Improved Identifiability and Sample Complexity Analysis of Complete Dictionary Learning

Yuchen Sun and Kejun Huang
Department of Computer and Information Science and Engineering
University of Florida
Gainesville, Florida 32611
Email: (yuchen.sun, kejun.huang)@ufl.edu

2024 IEEE 13th Sensor Array and Multichannel Signal Processing Workshop (SAM) | 979-8-3503-4481-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/SAM60225.2024.10636510

***Abstract*—It has been recently shown that dictionary learning is identifiable with a complete dictionary, with relatively loose assumptions that the dictionary is invertible and that a geometric body obtained from the sparse coefficient matrix is sufficiently scattered in the hypercube. In addition, if the sparse coefficients are generated from the Bernoulli-Gaussian model, then the number of samples required to guarantee identifiability is $O(k\log(k))$. In this paper, we further improve the identifiability results in two-folds. First, we extend the deterministic identifiability results from the real domain in the previous work to the complex domain, hence significantly expand its applications. Second, we consider the generative model for the sparse coefficients from the more specific (also widely adopted) Bernoulli-Gaussian to a much wider class where the nonzero entries can be drawn from any subexponential distributions, and show that the same sample complexity results still holds. The recovery performance is confirmed in a Monte-Carlo simulation where the sparse coefficients are drawn from Bernoulli-Laplacian model.**

***Index Terms*—dictionary learning, sufficiently scattered, Bernoulli-subexponential.**

I. Introduction

Dictionary learning (DL) amounts to factor a data matrix as $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$ where $\boldsymbol{S}$ is sparse [1]. Treating $\boldsymbol{X} \in \mathbb{R}^{k\times n}$ or $\mathbb{C}^{k\times n}$ as a collection of data samples as its columns, this factorization means that each sample is a *sparse* combination of the columns of $\boldsymbol{A}$, or in other words atoms of the dictionary. Unlike the task of compressive sensing or sparse vector recovery, in which case the dictionary matrix $\boldsymbol{A}$ is given, dictionary learning tries to find both $\boldsymbol{A}$ and $\boldsymbol{S}$, therefore the problem is a lot more challenging. Depending on the shape of the dictionary matrix $\boldsymbol{A}$, we may seek to find a complete dictionary if $\boldsymbol{A}$ is square or an overcomplete dictionary if $\boldsymbol{A}$ is wide. In this paper we focus on complete dictionary learning, therefore $\boldsymbol{A} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ (or $\mathbb{C}^{k\times k}$ and $\mathbb{C}^{k\times n}$, respectively).

Identifiability of dictionary learning has been an open question ever since the problem was posed. As an instance of matrix factorization, inherent ambiguities of permutation and scaling of the columns of $\boldsymbol{A}$ (and in turn same permutation and counter-scaling of the rows of $\boldsymbol{S}$) are unavoidable and inconsequential in practice, and the factorization is essentially unique or identifiable if all admissible factorization are permutation and scaling of each other. Earlier works focus on directly enforcing sparsity constraints on $\boldsymbol{S}$, which show that over-complete DL is identifiable if $n > O((k+1)\binom{k}{s})$ [2], [3] and complete DL is identifiable with $O(k^3/(k-s)^2)$ samples. Notice that these work require *all* columns of $\boldsymbol{S}$ to be strictly s-sparse—no dense outliers is allowed for the analysis to work. More recently, there have been work on identifiability of DL with ℓ_1 norm as the regularization to promote sparsity, although they are only able to show identifiability in a local region [4]–[6] if the dictionary $\boldsymbol{A}$ is incoherent and the sparse coefficients $\boldsymbol{S}$ follows some sparse generative model such as the Bernoulli-Gaussian model. The benefit is not only that algorithm design is easier with ℓ_1 norm regularization, but also that the number of samples n required is usually $O(k\log k)$, while a few outlying dense columns of $\boldsymbol{S}$ are allowed.

Supported in part by NSF ECCS-2237640 and NIH R01LM014027.

