqrt{\pi}(1+\mu(K-1))}\left(1-p-\mu\sqrt{p(K-1)}\right)$. In addition, for any*

$$\boldsymbol{D} \in \Big\{\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)\Big| |||\boldsymbol{D}|||_2 \leq 2\sqrt{1+\mu(K-1)}, \quad |||\boldsymbol{D}-\boldsymbol{D}^*|||_F^2 \leq \frac{1}{8\sqrt{2}(1+\mu(K-1))}\left(1-p-\mu\sqrt{p(K-1)}\right)\Big\},$$

$$\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 \geq \mathbb{E}\|\boldsymbol{\alpha}\|_1.$$

Corollary 14 *Let the reference dictionary be a constant collinearity dictionary with coherence μ. Assume that the coefficients are generated from the sparse Laplacian model $SL(s)$. If*

$$\frac{\mu s(K-1)}{(K-s)\iint_0^\infty |y-x|(xy)^{s-1}\exp(-(x+y))\Gamma(s)^{-2}dxdy} < 1,$$

then the reference dictionary is a sharp local minimum of $\mathbb{E}L(D)$.

Although the condition in Corollary 14 is quite convoluted, we can compare it with the sparse Gaussian case empirically. For sparse Gaussian distributions, there are two parameters: sparsity s and dimension K. Define the phase transition curve to be the asymptotic boundary that separates the region where local identifiability holds (the area below the curve) and the region where local identifiability fails (the area above the curve). When $K = 10$ and 20, the phase transition curve ($|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = 1$) for sparse Laplace distribution and sparse Gaussian distribution can be found in Fig. 9. As can be seen in the figure, the phase transition curve for sparse Laplace distribution is slightly higher than that for sparse Gaussian distribution, suggesting that the Laplace distribution has less stringent local identifiability condition. That is consistent with our intuition: while the density function of a standard Gaussian distribution is rotation symmetric, which implies that it does not prefer any direction, the density function of the Laplace distribution is not. For example, consider a simple two-dimensional case: let $\boldsymbol{D}^*$ be the identity matrix in $\mathbb{R}^{2\times 2}$. If the reference coefficient is from the standard Gaussian distribution with no sparsity, i.e. $s = K$, all the orthogonal dictionaries will have the same objective value $\sqrt{\frac{2}{\pi}}$. So local identifiability does not hold for Gaussian distribution under the setting $s = K$. However, for the Laplace distribution, even if $s = K$, for an orthogonal dictionary $\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ with $\theta \in [0, \pi/2]$, its ℓ_1 objective function value is $2(\sin\theta + \cos\theta + \frac{1}{\sin\theta+\cos\theta})$, which attains its minimum when $\theta = 0$ or $\frac{\pi}{2}$. This means the local identifiability still holds.

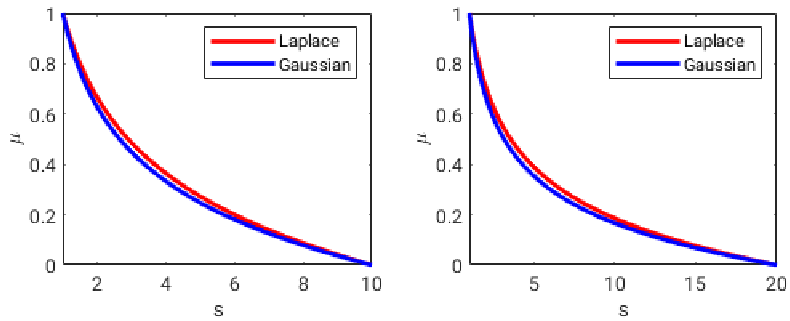

Figure 9: The theoretical phase transition curve for constant collinearity dictionary with coherence μ and sparsity s for $K = 10$ (Left) and $K = 20$ (Right).

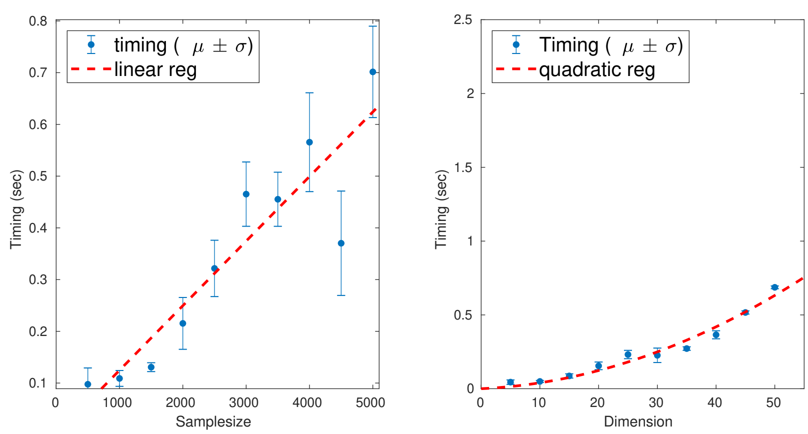

Figure 10: Computation time of Algorithm 1. $p = 0.1$ and $\mu = 0.1$. Left: For $K = 20$ and $n = 500, \ldots, 5000$. Right: For $K = 5, \ldots, 50$ and $n = 400$.

9. Appendix B: Additional Simulations

9.1. Running time complexity

Following the simulation in Section 5.1, we carry out the same simulation for different values of μ and p. Let the reference dictionary be a constant collinearity dictionary with coherence $\mu = 0.1$ and $\mu = 0.9$. The sparse linear coefficients are generated from the Bernoulli Gaussian distribution $BG(p)$ where $p = 0.1$ and $p = 0.9$. The simulation results are shown in Fig. 10 and Fig. 11. We find that for a fixed dimension, the computation time scales roughly linearly with sample size, and for fixed sample size, the computation time scales quadratically with dimension K. That reveals the same trend as in Section 5.1.

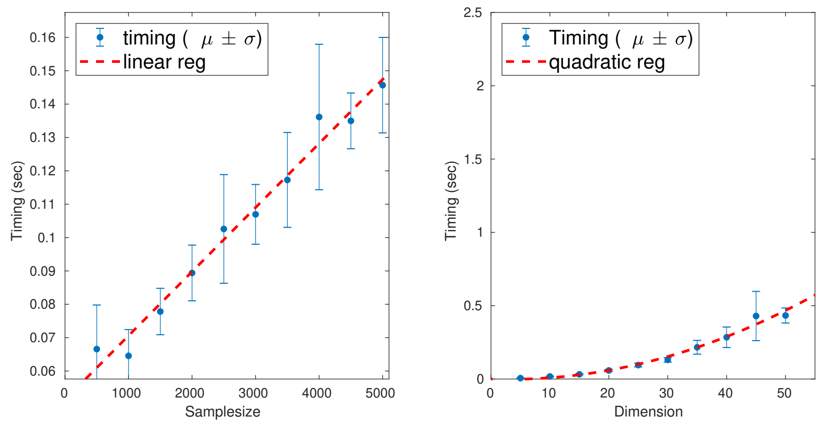

Figure 11: Computation time of Algorithm 1. $p = 0.9$ and $\mu = 0.9$. Left: For $K = 20$ and $n = 500, \ldots, 5000$. Right: For $K = 5, \ldots, 50$ and $n = 400$.

9.2. Sensitivity analysis for ρ

We test the sensitivity of Algorithm 1 by varying the parameter ρ. Let dictionary dimension $K = 20$, sparsity parameter $s = 10$ and sample size $n = 1600$. We consider constant collinearity dictionaries with $\mu = \frac{1}{\sqrt{s}}(\frac{K-s}{K-1} + 0.05)$ (Fig. 12 Left) and $\mu = \frac{1}{\sqrt{s}}(\frac{K-s}{K-1} - 0.1)$ (Fig. 12 Right). For the first experiment, the reference dictionary is not a sharp local minimum of the objective function given large enough samples. Hence a small perturbation to Algorithm 1 will result in a large r defined in the Algorithm 1. Similarly, in the second experiment, the reference dictionary is sharp, indicating r in the Algorithm 1 should be small with respect to the perturbation. The results are in Fig. 12. This experiment shows for parameter ρ values ranging from 0.05 to 0.1, Algorithm 1 succeeds in distinguishing between the sharp and not-sharp local minima.

10. Appendix C: Proofs

10.1. Proofs of Propositions

Proof [Proof of Proposition 2] To prove both models satisfy Assumption I, we just need to prove $|||\cdot|||_{\boldsymbol{\alpha}}$ is lower bounded by $|||\cdot|||_F$ in the linear subspace $H^K = \{A \in \mathbb{R}^{K\times K} | A_{k,k} = 0 \forall\ k\}$. If we can prove $|||\cdot|||_{\boldsymbol{\alpha}}$ is a norm on H^K, then we know it is equivalent to the Frobenious norm since H^K is a finite dimensional space.

In order to show that $|||\cdot|||_{\boldsymbol{\alpha}}$ is a norm, we need to prove three properties:

- Sub-additivity: for any $A, B \in H^K$, $|||A + B|||_{\boldsymbol{\alpha}} \le |||A|||_{\boldsymbol{\alpha}} + |||B|||_{\boldsymbol{\alpha}}$.
- Absolutely homogeneity: for any $A \in H^K$ and $\lambda > 0$, $|||\lambda A|||_{\boldsymbol{\alpha}} = \lambda |||A|||_{\boldsymbol{\alpha}}$.
- Positive definiteness: If $|||A|||_{\boldsymbol{\alpha}} = 0$ and $A \in H^K$, we know $A = 0$.

The first two properties are quite straightforward so we leave the details to readers. Here we focus on proving the third property. Note that $|||A|||_{\boldsymbol{\alpha}}$ is a sum of K non-negative terms, if

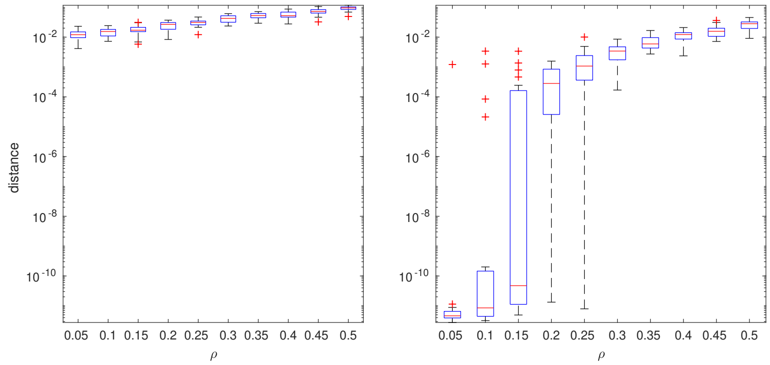

Figure 12: Sensitivity analysis of perturbation parameter ρ in Algorithm 1. Left: constant collinearity dictionary with $\mu = \frac{1}{\sqrt{s}}(\frac{K-s}{K-1} + 0.05)$; Right: constant collinearity dictionary with coherence $\mu = \frac{1}{\sqrt{s}}(\frac{K-s}{K-1} - 0.1)$.