The best known result regarding the identifiability of DL is in [7], which shows that complete dictionary learning is *globally* identifiable if: 1) the complete dictionary $\boldsymbol{A} \in \mathbb{R}^{k\times k}$ is invertible (hence no incoherence is necessary for complete dictionaries), and 2) the sparse coefficients $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ satisfies some sufficiently scattered condition, which we will elaborate in the next section. It was also shown that if $\boldsymbol{S}$ is generated from the Bernoulli-Gaussian model, it will satisfy the sufficiently scattered condition with overwhelming probability as long as the sample size n is larger than $O(k\log(k))$, which confirms the sample complexity analysis in [4]–[6] but identifiability is guaranteed globally, not just in a local region near the groundtruth factors. This new framework for identifiable DL is based on a new problem formulation by constraining the ℓ_1 norm of each row of $\boldsymbol{S}$ while minimizing the matrix volume of the dictionary, i.e., $|\det \boldsymbol{A}|$.

In this paper, we aim at further improving the identifiability result for complete dictionary learning. First, noticing that all the results in [7] applies only to real dictionary learning, we extend the identifiability result from the real case to the complex case, i.e., allowing both the dictionary matrix $\boldsymbol{A}$ and the sparse coefficients matrix $\boldsymbol{S}$ to have complex values. A complex matrix $\boldsymbol{S}$ is sparse if a significant portion of its values are equal to zero, but for its nonzero values we do not care if their real or imaginary parts are zero—they are simply treated as nonzeros. Second and more importantly, we extend the sample analyses to a much wider class of generative models for $\boldsymbol{S}$ called Bernoulli-subexponential models; compared to the Bernoulli-Gaussian model that is widely adopted in DL research, we now allow the nonzero values of $\boldsymbol{S}$ to be drawn from any subexponential distributions, which includes not only Gaussian but also a large variety of distributions such as Laplacian, Cauchy, chi, and any bounded ones [8].

II. Identifiability of dictionary learning

We denote the groundtruth factorization of the DL model as $\boldsymbol{X} = \boldsymbol{A}_\natural\boldsymbol{S}_\natural$. Suppose the rows of $\boldsymbol{X}$ are linearly independent, which is a valid assumption since $k > n$, then the row space of $\boldsymbol{X}$ is the same as that of $\boldsymbol{S}_\natural$. Consider another factorization $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$, then the same rule applies, which means that there exists an invertible matrix $\boldsymbol{W}$ such that $\boldsymbol{S} = \boldsymbol{W}\boldsymbol{S}_\natural$. Furthermore, both $\boldsymbol{S}$ and $\boldsymbol{S}_\natural$ are right-invertible, and they are related as $\boldsymbol{S}^\dagger = \boldsymbol{S}_\natural^\dagger\boldsymbol{W}^{-1}$; multiplying it from the right on both sides of $\boldsymbol{A}\boldsymbol{S} = \boldsymbol{A}_\natural\boldsymbol{S}_\natural$ shows that $\boldsymbol{A} = \boldsymbol{A}_\natural\boldsymbol{W}^{-1}$. This shows that matrix factorization in general is not identifiable without any structural constraints. On the other hand, people have discovered a plethora of structural constraints that do guarantee identifiability, i.e., restricting the choice of $\boldsymbol{W}$ to be the product of a permutation matrix

and a diagonal matrix, such as nonnegative [9]–[11], simplicial [12], [13], or bounded [14]–[16].

A breakthrough result regarding identifiability of DL has been shown in [7] using the following formulation:

$$\begin{aligned} \underset{\boldsymbol{A},\boldsymbol{S}}{\text{minimize}} \quad & |\det \boldsymbol{A}| \\ \text{subject to} \quad & \boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}, \|\boldsymbol{S}_{j,:}\|_1 \le 1, j = 1, \dots, k, \end{aligned} \tag{1}$$

In the remainder of this section, we present the identifiability result in the complex domain, although the original result in [7] was in the real domain, and show that similar identifiability result can be obtained in a much more general complex case.

First we introduce some definitions.

Definition 1 (Cellular hull)**.** The cellular hull of a finite set of vectors $\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_n\}$, stacked as the columns of the matrix $\boldsymbol{S}$, is

$$\text{cell}(\boldsymbol{S}) = \left\{ \boldsymbol{S}\boldsymbol{\theta} \;\middle|\; \|\boldsymbol{\theta}\|_\infty \le 1 \right\}.$$

Notice that the definition works in both real and complex cases: in the real domain, it restricts the absolute values of the coefficients θ_i, and in the complex domain, it restricts the magnitudes of the complex coefficients θ_i.