$|||A|||_{\boldsymbol{\alpha}} = 0$, then for any $k \in \{1, \ldots, K\}$, each term should be zero, i.e. $\mathbb{E}|\sum_j A_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0) = 0$. If $\boldsymbol{\alpha}$ is from Bernoulli type models $\mathcal{B}(p_1, \ldots, p_K; f)$, then we could further decompose $\mathbb{E}|\sum_j A_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0) = 0$ into:

$$\mathbb{E}|\sum_j A_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0) = 0 \Leftrightarrow P(\eta_k = 0)\mathbb{E}|\sum_j A_{k,j}\eta_j z_j| = 0 \Leftrightarrow \mathbb{E}|\sum_j A_{k,j}\eta_j z_j| = 0.$$

The second "$\Leftrightarrow$" is because $P(\eta_k = 0) = 1 - p_k > 0$ and $A_{k,k} = 0$. Since $\mathbb{E}|\sum_j A_{k,j}\eta_j z_j| = 0 > P(\eta_1 = \ldots = \eta_K = 1)\mathbb{E}|\sum_j A_{k,j} z_j| \geq 0$ for $p_1, \ldots, p_K \neq 0$, we know $\mathbb{E}|\sum_j A_{k,j} z_j| = 0$. Define $X = \sum_j A_{k,j} z_j$, since $\mathbb{E}|X| = 0$, we know $X = 0$ almost surely. If $A_{j,k}$ are not all zeros, this means $z_1, \ldots, z_K$ are linearly dependent. In other words, z lies in a linear subspace of $\mathbb{R}^K$ almost surely. However, that contradicts the fact that z has a density probability function in $\mathbb{R}^K$. So A must be zero. That completes the proof for Bernoulli type models. For exact sparse models, the approach is essentially the same.

Now for sparse Gaussian and Bernoulli Gaussian distributions, we can obtained the constant $c_{\boldsymbol{\alpha}}$. We first derive the constant for the sparse Gaussian distribution. For $X \in H^K$,

$$\begin{aligned}
\|\|X\|\|_{\boldsymbol{\alpha}} =& \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \binom{K}{s}^{-1} \sum_{\substack{S \subset \{1,\ldots,K\} \\ k \notin S, |S|=s}} \sqrt{\sum_{j \in S} X_{k,j}^2} \\
=& \frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \binom{K-2}{s-1}^{-1} \sum_{\substack{S \subset \{1,\ldots,K\} \\ k \notin S, |S|=s}} \sqrt{\sum_{j \in S} X_{k,j}^2} \\
\text{(Lemma 6.5 in Wu and Yu (2018))} \geq& \frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \sqrt{\sum_{j=1}^{K} X_{k,j}^2} \\
(\|x\|_1 \geq \|x\|_2) \quad \geq& \frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}} \|\|X\|\|_F.
\end{aligned}$$

Here, we need to use Lemma 6.5 in Wu and Yu (2018). For the completeness of this paper, we rewrite the lemma below:

Lemma 6.5 in Wu and Yu (2018) Let $z \in \mathbb{R}^{K-1}$, then for $1 \leq l \leq m \leq K-1$,

$$\binom{K-2}{l-1}^{-1} \sum_{\substack{S \subset \{1,\ldots,K-1\} \\ |S|=l}} \sqrt{\sum_{j \in S} z_j^2} \geq \binom{K-2}{m-1}^{-1} \sum_{\substack{S \subset \{1,\ldots,K-1\} \\ |S|=m}} \sqrt{\sum_{j \in S} z_j^2}.$$

Then the first inequality holds by setting $l = s$ and $m = K-1$. In summary, we have shown that for $X \in H^K$, $\|\|X\|\|_{\boldsymbol{\alpha}} \geq \frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}} \|\|X\|\|_F$, which means $c_{\boldsymbol{\alpha}}$ is at least $\frac{s(K-s)}{K(K-1)} \sqrt{\frac{2}{\pi}}$.

Now, we will compute the constant $c_{\boldsymbol{\alpha}}$ for Bernoulli Gaussian distribution. For $X \in H^K$, if we define $\tilde{s} = \lceil (K-2)p+1 \rceil$,

$$\begin{aligned}
\|\|X\|\|_{\boldsymbol{\alpha}} =& \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \sum_{s=0}^{K-1} \sum_{\substack{S \subset \{1,\ldots,K\} \\ |S|=s, k \notin S}} p^s (1-p)^{K-s} \sqrt{\sum_{j \in S} X_{k,j}^2} \\
\text{(Wu and Yu, 2018, Lemma 6.6)} \geq& (1-p) \sqrt{\frac{2}{\pi}} \sum_{k=1}^{K} \binom{K-1}{\tilde{s}}^{-1} \sum_{\substack{S \subset \{1,\ldots,K\} \\ k \notin S, |S|=\tilde{s}}} \sqrt{\sum_{j \in S} X_{k,j}^2} \\
\text{(Wu and Yu, 2018, Lemma 6.5)} \geq& (1-p) \frac{\lceil (K-2)p+1 \rceil}{K-1} \sqrt{\frac{2}{\pi}} \|\|X\|\|_F \\
\geq& p(1-p) \sqrt{\frac{2}{\pi}} \|\|X\|\|_F.
\end{aligned}$$

Here, we have used Lemma 6.6 in Wu and Yu (2018). We rewrite that Lemma using the notations in our paper as follows:

Lemma 6.6 in Wu and Yu (2018) Let $p \in (0,1)$ and $\tilde{s} = \lceil (K-2)p+1 \rceil$. For any $z \in \mathbb{R}^{K-1}$,

$$\sum_{s=0}^{K-1} \sum_{\substack{S\subset\{1,\dots,K-1\}\\|S|=s}} p^s(1-p)^{K-1-s} \sqrt{\sum_{j\in S} z_j^2} \geq \binom{K-1}{\tilde{s}}^{-1} \sum_{\substack{S\subset\{1,\dots,K-1\}\\|S|=\tilde{s}}} \sqrt{\sum_{j\in S} z_j^2}.$$

In summary, we have shown that for $X \in H^K$, $|||X|||_{\boldsymbol{\alpha}} \geq p(1-p)\sqrt{\frac{2}{\pi}}|||X|||_F$, which means $c_{\boldsymbol{\alpha}}$ is at least $p(1-p)\sqrt{\frac{2}{\pi}}$. ■

Proof [Proof of Proposition 3] In order to prove Assumption II, we only need to show that for any $c_1,\dots,c_K$, $P(\sum_{l=1}^d c_l\boldsymbol{\alpha}_l = 0, \text{ and } \exists\, l,\ c_l\boldsymbol{\alpha}_l \neq 0) = 0$. Note that $\boldsymbol{\alpha}_j = \xi_j z_j$ for $j = 1,\dots,K$,

$$\begin{aligned}
&P(\sum_{l=1}^d c_l\boldsymbol{\alpha}_l = 0, \text{ and } \exists\, l,\ c_l\boldsymbol{\alpha}_l \neq 0)\\
\leq &\sum_{S\subset\{1\dots,K\}} P(\xi_l = 1 \text{ if } l\in S \text{ and } 0 \text{ if } l\notin S)\cdot P(\sum_{l\in S} c_l z_l = 0, \text{ and } \sum_{l\in S} c_l^2 > 0).
\end{aligned}$$

The inequality holds because $\boldsymbol{\alpha}_k = \eta_k \cdot z_k$ for $k = 1,\dots,K$ and η and z are independent for exact sparse models or Bernoulli type models. Since z has a density function, $z_1,\dots,z_K$ are linearly independent, i.e., $P(\sum_{l\in S} c_l z_l = 0, \text{ and } \sum_{l\in S} c_l^2 > 0) = 0$ for any S. ■

10.2. Proofs of Corollaries

Before proving the corollaries, we need the following lemma.

Lemma 15 *If X equals to $c\cdot\mathbf{1}\mathbf{1}^T$, and* $|||A|||_{\boldsymbol{\alpha}} = \sum_{k=1}^K \sqrt{\frac{2}{\pi}}\frac{s}{K}\binom{K-1}{s-1}^{-1} \sum_{\substack{S\subset\{1,\dots,K\}\\k\notin S,|S|=s}} \sqrt{\sum_{j\in S} A_{k,j}^2}$, *then* $|||X|||_{\boldsymbol{\alpha}}^* = \frac{cK(K-1)}{\sqrt{s}(K-s)}\sqrt{\frac{\pi}{2}}$.

Proof [Proof of Lemma 15] Essentially, we are trying to prove that

$$\begin{aligned}
\max_{A\neq 0, A\in H^K} \frac{\mathrm{tr}(A^T X)}{|||A|||_{\boldsymbol{\alpha}}} &= \max_{A\neq 0, A\in H^K} \frac{c\sum_{k=1}^K \sum_{j\neq k} A_{k,j}}{\sum_{k=1}^K \frac{s}{K}\binom{K-1}{s-1}^{-1} \sum_{\substack{S\subset\{1,\dots,K\}\\k\notin S,|S|=s}} \sqrt{\sum_{j\in S} A_{k,j}^2}}\sqrt{\frac{\pi}{2}}\\
&= \frac{cK(K-1)}{\sqrt{s}(K-s)}\sqrt{\frac{\pi}{2}}.
\end{aligned}$$

Note that this is equivalent to the fact that the following convex optimization problem attains the minimum $(K-s)\sqrt{s}$:

$$\min \sum_{k=1}^{K} \frac{s}{K}\binom{K-1}{s-1}^{-1} \sum_{\substack{S\subset\{1,\ldots,K\}\\ k\notin S, |S|=s}} \sqrt{\sum_{j\in S} A_{k,j}^2}$$
$$\text{subject to } \sum_{k=1}^{K}\sum_{j\neq k} A_{k,j} = K(K-1).$$

First of all, note that the problem can be split into K sub-problems: For $k = 1, \ldots, K$,

$$\min \frac{s}{K}\binom{K-1}{s-1}^{-1} \sum_{\substack{S\subset\{1,\ldots,K\}\\ k\notin S, |S|=s}} \sqrt{\sum_{j\in S} A_{k,j}^2}$$
$$\text{subject to } \sum_{j\neq k} A_{k,j} = K-1.$$

Furthermore, note that both the objective and the constraint are permutation symmetric: if $\tilde{A}$ is obtained by permuting off-diagonal elements from each row in A, then the objective function remains the same. It is not hard to show for the optimal solution A^* must satisfy that for any k, $j_1 \neq k$, and $j_2 \neq k$, $A^*_{k,j_1} = A^*_{k,j_2}$. Therefore, $A^*_{k,j} = 1$ and the objective function is $s\binom{K-1}{s-1}^{-1}\binom{K-1}{s}\sqrt{s} = (K-s)\sqrt{s}$. That completes the proof. ■

Proof [Proof of Corollary 6] (local identifiability for constant collinearity reference dictionary and sparse Gaussian coefficients) The coefficients are generated from sparse Gaussian distribution $SG(s)$. First, the collinearity matrix $M^* = (\boldsymbol{D}^*)^T\boldsymbol{D}^* = (1-\mu)\mathbb{I} + \mu\mathbf{1}\mathbf{1}^T$. Because $\boldsymbol{\alpha}$ is sparse Gaussian, we know $\mathbb{E}\boldsymbol{\alpha}_j\text{sign}(\boldsymbol{\alpha}_k) = 0$ for any $j \neq k$ and $\mathbb{E}|\boldsymbol{\alpha}_j| = \sqrt{\frac{2}{\pi}}\frac{s}{K}$. The bias matrix B is

$$(B(\boldsymbol{\alpha}, M^*))_{k,j} = \begin{cases} -M_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j| = -M_{j,k}\sqrt{\frac{2}{\pi}}\frac{s}{K} = -\sqrt{\frac{2}{\pi}}\frac{\mu s}{K} & \text{for } j \neq k \\ \mathbb{E}|\boldsymbol{\alpha}_j| - \mathbb{E}|\boldsymbol{\alpha}_j| = 0 & \text{if } j = k \end{cases}.$$

That means $B(\boldsymbol{\alpha}, M^*)$ is a constant matrix except for the diagonal elements. In the proof of Proposition 2, we showed $|||X|||_{\boldsymbol{\alpha}} = \sqrt{\frac{2}{\pi}}\sum_{k=1}^{K}\frac{s}{K}\binom{K-1}{s-1}^{-1}\sum_{k\notin S,|S|=s}\sqrt{\sum_{j\in S}X_{k,j}^2}$. In general, $|||\cdot|||^*_{\boldsymbol{\alpha}}$ does not have an explicit formula, but for constant matrices, there is a closed form formula (see Lemma 15). Using Lemma 15, we know

$$|||B(\boldsymbol{\alpha}, M^*)|||^*_{\boldsymbol{\alpha}} = \frac{\mu\sqrt{s}(K-1)}{K-s}.$$