Assumption 1 (Sufficiently scattered in the complex hypercube)**.** Let $\mathcal{B}$ denote the complex Euclidean ball $\mathcal{B} = \{\boldsymbol{x} \in \mathbb{C}^k \mid \|\boldsymbol{x}\| \le 1\}$ and $\mathcal{C}$ denote the complex hypercube $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{C}^k \mid \|\boldsymbol{x}\|_\infty \le 1\}$. A set $\mathcal{S}$ is sufficiently scattered in the complex hypercube if:

1) $\mathcal{B} \subseteq \mathcal{S} \subseteq \mathcal{C}$;
2) $\partial\mathcal{B} \cap \partial\mathcal{S} = \{\alpha \boldsymbol{e}_i \mid |\alpha| = 1, i = 1, \dots, k\}$, where ∂ denotes the boundary of the set, and $\boldsymbol{e}_1, \dots, \boldsymbol{e}_k$ are the k unit vectors in $\mathbb{R}^k$.

Our main result on the identifiability of complex DL is presented as follows:

Theorem 1. *Consider the complex DL model $\boldsymbol{X} = \boldsymbol{A}_\natural \boldsymbol{S}_\natural$, where $\boldsymbol{A}_\natural \in \mathbb{C}^{d\times k}$ is the groundtruth mixing matrix and $\boldsymbol{S}_\natural$ is the groundtruth sparse coefficient matrix. Let $\widetilde{\boldsymbol{S}}_\natural$ denote the matrix obtained from rescaling the rows of $\boldsymbol{S}_\natural$ to have unit ℓ_1 norms. If* $\text{rank}(\boldsymbol{A}_\natural) = k$ *and* $\text{cell}(\widetilde{\boldsymbol{S}}_\natural)$ *is sufficiently scattered in the complex hypercube as in Assumption 1, then for any solution of* (1)*, denoted as $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$, there exist a permutation matrix $\boldsymbol{\Pi}$ and a complex diagonal matrix $\boldsymbol{D}$ such that*

$$\boldsymbol{A}_\natural = \boldsymbol{A}_\star \boldsymbol{D}\boldsymbol{\Pi} \qquad \textit{and} \qquad \boldsymbol{S}_\natural = \boldsymbol{\Pi}^\top \boldsymbol{D}^{-1} \boldsymbol{S}_\star.$$

In other words, complex DL is identifiable if the groundtruth $\boldsymbol{A}_\natural$ has full column rank and the cellular hull of $\boldsymbol{S}_\natural$ is sufficiently scattered.

Proof. Let $\widetilde{\boldsymbol{S}}_\natural$ denote the matrix obtained from rescaling the rows of $\boldsymbol{S}_\natural$ to have unit ℓ_1 norms, and counter-scale the columns of $\boldsymbol{A}_\natural$ to obtain $\widetilde{\boldsymbol{A}}_\natural$. Since both $(\widetilde{\boldsymbol{A}}_\natural, \widetilde{\boldsymbol{S}}_\natural)$ and $(\boldsymbol{A}_\star, \boldsymbol{S}_\star)$ are feasible for (1), we immediately have that $|\det \boldsymbol{A}_\star| \le |\det \widetilde{\boldsymbol{A}}_\natural|$. Define $\boldsymbol{W} = \boldsymbol{A}_\star^{-1}\widetilde{\boldsymbol{A}}_\natural$, then

$$|\det \boldsymbol{W}| = |\det \boldsymbol{A}_\star^{-1}\widetilde{\boldsymbol{A}}_\natural| = |\det \widetilde{\boldsymbol{A}}_\natural| / |\det \boldsymbol{A}_\star| \ge 1. \tag{2}$$

On the other hand, since $\boldsymbol{W}\widetilde{\boldsymbol{S}}_\natural = \boldsymbol{A}_\star^{-1}\widetilde{\boldsymbol{A}}_\natural\widetilde{\boldsymbol{S}}_\natural = \boldsymbol{S}_\star$, we also have that $\|[\boldsymbol{W}\widetilde{\boldsymbol{S}}_\natural]_{j,:}\| \le 1$. Let $\boldsymbol{w}^\mathsf{H}$ be any row of $\boldsymbol{W}$, then $\|\boldsymbol{w}^\mathsf{H}\widetilde{\boldsymbol{S}}_\natural\|_1 \le 1$. This is equivalent to a set of linear inequalities

$$\boldsymbol{w}^\mathsf{H}\widetilde{\boldsymbol{S}}_\natural\boldsymbol{\theta} \le 1, \quad \forall \|\boldsymbol{\theta}\|_\infty \le 1.$$