Now we will calculate the sharpness constant and the region bound. First of all, $|||\boldsymbol{D}^*|||_2^2 = |||M^*|||_2 = 1 + \mu(K-1)$. Secondly, $|||B(\boldsymbol{\alpha}, M^*)|||^*_{\boldsymbol{\alpha}} = \frac{\mu\sqrt{s}(K-1)}{K-s}$ and $c_{\boldsymbol{\alpha}} \geq \frac{s(K-s)}{K(K-1)}\sqrt{\frac{2}{\pi}}$. Combining those formulas, the sharpness is at least:

$$\frac{1}{\sqrt{\pi}(1+\mu(K-1))}\frac{s}{K}\left(\frac{K-s}{K-1} - \mu\sqrt{s}\right) \approx \frac{s}{\sqrt{\pi}\mu K^2}\left(1-\mu\sqrt{s}\right) \text{ for large } K.$$

For sparse Gaussian distributions, $\max_j \mathbb{E}|\boldsymbol{\alpha}_j| = \sqrt{\frac{2}{\pi}}\frac{s}{K}$, the set S in Theorem 5 is

$$\begin{aligned}
S &= \left\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \middle| |||\boldsymbol{D}|||_2 \le 2|||\boldsymbol{D}^*|||_2, |||\boldsymbol{D}-\boldsymbol{D}^*|||_F \le \frac{(1-|||B(\boldsymbol{\alpha},M^*)|||^*_{\boldsymbol{\alpha}})\cdot c_{\boldsymbol{\alpha}}}{8\sqrt{2}|||\boldsymbol{D}^*|||_2^2 \max_j \mathbb{E}|\boldsymbol{\alpha}_j|} \right\} \\
&= \Big\{ \boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K) \mid |||\boldsymbol{D}|||_2 \le 2\sqrt{1+\mu(K-1)}, \\
&\qquad |||\boldsymbol{D}-\boldsymbol{D}^*|||_F \le \frac{1}{8\sqrt{2}(1+\mu(K-1))}\left(\frac{K-s}{K-1} - \mu\sqrt{s}\right) \Big\},
\end{aligned}$$

which completes the proof. ■

Proof [Proof of Corollary 7] Assume the reference dictionary is a constant collinearity dictionary with coherence μ and the coefficients are generated from non-negative sparse Gaussian distribution $|SG(s)|$. Since $\mathbb{E}\boldsymbol{\alpha}_k \text{sign}(\boldsymbol{\alpha}_j) = \mathbb{E}\eta_k\eta_j z_k = \sqrt{\frac{2}{\pi}}\frac{s(s-1)}{K(K-1)}$ when $j \ne k$, it can be shown that

$$(B(\boldsymbol{\alpha}, M^*))_{k,j} = \begin{cases} -\sqrt{\frac{2}{\pi}}\left(\frac{\mu s}{K} - \frac{s(s-1)}{K(K-1)}\right) & \text{for } j \ne k. \\ 0 & \text{if } j = k. \end{cases}$$

This shows $B(\boldsymbol{\alpha}, M^*)$ is still a constant matrix except the diagonal elements. However, compared with standard sparse Gaussian coefficients, the constant here is $\sqrt{\frac{2}{\pi}}\left(\frac{\mu s}{K} - \frac{s(s-1)}{K(K-1)}\right)$, which is smaller than $\sqrt{\frac{2}{\pi}}\frac{\mu s}{K}$ in Corollary 6. Recall the explanation of the matrix B after Theorem 4, that is because for non-negative sparse Gaussian coefficients, the bias matrix B_1 introduced by the coefficient is of different signs compared to the bias matrix B_2 introduced by the reference dictionary and they cancel with each other. In standard sparse Gaussian case, $B = 0$ if $\mu = 0$, which means the reference dictionary is orthogonal. For this non-negative case, $B = 0$ if $\mu = s/K$, which means the atoms in the reference dictionary should have positive collinearity s/K. As will be shown next, this significantly relaxes the local identifiability condition for non-negative coefficients.

We now compute the closed form formula for the dual semi-norm. By definition, for any matrix X whose elements are all non-negative, $|||X|||_{\boldsymbol{\alpha}} = \sum_{k=1}^K \mathbb{E}|\sum_{j=1}^K X_{k,j}\boldsymbol{\alpha}_j|\mathbf{1}(\boldsymbol{\alpha}_k = 0) = \sum_{k=1}^K\sum_{j=1}^K X_{k,j}\mathbb{E}\boldsymbol{\alpha}_j\mathbf{1}(\boldsymbol{\alpha}_k = 0) = \sqrt{\frac{2}{\pi}}\frac{s(K-s)}{K(K-1)}\sum_{k=1}^K\sum_{j=1,j\ne k}^K X_{j,k}$. Thus we have

$$|||B(\boldsymbol{\alpha}, M^*)|||^*_{\boldsymbol{\alpha}} = \frac{\sqrt{\frac{2}{\pi}}\frac{s}{K}\cdot\left|\mu - \frac{s-1}{K-1}\right|}{\sqrt{\frac{2}{\pi}}\frac{s(K-s)}{K(K-1)}} = \frac{K-1}{K-s}\cdot\left|\mu - \frac{s-1}{K-1}\right|.$$

■

Proof [Proof of Corollary 13] First of all,

$$(B(\boldsymbol{\alpha}, M^*))_{k,j} = \begin{cases} -M_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j| = -M_{j,k}\sqrt{\frac{2}{\pi}}p = -\sqrt{\frac{2}{\pi}}\mu p & \text{for } j \ne k. \\ 0 & \text{if } j = k. \end{cases}$$

Because all the elements in the matrix are constant except the diagonal ones, similar to Lemma 15, we can show the optimal A that attains the maximum of $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = \max \frac{\mathrm{tr}(A^T B(\boldsymbol{\alpha}, M^*))}{|||A|||_{\boldsymbol{\alpha}}}$ is a constant matrix $\mathbf{1}\mathbf{1}^T$. Thus, we have

$$|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = \frac{\mu p(K-1)}{(1-p)\sum_{s=0}^{K-1}\binom{K-1}{s} p^s (1-p)^{K-1-s}\sqrt{s}} \leq \frac{\mu\sqrt{p(K-1)}}{1-p}.$$

Here we are using the Jensen inequality that

$$\sum_{s=0}^{K-1}\binom{K-1}{s} p^s (1-p)^{K-1-s}\sqrt{s} > \sqrt{\sum_{s=0}^{K-1}\binom{K-1}{s} p^s (1-p)^{K-1-s} s} = \sqrt{(K-1)p}.$$

Thus RHS < 1 when μ and p are small. The sharpness is at least

$$\frac{p}{\sqrt{\pi}(1+\mu(K-1))}\left(1 - p - \mu\sqrt{p(K-1)}\right),$$

Because $\mathbb{E}|\boldsymbol{\alpha}_j| = p\sqrt{\frac{2}{\pi}}$ for any j, the set S in Theorem 5 is

$$\Big\{\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)\Big| |||\boldsymbol{D}|||_2 \leq 2\sqrt{1+\mu(K-1)}, \\ |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \leq \frac{1}{8\sqrt{2}(1+\mu(K-1))}\left(1 - p - \mu\sqrt{p(K-1)}\right)\Big\}.$$

■

Proof [Proof of Corollary 14] We compute the local identifiability condition when the reference dictionary is a constant collinearity dictionary with coherence μ and the coefficients are generated from sparse Laplace distribution, i.e., for any j $\boldsymbol{\alpha}_j = \xi_j z_j$ where z_j is from a standard Laplace distribution and ξ is a random 0-1 vector with s nonzeros. For standard Laplace distributions, since $\mathbb{E}|\boldsymbol{\alpha}_j| = \frac{s}{K}$, we have

$$(B(\boldsymbol{\alpha}, M^*))_{k,j} = \begin{cases} -\mu\frac{s}{K} & \text{for } j \neq k. \\ 0 & \text{if } j = k. \end{cases}$$

Similar to Lemma 15, we can show the optimal A that attains the maximum of $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = \max_A \frac{\mathrm{tr}(A^T B(\boldsymbol{\alpha}, M^*))}{|||A|||_{\boldsymbol{\alpha}}}$ is a constant matrix $\mathbf{1}\mathbf{1}^T$.

$$|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* = \frac{\mu s(K-1)}{(K-s)\iint_0^\infty |y-x|(xy)^{s-1}\exp(-(x+y))\Gamma(s)^{-2}dxdy}.$$

To derive this denominator, we need to give an explicit formula for a linear combination of Laplace random variables. The formula can be found in a few papers, e.g., Nadarajah and Kotz (2005). ■

10.3. Proofs of Theorems

The following lemmas are useful for proving Theorem 4.

Lemma 16 *Given two dictionaries $\boldsymbol{D}$ and $\boldsymbol{D}' \in \mathbb{B}(\mathbb{R}^K)$, we have the decomposition:*

$$\boldsymbol{D}^{-1}\boldsymbol{D}' = \mathbb{I} + (\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')) + \Lambda(\boldsymbol{D}, \boldsymbol{D}').$$

where $\Lambda(\boldsymbol{D}, \boldsymbol{D}')$ is a diagonal matrix whose j-th element is $-\frac{1}{2}\|\boldsymbol{D}_j - \boldsymbol{D}'_j\|_2^2$. Then we know

1. *For any $j = 1, \ldots, K$, $M[j,](\boldsymbol{D}^{-1}\boldsymbol{D}'_j - \mathbb{I}_j - \Lambda_j(\boldsymbol{D}, \boldsymbol{D}')) = 0$ where $M = \boldsymbol{D}^T\boldsymbol{D}$.*
2. *$|||\Lambda(\boldsymbol{D})|||_F = \Theta(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2)$:*
$$\frac{1}{2\sqrt{K}}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \le |||\Lambda(\boldsymbol{D})|||_F \le \frac{1}{2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2.$$
3. *When $\langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle \ge 0$ for any $j = 1, \ldots, K$, $|||\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F = \Theta(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F)$:*
$$\frac{|||\boldsymbol{D} - \boldsymbol{D}'|||_F}{\sqrt{2}|||\boldsymbol{D}|||_2} \le |||\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}')|||_F \le |||\boldsymbol{D}^{-1}|||_2 \cdot |||\boldsymbol{D} - \boldsymbol{D}'|||_F$$
4. *Let $M' = (\boldsymbol{D}')^T\boldsymbol{D}'$, for any A satisfying $M'[j,]A_j = 0$ for $j = 1, \ldots, K$ and $|||A|||_F$ sufficiently small, there is a $\boldsymbol{D} \in \mathbb{B}(\mathbb{R}^K)$ such that $\boldsymbol{D}^{-1}\boldsymbol{D}' - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}') = A$.*

Proof [Proof of Lemma 16]

(1):

$$\begin{aligned}
&M[j,](\boldsymbol{D}^{-1}\boldsymbol{D}'_j - \mathbb{I}_j - \Lambda_j(\boldsymbol{D}, \boldsymbol{D}')) \\
=&\langle \boldsymbol{D}_j, \boldsymbol{D}(\boldsymbol{D}^{-1}\boldsymbol{D}'_j - \mathbb{I}_j - \Lambda_j(\boldsymbol{D}, \boldsymbol{D}'))\rangle \\
=&\langle \boldsymbol{D}_j, \boldsymbol{D}'_j - \boldsymbol{D}_j + \frac{1}{2}\boldsymbol{D}_j\|\boldsymbol{D}_j - \boldsymbol{D}'_j\|_2^2\rangle \\
=&\langle \boldsymbol{D}_j, \boldsymbol{D}'_j - \boldsymbol{D}_j\rangle + \frac{1}{2}\|\boldsymbol{D}_j - \boldsymbol{D}'_j\|_2^2 \\
=&\langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle - 1 + 1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j\rangle = 0.
\end{aligned}$$

(2): $|||\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)|||_F = \frac{1}{2}\sqrt{\sum_j \|\boldsymbol{D}_j - \boldsymbol{D}^*_j\|_2^4} \le \frac{1}{2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2$. On the other hand, because of the power inequality $\|x\|_2 \ge \frac{1}{\sqrt{K}}\|x\|_1$, we have

$$|||\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)|||_F = \frac{1}{2}\sqrt{\sum_j \|\boldsymbol{D}_j - \boldsymbol{D}^*_j\|_2^4} \ge \frac{1}{2\sqrt{K}}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2.$$

(3): Firstly, consider $|||\boldsymbol{D'} - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_F$, we have

$$\begin{aligned} |||\boldsymbol{D'} - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_F^2 =& \sum_{j=1}^{K} \|\boldsymbol{D}'_j - \boldsymbol{D}_j \langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle \|_2^2 \\ =& \sum_{j=1}^{K} 1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle^2 \\ =& \sum_{j=1}^{K} \min_{t_j \in \mathbb{R}} \|\boldsymbol{D}'_j - t_j \cdot \boldsymbol{D}_j\|_2^2. \end{aligned}$$

Then by taking $t_j = 1$ for all $j = 1, \ldots, K$, we have

$$|||\boldsymbol{D'} - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_F^2 \leq |||\boldsymbol{D'} - \boldsymbol{D}|||_F^2.$$

On the other hand, when $\langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle \geq 0$.

$$\sum_{j=1}^{K} 1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle^2 = \sum_{j=1}^{K} (1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle)(1 + \langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle) \geq \sum_{j=1}^{K} (1 - \langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle) = \frac{1}{2} |||\boldsymbol{D} - \boldsymbol{D'}|||_F^2.$$

Then for $(\boldsymbol{D}^{-1}\boldsymbol{D'} - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D'}))$, using the above inequalities, we have:

$$\begin{aligned} |||\boldsymbol{D'} - \boldsymbol{D}|||_F \leq& \sqrt{2} |||\boldsymbol{D'} - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_F \\ \leq& \sqrt{2} |||\boldsymbol{D}|||_2 |||\boldsymbol{D}^{-1}\boldsymbol{D'} - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_F, \end{aligned}$$

which proves the first inequality. The second inequality follows from

$$\begin{aligned} & |||\boldsymbol{D}^{-1}\boldsymbol{D'} - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_F \\ \leq& |||\boldsymbol{D}^{-1}|||_2 |||\boldsymbol{D'} - \boldsymbol{D} - \boldsymbol{D}\Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_F \\ \leq& |||\boldsymbol{D}^{-1}|||_2 |||\boldsymbol{D'} - \boldsymbol{D}|||_F. \end{aligned}$$

(4): Consider a differentiable mapping $F(\boldsymbol{D}) = \boldsymbol{D}^{-1}\boldsymbol{D'} - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D'})$ from $\mathbb{B}(\mathbb{R}^K)$ to a linear manifold

$$H = \{A \in \mathbb{R}^{K \times K} \Big| M'[j,]A_j = 0 \text{ for any } j = 1, \ldots, K.\}$$

Since $F(\boldsymbol{D'}) = \boldsymbol{0}$, if we can prove the differential of F at $\boldsymbol{D'}$, namely dF, is bijective from the tangent space $T\mathbb{B}(\mathbb{R}^K)\Big|_{\boldsymbol{D'}} = \{A \in \mathbb{R}^{K \times K} \Big| \langle D'_j, A_j \rangle = 0 \text{ for any } j = 1, \ldots, K.\}$ to the tangent space $TH\Big|_0 = H$, then by the inverse function theorem on the manifold, we have the conclusion. To prove it is indeed bijective, we note that $dF(\Delta)\Big|_{\boldsymbol{D'}}$ is $\sum_{k,j} (\boldsymbol{D'})_j^{-1} \mathbb{I}[k,]\Delta_{j,k} = (\boldsymbol{D'})^{-1}\Delta$. Clearly dF is injective: $(\boldsymbol{D'})^{-1}\Delta = 0$ implies $\Delta = 0$. To show it is also surjective, first of all for any $\Delta \in T\mathbb{B}(\mathbb{R}^K)\Big|_{\boldsymbol{D'}}$, its image under dF is in H:

$$\begin{aligned} M'[j,](\boldsymbol{D'})^{-1}\Delta_j =& \langle \boldsymbol{D}'_j, \boldsymbol{D'}((\boldsymbol{D'})^{-1}\Delta_j \rangle \\ =& \langle \boldsymbol{D}'_j, \Delta_j \rangle = 0. \end{aligned}$$

Because these two linear manifolds have the same dimension, dF must be one-on-one. This concludes the proof. ∎

Lemma 17 *If $|||\cdot|||_{\boldsymbol{\alpha}}$ is regular with constant $c_{\boldsymbol{\alpha}}$, then we know for any $\boldsymbol{D}, \boldsymbol{D'}$ such that $\langle \boldsymbol{D}_j, \boldsymbol{D}'_j \rangle \geq 0$ for any $j = 1, \ldots, K$, $|||(\boldsymbol{D})^{-1}\boldsymbol{D'}|||_{\boldsymbol{\alpha}} \geq \frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}|||_2^2}|||\boldsymbol{D} - \boldsymbol{D'}|||_F$.*

Proof First of all, because for any $A \in \mathbb{R}^{K\times K}$, by definition of $|||\cdot|||_{\boldsymbol{\alpha}}$, $|||A|||_{\boldsymbol{\alpha}}$ does not depend on diagonal elements $A_{j,j}$ for any $j = 1, \ldots, K$. Thus, $|||(\boldsymbol{D})^{-1}\boldsymbol{D'}|||_{\boldsymbol{\alpha}} = |||(\boldsymbol{D})^{-1}\boldsymbol{D'} - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_{\boldsymbol{\alpha}}$, where Λ is defined in Lemma 16. If we denote A as $(\boldsymbol{D})^{-1}\boldsymbol{D'} - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D'})$, then Lemma 16 shows $M[j,]A_j = 0$. Since $M_{j,j} = 1$, $A_{j,j} = -M[j,-j]A[-j,j]$. Thus $\|A_j\|_2^2 \leq (M[j,-j]A[-j,j])^2 + \|A[-j,j]\|_2^2 \leq (\|M[j,-j]\|_2^2 + 1)\|A[-j,j]\|_2^2 = \|M[j,]\|_2^2\|A[-j,j]\|_2^2$. Summing over j, we have

$$|||A|||_F \leq \max_j \|M[j,]\|_2 \sqrt{\sum_j \|A[-j,j]\|_2^2}.$$

Note that for any j, $\|M[j,]\|_2 = \|\boldsymbol{D}_j^T\boldsymbol{D}\|_2 \leq |||\boldsymbol{D}|||_2$, thus we have:

$$|||A|||_F \leq |||\boldsymbol{D}|||_2 \sqrt{\sum_j \|A[-j,j]\|_2^2}.$$

On the other hand, by Lemma 16, we know $|||A|||_F \geq \frac{1}{\sqrt{2}|||\boldsymbol{D}|||_2}|||\boldsymbol{D} - \boldsymbol{D'}|||_F$. Combining those together, we have

$$\begin{aligned}
|||(\boldsymbol{D})^{-1}\boldsymbol{D'}|||_{\boldsymbol{\alpha}} =& |||(\boldsymbol{D})^{-1}\boldsymbol{D'} - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D'})|||_{\boldsymbol{\alpha}} \\
\text{(Because of Assumption I))} \geq& c_{\boldsymbol{\alpha}} \sqrt{\sum_j \|A[-j,j]\|_2^2} \\
\geq& c_{\boldsymbol{\alpha}} \frac{|||A|||_F}{|||\boldsymbol{D}|||_2} \\
\geq& c_{\boldsymbol{\alpha}} \frac{|||\boldsymbol{D} - \boldsymbol{D'}|||_F}{\sqrt{2}|||\boldsymbol{D}|||_2^2}.
\end{aligned}$$

∎

Lemma 18 *For $x, y \in \mathbb{R}$, $y \cdot \text{sign}(x) + |x| + |y| \cdot \mathbf{1}(x = 0) \leq |y + x| \leq y \cdot \text{sign}(x) + |x| + |y| \cdot \mathbf{1}(x = 0) + 2|y| \cdot \mathbf{1}(|y| > |x| > 0)$.*

Proof If $x = 0$, the above inequality definitely holds. So without loss of generality, let's assume $x \neq 0$. When $|y| < |x|$, $\text{sign}(x + y) = \text{sign}(x)$, so $|y + x| = \text{sign}(x)(x + y) = |x| + \text{sign}(x)y$. When $|y| > |x|$, $\text{sign}(x + y) = \text{sign}(y)$, if $\text{sign}(x) = \text{sign}(y)$, clearly we have $|x + y| = |x| + |y| = y\text{sign}(x) + |x|$. If $\text{sign}(x) \neq \text{sign}(y)$, $|y + x| = |y| - |x| > |x| - |y| = |x| + y\text{sign}(x)$. So in summary, we prove the first inequality. The second inequality comes from: $|y + x| \leq |y| + |x| \leq |x| + 2|y| + y\text{sign}(x)$, which completes the proof. ∎

Lemma 19 *We have the upper and lower bound of the objective function:*

$$\begin{aligned}
&\mathbb{E}\|(\boldsymbol{D}^*)^{-1}\boldsymbol{x}\|_1 + |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) + o(|||\boldsymbol{D}-\boldsymbol{D}^*|||_F) \\
&\geq \mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 \geq \\
&\mathbb{E}\|(\boldsymbol{D}^*)^{-1}\boldsymbol{x}\|_1 + |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1
\end{aligned}$$

Proof [Proof of Lemma 19] By Lemma 16, $(\boldsymbol{D})^{-1}\boldsymbol{D}^*$ can be decomposed into

$$\boldsymbol{D}^{-1}\boldsymbol{D}^* = \mathbb{I} + \Delta(\boldsymbol{D}, \boldsymbol{D}^*) + \Lambda(\boldsymbol{D}, \boldsymbol{D}^*),$$

where $\Delta(\boldsymbol{D}, \boldsymbol{D}^*) = \boldsymbol{D}^{-1}\boldsymbol{D}^* - \mathbb{I} - \Lambda(\boldsymbol{D}, \boldsymbol{D}^*)$ and $\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)$ is defined in Lemma 16. In what follows, we use Λ, Δ without writing $\boldsymbol{D}, \boldsymbol{D}^*$ explicitly for notation ease.

Let $\Delta_{k,j}$ be the element of Δ at k-th row and j-th column. Then the objective function can be lower bounded by:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 &= \mathbb{E}\|(\boldsymbol{D}^*)^{-1}\boldsymbol{x} - (\mathbb{I} - \boldsymbol{D}^{-1}\boldsymbol{D}^*)(\boldsymbol{D}^*)^{-1}\boldsymbol{x}\|_1 \\
&= \mathbb{E}\|\boldsymbol{\alpha} + (\Delta + \Lambda)\boldsymbol{\alpha}\|_1 \\
(a)\ &\geq \mathbb{E}\|\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}\|_1 - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1 \\
(b)\ &\geq \mathbb{E}\sum_{k=1}^{K} |\boldsymbol{\alpha}_k| + \mathbf{1}(\boldsymbol{\alpha}_k = 0)|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| - \mathrm{sign}\boldsymbol{\alpha}_k \sum_j \Delta_{k,j}\boldsymbol{\alpha}_j - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1 \\
&\geq \mathbb{E}\|\boldsymbol{\alpha}\|_1 + |||\Delta|||_{\boldsymbol{\alpha}} - \mathbb{E}\sum_{k,j} \Delta_{k,j}\mathbb{E}\boldsymbol{\alpha}_j \mathrm{sign}\boldsymbol{\alpha}_k - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1.
\end{aligned}$$

(a) holds because of the triangle inequality. (b) holds because of Lemma 18 (let $x = \Delta[k,]\boldsymbol{\alpha}$ and $y = \boldsymbol{\alpha}_k$). Note that by the definition of $|||\cdot|||_{\boldsymbol{\alpha}}$, the diagonal elements of Δ do not matter, so $|||\Delta|||_{\boldsymbol{\alpha}} = |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}$.

Recall $M_{j,k} = \langle \boldsymbol{D}_j, \boldsymbol{D}_k \rangle$, by Lemma 16, $\Delta_{k,j}$ satisfies: $M[j,]\Delta_j = \sum_{k \neq j} M_{j,k}\Delta_{k,j} + \Delta_{j,j} = 0$ (Because $M_{j,j} = 1$) for any j. Thus we have

$$\begin{aligned}
\sum_{j,k=1}^{K} \Delta_{k,j}\mathbb{E}\boldsymbol{\alpha}_j \mathrm{sign}\boldsymbol{\alpha}_k &= \sum_{j=1}^{K}\left(\sum_{k\neq j} \Delta_{k,j}\mathbb{E}\boldsymbol{\alpha}_j\mathrm{sign}\boldsymbol{\alpha}_k + \Delta_{j,j}\mathbb{E}|\boldsymbol{\alpha}_j|\right) \\
&= \sum_{j=1}^{K}\sum_{k\neq j} \Delta_{k,j}\left(\mathbb{E}\boldsymbol{\alpha}_j\mathrm{sign}\boldsymbol{\alpha}_k - M_{j,k}\mathbb{E}|\boldsymbol{\alpha}_j|\right) = \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\Delta).
\end{aligned}$$

Because the diagonal elements of $B(\boldsymbol{\alpha}, M)$ are all zeros, we know

$$\mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\Delta) = \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*).$$

In summary, we have shown that

$$\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 \geq \mathbb{E}\|\boldsymbol{\alpha}\|_1 + |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) - \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1.$$

In order to have an upper bound, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 =& \mathbb{E}\|(\boldsymbol{D}^*)^{-1}\boldsymbol{x} - (\mathbb{I} - \boldsymbol{D}^{-1}\boldsymbol{D}^*)(\boldsymbol{D}^*)^{-1}\boldsymbol{x}\|_1 \\
=& \mathbb{E}\|\boldsymbol{\alpha} + (\Delta + \Lambda)\boldsymbol{\alpha}\|_1 \\
\leq& \mathbb{E}\|\boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}\|_1 + \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1 \\
\text{(Lemma 18)} \leq& \mathbb{E}\|\boldsymbol{\alpha}\|_1 + |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) \\
&+ \sum_k 2\mathbb{E}\left|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j\right| \mathbf{1}\left(\left|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j\right| > |\boldsymbol{\alpha}_k| > 0\right) + \mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1.
\end{aligned}$$

Note that by Lemma 16, $\mathbb{E}\|\Lambda\boldsymbol{\alpha}\|_1 \leq |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \max_j \mathbb{E}|\boldsymbol{\alpha}_j| = o(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F)$. Furthermore,

$$\begin{aligned}
&\mathbb{E}\left|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j\right| \mathbf{1}\left(\left|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j\right| > |\boldsymbol{\alpha}_k| > 0\right) \qquad (10) \\
\leq& \sum_{k=1}^K \max_j |\Delta_{k,j}| \cdot \mathbb{E}\ \mathbf{1}(\boldsymbol{\alpha}_k \neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| \geq |\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1.
\end{aligned}$$

Because $\mathbf{1}(\boldsymbol{\alpha}_k \neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| \geq |\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1 \leq \|\boldsymbol{\alpha}\|_1$, $\mathbb{E}\|\boldsymbol{\alpha}\|_1 < \infty$, and

$$\lim_{\Delta_{k,j}\to 0} \mathbf{1}(\boldsymbol{\alpha}_k \neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| \geq |\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1 = 0 \quad a.s.,$$

by the dominant convergence theorem, we know

$$\begin{aligned}
&\lim_{\Delta\to 0} \mathbb{E}\mathbf{1}(\boldsymbol{\alpha}_k \neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| \geq |\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1 \\
=& \mathbb{E}\lim_{\Delta\to 0} \mathbf{1}(\boldsymbol{\alpha}_k \neq 0)\mathbf{1}(|\sum_j \Delta_{k,j}\boldsymbol{\alpha}_j| \geq |\boldsymbol{\alpha}_k|)\|\boldsymbol{\alpha}\|_1 \\
=& 0.
\end{aligned}$$

This means (10) is $o(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F)$, which proves the upper bound. ■

Proof [Proof of Theorem 4]

(i): We will first prove that if $|||\cdot|||_{\boldsymbol{\alpha}}$ is regular with constant $c_{\boldsymbol{\alpha}}$ and (6) holds, $\boldsymbol{D}^*$ is a sharp local minimum. When (6) is satisfied and $\boldsymbol{D} \to \boldsymbol{D}^*$, $|||B(\boldsymbol{\alpha}, M)|||_{\boldsymbol{\alpha}}^* \to |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* < 1$ and

$$\begin{aligned}
&|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) \\
=& |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M^*)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) + o(|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}) \\
\geq& (1 - |||B(\boldsymbol{\alpha}, M)|||_{\boldsymbol{\alpha}}^*)|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} + o(|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}}).
\end{aligned}$$

Because $|||\cdot|||_{\boldsymbol{\alpha}}$ is regular and Lemma 17, by appropriately choosing signs of each column in $\boldsymbol{D}^*$, we have

$$|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} \geq \frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}|||_2^2}|||\boldsymbol{D}^* - \boldsymbol{D}|||_F.$$

Combine those two inequalities, when $|||\boldsymbol{D} - \boldsymbol{D}^*|||_F$ is small enough,

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 - \mathbb{E}\|\boldsymbol{\alpha}\|_1 \\ \geq &(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*)\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}\cdot|||\boldsymbol{D}^*|||_2^2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F + o(|||\boldsymbol{D} - \boldsymbol{D}^*|||_F).\end{aligned}$$

By Definition 1, $\boldsymbol{D}^*$ is a sharp local minimum with sharpness at least

$$(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*)\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}^*|||_2^2}.$$

(ii) When (6) does not hold or $|||\cdot|||_{\boldsymbol{\alpha}}$ is not regular, $\boldsymbol{D}^*$ is not a sharp local minimum.

If $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* \geq 1$, then there exists Δ' such that $|||\Delta'|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta') \leq 0$. Note that the left hand side does not depend on diagonal elements of Δ', so we can find a matrix Δ that is the same as Δ' except the diagonal elements such that $M^*[j,]\Delta_j = 0$ for any j and $|||\Delta|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta) \leq 0$. For any $t > 0$, by Lemma 16 we can construct a series of dictionaries $\boldsymbol{D}(t)$ for a sufficiently small t such that

$$(\boldsymbol{D}(t))^{-1}\boldsymbol{D}^* = \mathbb{I} + \Delta + \Lambda(\boldsymbol{D}(t), \boldsymbol{D}^*).$$

Then by Lemma 19, we have the formula for the objective of $\boldsymbol{D}(t)$:

$$\mathbb{E}\|\boldsymbol{D}(t)^{-1}\boldsymbol{x}\|_1 = \mathbb{E}\|\boldsymbol{\alpha}\|_1 + \big(|||\Delta|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta)\big) + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F).$$

Because $|||\Delta|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta) \leq 0$, $\mathbb{E}\|\boldsymbol{D}(t)^{-1}\boldsymbol{x}\|_1 \leq \mathbb{E}\|\boldsymbol{\alpha}\|_1 + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F)$. By definition, $\boldsymbol{D}^*$ is not a sharp local minimum. If $|||\cdot|||_{\boldsymbol{\alpha}}$ is not regular, for any $c > 0$, there exists Δ such that $M^*[j,]\Delta_j = 0$ for any j and $|||\Delta|||_{\boldsymbol{\alpha}} < c\|\Delta\|_F$. Without loss of generality, assume $\mathrm{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta) \geq 0$, otherwise just take $-\Delta$. For sufficiently small t, there exists a dictionary $\boldsymbol{D}(t)$ such that

$$(\boldsymbol{D}(t))^{-1}\boldsymbol{D}^* = \mathbb{I} + \Delta + \Lambda(\boldsymbol{D}(t), \boldsymbol{D}^*).$$

Then by Lemma 19, we have the formula for the objective of $\boldsymbol{D}(t)$:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{D}(t)^{-1}\boldsymbol{x}\|_1 = &\mathbb{E}\|\boldsymbol{\alpha}\|_1 + \big(|||\Delta|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M^*)^T\Delta)\big) + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F) \\ \leq &\mathbb{E}\|\boldsymbol{\alpha}\|_1 + c|||\Delta|||_F + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F) \\ \leq &\mathbb{E}\|\boldsymbol{\alpha}\|_1 + c|||\boldsymbol{D}(t)^{-1}|||_2 \cdot |||\boldsymbol{D}(t) - \boldsymbol{D}^*|||_F + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F)\end{aligned}$$

Because that holds for any $c > 0$, by definition, we have shown $\boldsymbol{D}^*$ is not a sharp local minimum.

(iii): When $|||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* > 1$, $\boldsymbol{D}^*$ is not a local minimum. This part is essentially the same as (ii). The key is to construct a series of dictionaries $\boldsymbol{D}(t)$ using Lemma 16 as in (ii).

Then by using the upper bound in Lemma 19, we can find a small $t > 0$ and a small $c > 0$ such that

$$\mathbb{E}\|\boldsymbol{D}_t^{-1}\boldsymbol{x}\|_1 \le \mathbb{E}\|\boldsymbol{\alpha}\|_1 - c\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F + o(\|\boldsymbol{D}(t) - \boldsymbol{D}^*\|_F).$$

Thus by definition $\boldsymbol{D}^*$ is not a local minimum. ∎

Proof [Proof of Theorem 5] Note that by the definition of $\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)$ as in Lemma 16, we have

$$\mathbb{E}\|\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)\boldsymbol{\alpha}\|_1 \le \max_j \mathbb{E}|\boldsymbol{\alpha}_j| |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2$$

On the other hand, by Lemma 19, we know

$$\mathbb{E}\|\boldsymbol{D}^{-1}\boldsymbol{x}\|_1 - \mathbb{E}\|\boldsymbol{\alpha}\|_1 \ge |||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) - \mathbb{E}\|\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)\boldsymbol{\alpha}\|_1$$

Similar to the proof of Theorem 4, the right hand side is bounded by

$$\begin{aligned}
&|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathrm{tr}(B(\boldsymbol{\alpha}, M)^T\boldsymbol{D}^{-1}\boldsymbol{D}^*) - \mathbb{E}\|\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)\boldsymbol{\alpha}\|_1 \\
\ge&(1 - |||B(\boldsymbol{\alpha}, M)|||_{\boldsymbol{\alpha}}^*)|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \mathbb{E}\|\Lambda(\boldsymbol{D}, \boldsymbol{D}^*)\boldsymbol{\alpha}\|_1 \\
\ge&(1 - |||B(\boldsymbol{\alpha}, M)|||_{\boldsymbol{\alpha}}^*)|||\boldsymbol{D}^{-1}\boldsymbol{D}^*|||_{\boldsymbol{\alpha}} - \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \\
\ge&(1 - |||B(\boldsymbol{\alpha}, M)|||_{\boldsymbol{\alpha}}^*)\frac{c_{\boldsymbol{\alpha}}}{\sqrt{2}|||\boldsymbol{D}|||_2^2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F - \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \\
\ge&(1 - |||B(\boldsymbol{\alpha}, M)|||_{\boldsymbol{\alpha}}^*)\frac{c_{\boldsymbol{\alpha}}}{4\sqrt{2}|||\boldsymbol{D}^*|||_2^2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F - \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2. \qquad (11)
\end{aligned}$$

Because $|||M - M^*|||_F \le (|||\boldsymbol{D}|||_2 + |||\boldsymbol{D}^*|||_2) \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F \le 3|||\boldsymbol{D}^*|||_2 \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F$ and $|||\boldsymbol{D} - \boldsymbol{D}^*|||_F \le \frac{c_{\boldsymbol{\alpha}}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)}{8\sqrt{2}\max_j \mathbb{E}|\boldsymbol{\alpha}_j||||\boldsymbol{D}^*|||_2^2}$ we know $|||M - M^*|||_F \le \frac{c_{\boldsymbol{\alpha}}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)}{2\max_j \mathbb{E}|\boldsymbol{\alpha}_j||||\boldsymbol{D}^*|||_2} \le \frac{c_{\boldsymbol{\alpha}}(1-|||B(\boldsymbol{\alpha},M^*)|||_{\boldsymbol{\alpha}}^*)}{2\max_j \mathbb{E}|\boldsymbol{\alpha}_j|}$. The last inequality is because $|||\boldsymbol{D}^*|||_2 \ge 1$. Based on this chain of inequalities, we have:

$$\begin{aligned}
1 - |||B(\boldsymbol{\alpha}, M)|||_{\boldsymbol{\alpha}}^* \ge&1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* - ||||B(\boldsymbol{\alpha}, M)|||_{\boldsymbol{\alpha}}^* - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*| \\
\ge&1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* - |||B(\boldsymbol{\alpha}, M) - B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* \\
\ge&1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* - \frac{1}{c_{\boldsymbol{\alpha}}}|||B(\boldsymbol{\alpha}, M) - B(\boldsymbol{\alpha}, M^*)|||_F \\
\ge&1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^* - \frac{1}{c_{\boldsymbol{\alpha}}}\max_j \mathbb{E}|\boldsymbol{\alpha}_j| \cdot |||M - M^*|||_F \\
\ge&\frac{1}{2}(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*).
\end{aligned}$$

Based on this, (11) is bounded by:

$$\begin{aligned}
&\frac{1}{2}(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*)\frac{c_{\boldsymbol{\alpha}}}{4\sqrt{2}|||\boldsymbol{D}^*|||_2^2}|||\boldsymbol{D} - \boldsymbol{D}^*|||_F - \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \cdot |||\boldsymbol{D} - \boldsymbol{D}^*|||_F^2 \\
\ge&\left(\frac{c_{\boldsymbol{\alpha}}(1 - |||B(\boldsymbol{\alpha}, M^*)|||_{\boldsymbol{\alpha}}^*)}{8\sqrt{2}\max_j \mathbb{E}|\boldsymbol{\alpha}_j||||\boldsymbol{D}^*|||_2^2} - |||\boldsymbol{D} - \boldsymbol{D}^*|||_F\right)|||\boldsymbol{D} - \boldsymbol{D}^*|||_F \max_j \mathbb{E}|\boldsymbol{\alpha}_j| \ge 0.
\end{aligned}$$

This shows the LHS is positive when $\boldsymbol{D} \neq \boldsymbol{D}^*$ and we have completed the proof. ■

Proof [Proof of Theorem 8] In order to prove Theorem 8, it suffices to prove any dictionary $\boldsymbol{D}$ in $\mathbb{B}(\mathbb{R}^K)$ other than $\boldsymbol{D}^*$ will not be a sharp local minimum. Recall $\boldsymbol{\beta}(\boldsymbol{D})$ is the coefficient of the samples under dictionary $\boldsymbol{D}$, i.e., $\boldsymbol{\beta}(\boldsymbol{D}) = \boldsymbol{D}^{-1}\boldsymbol{x}$. For notation ease, we omit $\boldsymbol{D}$ and simply write $\boldsymbol{\beta}$.

The following lemma provides a necessary condition for a dictionary to be a sharp local minimum.

Lemma 20 *For any dictionary $\boldsymbol{D}$, if $\boldsymbol{D}$ is a sharp local minimum of optimization form* (5), *then for any $k = 1, \dots, K$, $\boldsymbol{\beta} \cdot \mathbf{1}(\boldsymbol{\beta}_k = 0)$ does not lie in any linear subspace of dimension $K - 2$.*

Proof [Proof of Lemma 20] If $\boldsymbol{D}$ is a sharp local minimum, by the proof of Theorem 4, it should satisfy (12).

$$\sum_{j,k} \Delta_{k,j}(\mathbb{E}\boldsymbol{\beta}_j \text{sign}(\boldsymbol{\beta}_k) - M_{j,k}\mathbb{E}|\boldsymbol{\beta}_j|) < \sum_k \mathbb{E}|\sum_j \Delta_{k,j}\boldsymbol{\beta}_j|\mathbf{1}(\boldsymbol{\beta}_k = 0). \tag{12}$$

For any $\Delta_{k,j}$, let $\Delta'_{k,j} \triangleq -\Delta_{k,j}$, it should also satisfy (12). That makes

$$-\sum_{j,k} \Delta_{k,j}(\mathbb{E}\boldsymbol{\beta}_j \text{sign}(\boldsymbol{\beta}_k) - M_{j,k}\mathbb{E}|\boldsymbol{\beta}_j|) < \sum_k \mathbb{E}|\sum_j \Delta_{k,j}\boldsymbol{\beta}_j|\mathbf{1}(\boldsymbol{\beta}_k = 0).$$

Thus we have

$$\mathbb{E}|\sum_{j=1, j\neq k}^{K} \Delta_{k,j}\boldsymbol{\beta}_j|\mathbf{1}(\boldsymbol{\beta}_k = 0) > 0. \tag{13}$$

If $\boldsymbol{\beta}\mathbf{1}(\boldsymbol{\beta}_k = 0)$ lies in a linear subspace of dimension $K-2$, because there are $K-1$ free parameters in $\Delta_{j,k}$ for $j \neq k$, we can find a set of nonzero $\Delta_{j,k}$ such that $\sum_{j=1,j\neq k}^{K} \Delta_{k,j}\boldsymbol{\beta}_j \cdot \mathbf{1}(\boldsymbol{\beta}_k = 0) = 0$ a.s.. That contradicts (13). Therefore, $\boldsymbol{\beta}\mathbf{1}(\boldsymbol{\beta}_k = 0)$ does not lie in any linear subspace of dimension $K - 2$. ■

In order to show $\boldsymbol{D} \neq \boldsymbol{D}^*$ up to sign-permutation is not a sharp local minimum, by Lemma 20, it suffices to find a k such that the random vector $\boldsymbol{\beta} \cdot \mathbf{1}(\boldsymbol{\beta}_k = 0)$ lies in a linear manifold of dimension at most $K - 2$.

Note that $\boldsymbol{\beta} = \boldsymbol{D}^{-1}\boldsymbol{D}^*\boldsymbol{\alpha}$ is linear transform of $\boldsymbol{\alpha}$. For $\boldsymbol{D} \neq \boldsymbol{D}^*$ up to the sign-permutation sense, $\boldsymbol{D}^{-1}\boldsymbol{D}^* \neq \mathbb{I}$, which means there exists k such that $\boldsymbol{\beta}_k \neq \boldsymbol{\alpha}_{k'}$ for any $k' = 1, \dots, K$. This means $\boldsymbol{\beta}_k$ is the linear combination of at least two elements in $\boldsymbol{\alpha}$. Without loss of generality, $\boldsymbol{\beta}_k = \sum_{l=1}^{T} c_l\boldsymbol{\alpha}_l$ such that $c_1, \dots, c_T \neq 0$ and $T \geq 2$. Because of Assumption II, $\boldsymbol{\beta}_k = 0$ implies $\boldsymbol{\alpha}_1 = \ldots = \boldsymbol{\alpha}_T = 0$. Thus, $\boldsymbol{\beta} \cdot \mathbf{1}(\boldsymbol{\beta}_k = 0) = \boldsymbol{D}^{-1}\boldsymbol{D}^*\boldsymbol{\alpha}\mathbf{1}(\boldsymbol{\alpha}_1 = \ldots = \boldsymbol{\alpha}_T = 0)$, we know $\boldsymbol{\beta} \cdot \mathbf{1}(\boldsymbol{\beta}_k = 0)$ lies in a linear manifold of dimension $K - T$ almost surely. ■

Proof [Proof of Theorem 9] The whole proof consists of two major steps. The first step is to show that the finite population satisfies the Assumption I with high probability: for any $\epsilon > 0$,

$$P\left(\sup_{c_1,\dots,c_K} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\sum_j c_j \boldsymbol{\alpha}_j^{(i)} = 0 \text{ and } \sum_{j=1}^{K}\left(c_j\boldsymbol{\alpha}_j^{(i)}\right)^2 > 0) \geq \epsilon\right)$$
$$\leq 4\exp\left(2K\left(\ln\frac{n}{2K}+1\right) - n\left(\epsilon - \frac{1}{n}\right)^2\right). \tag{14}$$

In order to prove (14), define

$$f_c(\boldsymbol{\alpha}) \triangleq \mathbf{1}(\sum_{j=1}^{K} c_j\boldsymbol{\alpha}_j = 0 \text{ and } \sum_{j=1}^{K}(c_j\boldsymbol{\alpha}_j)^2 > 0),$$

$\mathcal{F}(\boldsymbol{\alpha}) \triangleq \{f_c(\cdot)|c \in \mathbb{R}^K\}$ and consider its VC dimension. We will prove the VC dimension of $\mathcal{F}$ is no bigger than $2K$, namely, for any $\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(2K)}$, define a set

$$\mathcal{F}^{(2K)}(\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(2K)}) \triangleq \{(f_c(\boldsymbol{\alpha}^{(1)}), \dots, f_c(\boldsymbol{\alpha}^{(2K)}))|c \in \mathbb{R}^K\},$$

The cardinality of $\mathcal{F}^{(2K)}$ is not 2^{2K}. If $\underbrace{(1,\dots,1)}_{2K}$ is not in $\mathcal{F}^{(2K)}$, then we are done. Otherwise, there exists c s.t. $f_c(\boldsymbol{\alpha}^{(i)}) = 1$ for any $i = 1, \dots, 2K$. That means $\sum_j c_j\boldsymbol{\alpha}_j^{(i)} = 0$ for any $i = 1, \dots, 2K$. Therefore, the dimension of the linear space spanned by $\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(2K)}$ is at most $K-1$. So we can find $K-1$ coefficients such that all other coefficients are their linear combinations. Without loss of generality, assume those coefficients are $\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(K-1)}$. Define the support of a vector to be the entries where it is nonzero. For $\boldsymbol{\alpha}^{(K)}, \dots, \boldsymbol{\alpha}^{(2K)}$, there will be one coefficient whose support is contained in the union of all the other coefficients. If this is not the case, each coefficient can be mapped to one entry which is only contained in its own support but not any support of other coefficients. But there are $K+1$ coefficient and only K entries, which leads to a contradiction. Without loss of generality, assume that coefficient is $\boldsymbol{\alpha}^{(K)}$. Now we will show that $(\underbrace{1,\dots,1}_{K}, 0, \dots, 0) \notin \mathcal{F}^{(2K)}(\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(2K)})$.

Since $f_c(\boldsymbol{\alpha}^{(i)}) = 1$ for $i = 1, \dots, K-1$, we have

$$\sum_j c_j\boldsymbol{\alpha}_j^{(i)} = 0 \quad \forall\ i = 1, \dots, K-1.$$

Because $\boldsymbol{\alpha}^{(K)}, \dots, \boldsymbol{\alpha}^{(2K)}$ are linear combinations of $\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(K-1)}$, we know

$$\sum_j c_j\boldsymbol{\alpha}_j^{(i)} = 0 \quad \forall\ i = K, \dots, 2K.$$

If $f_c(\boldsymbol{\alpha}^{(i)}) = 0$ for $i = K+1, \dots, 2K$, it means

$$\sum_j \left(c_j\boldsymbol{\alpha}_j^{(i)}\right)^2 = 0 \quad \forall\ i = K+1, \dots, 2K,$$

which means the support of c does not overlap with the support of $\boldsymbol{\alpha}^{(K+1)}, \dots, \boldsymbol{\alpha}^{(2K)}$. However, the support of $\boldsymbol{\alpha}^{(K)}$ is contained in the union of the supports of $\boldsymbol{\alpha}^{(K+1)}, \dots, \boldsymbol{\alpha}^{(2K)}$. That means $f_c(\boldsymbol{\alpha}^{(K)} = 0$ not 1, a contradiction.

Then by the classic statistical learning theory, for example, see Theorem 4.1 in Vapnik (1998), we know (14) holds true.

Now comes the second major step: we want to show that

$$A(\epsilon, \rho_1, \rho_2) \Rightarrow \sup_{c_1,\dots,c_K} \frac{1}{n} \sum_{i=1}^{n} f_c(\boldsymbol{\alpha}^{(i)}) > \frac{\rho_1^3 \epsilon}{2L\rho_2}.$$

Then, using (14), we get the desired conclusion.

For any $\epsilon, \rho_1, \rho_2 > 0$, if $\boldsymbol{D} \neq \boldsymbol{D}^*$ is a local min with sharpness at least ϵ and eigenvalue$(\boldsymbol{D}) \in [\rho_1, \rho_2]$, then $\sup_{c_1,\dots,c_K} \frac{1}{n} \sum_{i=1}^{n} f_c(\boldsymbol{\alpha}^{(i)}) > \frac{\rho_1^3 \epsilon}{L\rho_2}$. Since $\boldsymbol{D} \neq \boldsymbol{D}^*$ up to sign-permutation ambiguity, at least one row of $\boldsymbol{D}^{-1}\boldsymbol{D}^*$ contains two nonzero elements. Without loss of generality, assume the k-th row of $\boldsymbol{D}^{-1}\boldsymbol{D}^*$, denoted as $c^{(k)}$, has at least two nonzero entries. We will prove that it satisfies the desired condition:

$$\frac{1}{n} \sum_{i=1}^{n} f_{c^{(k)}}(\boldsymbol{\alpha}^{(i)}) > \frac{\rho_1^3 \epsilon}{2L\rho_2}.$$

Recall that $\boldsymbol{\beta}^{(i)} = \boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}$ for $i = 1, \dots, n$. Because $|||\boldsymbol{D}^{-1}|||_2 \leq \rho_1^{-1}$ and $\|\boldsymbol{x}^{(i)}\|_2$ is bounded by L, for any vector w such that $\|w\|_2 = 1$, we have $|\sum_j w_j \boldsymbol{\beta}_j^{(i)}| \leq L\rho_1^{-1}$ by Cauchy inequality. We have

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} f_{c^{(k)}}(\boldsymbol{\alpha}^{(i)}) \geq & \frac{\rho_1}{L} \max_i \{ \Big| \sum_j w_j \boldsymbol{\beta}_j^{(i)} \Big| \} \frac{1}{n} \sum_i f_{c^{(k)}}(\boldsymbol{\alpha}^{(i)}) \\
\geq & \frac{\rho_1}{L} \frac{1}{n} \sum_i | \sum_j w_j \boldsymbol{\beta}_j^{(i)} | f_{c^{(k)}}(\boldsymbol{\alpha}^{(i)}) \\
= & \frac{\rho_1}{L} \frac{1}{n} \sum_i | \sum_j w_j \boldsymbol{\beta}_j^{(i)} | \mathbf{1}(\boldsymbol{\beta}_k^{(i)} = 0, \sum_j (c_j^{(k)} \boldsymbol{\alpha}_j^{(i)})^2 > 0).
\end{aligned} \tag{15}$$

Note that this inequality holds for any w with unit norm. Recall that $c_j^{(k)}$ has at least two non-zero entries. Thus, for all the i's such that $\sum_j (c_j^{(k)} \boldsymbol{\alpha}_j^{(i)})^2 = 0$, $\boldsymbol{\alpha}_j^{(i)}$ must satisfy at least two linear constraints, which implies the corresponding $\boldsymbol{\beta}^{(i)}$'s must lie in a linear subspace of dimensions at most $K - 2$. Therefore, we can always select w such that $w_k = 0$ and for any i such that $\sum_j (c_j^{(k)} \boldsymbol{\alpha}_j^{(i)})^2 = 0$, we have $\sum_j w_j \boldsymbol{\beta}_j^{(i)} = 0$. Then, by using this specific w (this w satisfies $\|w\|_2 = 1, w_k = 0$, and for any $i = 1, \dots, n$, $\sum_j w_j \boldsymbol{\beta}_j^{(i)} \mathbf{1} \left(\boldsymbol{\beta}_k^{(i)} = 0, \sum_j \left(c_j^{(k)} \boldsymbol{\alpha}_j^{(i)} \right)^2 = 0 \right) = 0$), we have:

$$\frac{\rho_1}{L} \frac{1}{n} \sum_i | \sum_j w_j \boldsymbol{\beta}_j^{(i)} | \mathbf{1}(\boldsymbol{\beta}_k^{(i)} = 0, \sum_j (c_j^{(k)} \boldsymbol{\alpha}_j^{(i)})^2 > 0) = \frac{\rho_1}{L} \frac{1}{n} \sum_i | \sum_j w_j \boldsymbol{\beta}_j^{(i)} | \mathbf{1}(\boldsymbol{\beta}_k^{(i)} = 0) \tag{16}$$

Using the parametrization in Proposition 11, for $t > 0$ sufficiently small, $t \cdot w$ can map to a $\boldsymbol{D'}$ in the neighborhood of $\boldsymbol{D}$ such that $|||\boldsymbol{D'} - \boldsymbol{D}|||_F \geq \rho_1^2 |||\boldsymbol{D'}^{-1} - \boldsymbol{D}^{-1}|||_F \geq \rho_1^2 \cdot t \cdot \|w^T \boldsymbol{D}^{-1}\|_2 \geq \frac{\rho_1^2 t}{\rho_2}$. Because $\boldsymbol{D}$ is sharp local minimum with sharpness at least ϵ, we have

$$\frac{1}{n}\sum_{i=1}^{n} \|(\boldsymbol{D'})^{-1}\boldsymbol{x}^{(i)}\|_1 - \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{D}^{-1}\boldsymbol{x}^{(i)}\|_1 \geq \epsilon |||\boldsymbol{D'} - \boldsymbol{D}|||_F + o(|||\boldsymbol{D'} - \boldsymbol{D}|||_F).$$

By Lemma 16, we know the left hand side of the above inequality is equivalent to

$$|||(\boldsymbol{D'})^{-1}\boldsymbol{D}|||_{\boldsymbol{\beta}} - \mathrm{tr}(((\boldsymbol{D'})^{-1}\boldsymbol{D})^T B(\boldsymbol{\beta}, \boldsymbol{D}^T\boldsymbol{D}) \geq \epsilon |||\boldsymbol{D'} - \boldsymbol{D}|||_F + o(|||\boldsymbol{D'} - \boldsymbol{D}|||_F).$$

Without loss of generality, we could select $\boldsymbol{D'}$ (or $-\boldsymbol{D'}$) such that

$$\mathrm{tr}(((\boldsymbol{D'})^{-1}\boldsymbol{D})^T B(\boldsymbol{\beta}, \boldsymbol{D}^T\boldsymbol{D}) \geq 0.$$

This means the above inequality can be further rewritten as

$$|||(\boldsymbol{D'})^{-1}\boldsymbol{D}|||_{\boldsymbol{\beta}} \geq \frac{\rho_1^2\epsilon}{\rho_2} t + o(t).$$

Note that

$$\begin{aligned} |||(\boldsymbol{D'})^{-1}\boldsymbol{D}|||_{\boldsymbol{\beta}} =& \frac{1}{n}\sum_{i=1}^{n}\sum_{k'} \left|\boldsymbol{\beta}'^{(i)}_{k'}\right| \cdot \mathbf{1}(\boldsymbol{\beta}^{(i)}_{k'} = 0) \\ =& t \cdot \frac{1}{n}\sum_{i} |\sum_{j} w_j \boldsymbol{\beta}^{(i)}_j| \mathbf{1}(\boldsymbol{\beta}^{(i)}_k = 0). \end{aligned}$$

That means that when t is small,

$$\frac{1}{n}\sum_{i} |\sum_{j} w_j \boldsymbol{\beta}^{(i)}_j| \mathbf{1}(\boldsymbol{\beta}^{(i)}_k = 0) > \frac{\rho_1^2\epsilon}{2\rho_2}. \tag{17}$$

Combining (15), (16), and (17), we complete the proof. ■

Proof [Proof of Proposition 10] 1) ↔ 2): First observe that 2) is equivalent to the property that the directional derivative of the optimization (7) at $\mathbb{I}_k$ along any direction is always positive. By Theorem 4, we know 1) is equivalent to

$$|||B(\boldsymbol{\beta}, M)|||^*_{\boldsymbol{\alpha}} < 1.$$

Because of the definition of $|||\cdot|||_{\boldsymbol{\alpha}}$, this condition is equivalent to for any $k = 1, \ldots, K$, and $\delta_{k,j} \in \mathbb{R}$ for $j = 1, \ldots, K$, $j \neq k$ such that $\sum_{j \neq k} \delta_{k,j}^2 > 0$,

$$\left|\sum_{j} \delta_{k,j}\beta_j\right| \mathbf{1}(\beta_k = 0) + \sum_{k,j} \delta_{k,j}\beta_j \mathrm{sign}(\beta_k) - M_{k,j}\mathbb{E}|\beta_k| > 0.$$

On the other hand, the left hand side is exactly the directional derivative of the optimization (7) at $\mathbb{I}_j$ along direction $(\delta_1, \ldots, \delta_K)$. Because every directional derivative is strictly positive, $\mathbb{I}_k$ is a sharp local minimum of the optimization.

2) $\leftrightarrow$ 3). We have already shown that 2) is equivalent to

$$\Big|\sum_j \delta_{k,j}\beta_j\Big|\mathbf{1}(\beta_k = 0) + \sum_{k,j} \delta_{k,j}\beta_j \mathrm{sign}(\beta_k) + M_{k,j}\mathbb{E}|\beta_k| > 0.$$

3) is equivalent to

$$\Big|\sum_j \delta_{k,j}\beta_j\Big|\mathbf{1}(\beta_k = 0) + \sum_{k,j} \delta_{k,j}\beta_j \mathrm{sign}(\beta_k) + \tilde{M}_{k,j}\mathbb{E}|\beta_k| \geq 0.$$

for any $|\tilde{M}_{k,h} - M_{k,h}| \leq \rho$. These two are clearly equivalent for a sufficiently small ρ. ■

Proof [Proof of Proposition 11] Without loss of generality, we only need to show $\|Q_j^{-1}\|_2 = 1$ for any $j = 1, \dots, K$ when $k = 1$. We can write $Q = \Gamma \boldsymbol{D}^{-1}$, where the matrix Γ is equal to:

$$\Gamma_{h,j} = \begin{cases} w_j & \text{if } h = 1 \\ \sqrt{(w_h - M_{1,h})^2 + 1 - m_h^2} & \text{if } j = h \neq 1 \\ 0 & \text{otherwise} \end{cases}.$$

Note that Γ is upper triangle so we can obtain its inverse easily. Then $Q^{-1} = \boldsymbol{D}\Gamma^{-1}$ and

$$\Gamma^{-1}_{h,j} = \begin{cases} 1 & h = 1, j = 1 \\ -w_j/(\sqrt{(w_h - M_{1,h})^2 + 1 - M_{1,h}^2}) & h = 1, j > 1 \\ 1/(\sqrt{(w_h - M_{1,h})^2 + 1 - M_{1,h}^2}) & h > 1, j = h \\ 0 & h > 1, j \neq h \end{cases}$$

Q^{-1}'s first column has the form: $\|Q_1^{-1}\|_2^2 = \|\boldsymbol{D}_1\|_2^2 = 1$. For any other column Q_j^{-1} where $j > 1$, $\|Q_j^{-1}\|_2^2 = \|w_j\boldsymbol{D}_1 - \boldsymbol{D}_j\|_2^2/((w_j - M_{1,j})^2 + 1 - M_{1,j}^2) = 1$. ■

Proof [Proof of Proposition 12] Recall $f(\boldsymbol{D}) = \sum_{i=1}^n \sum_{j=1}^K \min(|\boldsymbol{D}^{-1}[j,]\boldsymbol{x}^{(i)}|, \tau)$. Denote $\boldsymbol{\beta}^{(i)} = (\boldsymbol{D}^{(t,j)})^{-1}\boldsymbol{x}^{(i)}$ and define a new function $\tilde{f}(\boldsymbol{D}) = \sum_{j=1}^K \sum_{i=1,|\boldsymbol{\beta}_j^{(i)}|\leq\tau}^n |\boldsymbol{D}^{-1}[j,]\boldsymbol{x}^{(i)}| + \sum_{i=1,|\boldsymbol{\beta}_j^{(i)}|>\tau}^n \tau$. Note that for any $\boldsymbol{D}$, $\tilde{f}(\boldsymbol{D})$ is always no smaller than $f(\boldsymbol{D})$, that is, $\tilde{f}(\boldsymbol{D}) \geq f(\boldsymbol{D})$. Also, $\tilde{f}(\boldsymbol{D}^{(t,j)}) = f(\boldsymbol{D}^{(t,j)})$. Because of Proposition 11, we know the iterate $\boldsymbol{D}^{(t,j+1)}$ in Algorithm 2 is the optimal solution of the following optimization:

$$\begin{aligned} &\mathrm{argmin}_Q \ \tilde{f}(Q^{-1}) \\ &\text{subject to } Q \text{ is parameterized as in Proposition 11.} \end{aligned}$$

That means $\tilde{f}(\boldsymbol{D}^{(t,j+1)}) \leq \tilde{f}(\boldsymbol{D}^{(t,j)})$. Combining the fact that $f(\boldsymbol{D}^{(t,j)}) = \tilde{f}(\boldsymbol{D}^{(t,j)})$ and $f(\boldsymbol{D}^{(t,j+1)}) \leq \tilde{f}(\boldsymbol{D}^{(t,j+1)})$, we have $f(\boldsymbol{D}^{(t,j+1)}) \leq \tilde{f}(\boldsymbol{D}^{(t,j+1)}) \leq \tilde{f}(\boldsymbol{D}^{(t,j)}) = f(\boldsymbol{D}^{(t,j)})$. ■

11. Appendix D: Parameter settings of dictionary learning algorithms in Section 5.4

- EM-BiG-AMP: The outer loop that performs EM iterations is allowed up to 20 iterations. The inner loop is allowed a minimum of 30 and a maximum of 1500 iterations.
- K-SVD: K-SVD has two parameters: number of iterations and the enforced sparsity. The number of iterations is set to be 1000. The enforced sparsity is set to be the same as the true sparsity of the underlying model s.
- SPAMS: SPAMS optimizes an LASSO type objective iteratively. The number of iterations is set to be 1000 and the penalty parameter in front of the L1 norm is $\lambda = .1/\sqrt{N}$.
- DL-BCD: Our algorithm has an outer loop and an inner loop. The outer loop is set to be at most 3. The inner loop is allowed a maximum of 100 iterations. τ is either ∞ or 0.5.
- ER-SpUD: We use the default settings in the package developed by the authors of ER-SpUD.

References

Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. A Clustering Approach to Learn Sparsely-Used Overcomplete Dictionaries. *arXiv preprint*, arxiv:1309(1952):1–31, 2013. ISSN 0018-9448. doi: 10.1109/TIT.2016.2614684.

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, and Praneeth Netrapalli. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2014.

Michal Aharon, Michael Elad, and Alfred M Bruckstein. K-SVD and its Non-Negative Variant for Dictionary Design. *International Society for Optics and Photonics*, 5914: 591411, 2005.

Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. ISSN 1053587X. doi: 10.1109/TSP.2006.881199.

Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *arXiv preprint arXiv:1401.0579*, page 23, 2014a.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Conference on Learning Theory*, pages 779–806, 2014b.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, Efficient, and Neural Algorithms for Sparse Coding. *arXiv:1503.00778 [cs, stat]*, 2015. ISSN 15337928.

A. M. Bagirov, L. Jin, N. Karmitsa, A. Al Nuaimat, and N. Sultanova. Subgradient method for nonconvex nonsmooth optimization. *Journal of Optimization Theory and Applications*, 157(2):416–435, May 2013. ISSN 1573-2878. doi: 10.1007/s10957-012-0167-6. URL https://doi.org/10.1007/s10957-012-0167-6.

Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method. *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, pages 143–151, 2014. ISSN 07378017. doi: 10.1145/2746539.2746605.

Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, and Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0308531101.

Emmanuel J. Candes, Michael B. Wakin, and Stephen P. Boyd. Enhancing sparsity by reweighted L1 minimization. *Journal of Fourier analysis and applications*, 14(5):877–905, 2008.

Niladri S. Chatterji and Peter L. Bartlett. Alternating minimization for dictionary learning with random initialization. In *Advances in Neural Information Processing Systems*, pages 1994–2003, 2017.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

Anulekha Dhara and Joydeep Dutta. *Optimality conditions in convex optimization: a finite-dimensional view.* CRC Press, 2011.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *Image Processing, IEEE Transactions on*, 15(12):3736–3745, 2006.

J-J Fuchs. On sparse representations in arbitrary redundant bases. *IEEE transactions on Information theory*, 50(6):1341–1344, 2004.

Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix Completion has No Spurious Local Minimum. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1–27, 2016. ISSN 10495258.

Quan Geng, John Wright, and Huan Wang. On the local correctness of L1-minimization for dictionary learning. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3180–3184. IEEE, 2014.

Rémi Gribonval and Karin Schnass. Dictionary Identification - Sparse Matrix-Factorisation via L1-Minimisation. *Information Theory, IEEE Transactions on*, 56(7):3523–3539, 2010.

Rémi Gribonval, Rodolphe Jenatton, and Francis Bach. Sparse and Spurious: Dictionary Learning With Noise and Outliers. *IEEE Transactions on Information Theory*, 2015. ISSN 00189448. doi: 10.1109/TIT.2015.2472522.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Patrik O. Hoyer. Non-negative sparse coding. *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop*, 2002-Janua:557–565, 2002. ISSN 0780376161. doi: 10.1109/NNSP.2002.1030067.

Kenneth Kreutz-delgado, Joseph F. Murray, Bhaskar D. Rao, Kjersti Engan, Te-Won Lee, and Terrence J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–96, 2003. ISSN 0899-7667. doi: 10.1162/089976603762552951.

Daniel Lee and Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems 13*, 2001. ISBN 9781424418206. doi: 10.1109/IJCNN.2008.4634046.

Sylvain Lesage, Rémi Gribonval, Frédéric Bimbot, and Laurent Benaroya. Learning unions of orthonormal bases with thresholded singular value decomposition. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 5, pages v—-293. IEEE, IEEE, 2005. doi: 10.1109/ICASSP.2005.1416298⟩.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online Dictionary Learning for Sparse Coding. *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696, 2009a.

Julien Mairal, Francis R. Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040, 2009b.

Julien Mairal, Rodolphe Jenatton, Francis Bach, Jean Ponce, Guillaume Obozinski, Bin Yu, Guillermo Sapiro, and Zaid Harchaoui. SPAMS : a SPArse Modeling Software , v2.5, 2014.

Saralees Nadarajah and Samuel Kotz. On the linear combination of laplace random variables. *Probab. Eng. Inf. Sci.*, 19(4):463–470, October 2005. ISSN 0269-9648. doi: 10.1017/S0269964805050291.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, 37(23):3311–3325, 1997. ISSN 00426989. doi: 10.1016/S0042-6989(97)00169-7.

Jason T. Parker and Philip Schniter. Parametric Bilinear Generalized Approximate Message Passing. In *IEEE Journal on Selected Topics in Signal Processing*, volume 10, pages 795–808, 2016. ISBN 1053-587X. doi: 10.1109/JSTSP.2016.2539123.

Jason T. Parker, Philip Schniter, and Volkan Cevher. Bilinear generalized approximate message passing—Part I: Derivation. *IEEE Transactions on Signal Processing*, 62(22): 5839–5853, 2014a.

Jason T. Parker, Philip Schniter, and Volkan Cevher. Bilinear generalized approximate message passing—Part II: Applications. *IEEE Transactions on Signal Processing*, 62 (22):5854–5867, 2014b.

Gabriel Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

Boris Teodorovic Polyak. Sharp Minima. In *Institute of Control Sciences Lecture Notes, Moscow IIASA Workshop On Generalized Lagrangians and Their Applications, IIASA, Laxenburg, Austria*, 1979.

Qiang Qiu, Vishal M Patel, and Rama Chellappa. Information-theoretic Dictionary Learning for Image Classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(11):2173–2184, 2014.

Joseph Lee Rodgers, W. Alan Nicewander, and Larry Toothaker. Linearly independent, orthogonal, and uncorrelated variables. *American Statistician*, 38(2):133–134, 1984. ISSN 15372731. doi: 10.1080/00031305.1984.10483183.

Ron Rubinstein, Alfred M. Bruckstein, and Michael Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010. ISSN 00189219. doi: 10.1109/JPROC.2010.2040551.

Karin Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD. *Applied and Computational Harmonic Analysis*, 37(3):464–491, 2014. ISSN 1096603X. doi: 10.1016/j.acha.2014.01.005.

Karin Schnass. Local Identification of Overcomplete Dictionaries. *Journal of Machine Learning Research*, 16:1211–1242, 2015. ISSN 15337928.

Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3087–3090. AAAI Press, 2013.

Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *Proceedings of the Twentieth International Conference on Machine Learning*, volume 3, pages 720–727, 2003. ISBN 1-57735-189-4. doi: 10.1.1.5.3301.

Ju Sun, Qing Qu, and John Wright. Complete Dictionary Recovery Over the Sphere I: Overview and the Geometric Picture. *IEEE Transactions on Information Theory*, 63(2): 853–884, 2017a. ISSN 00189448. doi: 10.1109/TIT.2016.2632162.

Ju Sun, Qing Qu, and John Wright. Complete Dictionary Recovery Over the Sphere II: Recovery by Riemannian Trust-Region Method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017b. ISSN 00189448. doi: 10.1109/TIT.2016.2632149.

Vladimir Vapnik. Statistical learning theory, 1998.

Christoph Witzgall and R. Fletcher. Practical Methods of Optimization. *Mathematics of Computation*, 1989. ISSN 00255718. doi: 10.2307/2008742.

Siqi Wu and Bin Yu. Local identifiability of l 1 -minimization dictionary learning: a sufficient and almost necessary condition. *Journal of Machine Learning Research*, 18:1–56, 2018.

Siqi Wu, Antony Joseph, Ann S. Hammonds, Susan E. Celniker, Bin Yu, and Erwin Frise. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences*, 113(16): 201521171, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1521171113.

Michael Zibulevsky and Barak A. Pearlmutter. Blind Source Separation by Sparse Decomposition in a Signal Dictionary. *Neural computation*, 4(13):863–882, 2001.