Now we invoke the assumption that $\text{cell}(\widetilde{\boldsymbol{S}}_\natural)$ is sufficiently scattered, then for any $\boldsymbol{v}$ with unit norm $\|\boldsymbol{v}\| = 1$, there must exist $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\|_\infty \le 1$ such that $\boldsymbol{v} = \widetilde{\boldsymbol{S}}_\natural\boldsymbol{\theta}$. Therefore

$$\left|\boldsymbol{w}^\mathsf{H}\boldsymbol{v}\right| = \left|\boldsymbol{w}^\mathsf{H}\widetilde{\boldsymbol{S}}_\natural\boldsymbol{\theta}\right| \le 1. \tag{3}$$

Since (3) holds for every $\boldsymbol{v}$ with unit norm, we must have $\|\boldsymbol{w}\| \le 1$ as well (otherwise we would let $\boldsymbol{v} = \boldsymbol{w}/\|\boldsymbol{w}\|$ and have that $|\boldsymbol{w}^\mathsf{H}\boldsymbol{v}| = \|\boldsymbol{w}\| > 1$, contradicting (3)), which means every row of $\boldsymbol{W}$ has norm no greater than 1. This gives us

$$|\det \boldsymbol{W}| \le \prod_{j=1}^{k} \|\boldsymbol{w}_j\| \le 1, \tag{4}$$

where the first inequality is due to the Hadamard inequality. Combining (2) and (4) shows that $(\boldsymbol{A}^\natural, \boldsymbol{S}^\natural)$ or any of their column permutation and unimodular scaling is in the set of optimal solutions of (1). The second requirement of Assumption 1 further ensures that they are the only possible solutions. This shows the identifiability of complete dictionary learning in the complex domain. □

III. Improved sample complexity analysis

The analysis given in the previous section gives an exact characterization of when dictionary learning is identifiable using the proposed formulation (1) with a matrix volume identification criterion. The analysis is inspired by the line of work from nonnegative matrix factorization and simplicial component analysis [9], [11], [13] using geometric interpretations, but with the introduction of cellular hulls in the complex domain, the geometric interpretation becomes harder to visualize. Nevertheless, there is a simple algebraic representation: $\text{cell}(\widetilde{\boldsymbol{S}})$ is sufficiently scattered in the complex hypercube if the solution set of the following optimization problem is $\{\alpha\boldsymbol{e}_j \mid |\alpha| = 1, j = 1, \dots, k\}$:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{maximize}} \quad & \|\boldsymbol{w}\|^2 \\ \text{subject to} \quad & \|\widetilde{\boldsymbol{S}}^\mathsf{H}\boldsymbol{w}\|_1 \le 1. \end{aligned} \tag{5}$$

Solving (5) exactly is NP-hard. In this section, we assume that the sparse coefficient matrix $\boldsymbol{S}$ is generated from a probabilistic generative model and show that it would satisfy guarantee identifiability with high probability, provided that the number of data points n is $O(k \log k)$. What differs from the previous work [7] as well as several earlier work on local identifiability is that here we consider a much broader class of generative model called Bernoulli-Subexponential model as introduced here.

Assumption 2 (Bernoulli-Subexponential model)**.** The matrix $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ or $\mathbb{C}^{k\times n}$ is generated from a Bernoulli-Subexponential model with parameter $p \in (0, 1)$, denoted as $\boldsymbol{S} \sim \mathcal{BE}(p)$, if its elements are i.i.d. with $S_{ij} = b_{ij}g_{ij}$, where $b_{ij} \in \{0, 1\}$ are i.i.d. Bernoulli random variables with $\Pr[b_{ij} = 1] = p$ (thus $\Pr[b_{ij} = 0] = 1 - p$) and g_{ij} are i.i.d. random variables drawn from a subexponential distribution with subexponential norm ν.

First let us review the definition of subexponential distributions. As the name suggests, a random variable is called subexponential if its tail distribution can be upperbounded by an exponential function. There are several equivalent characterizations of subexponential distributions [8]. One of the more explicit definitions, which involves the notion of subexponential norms in Assumption 2, is as follows:

Definition 2 (Subexponential random variables)**.** A random variable $Z \in \mathbb{R}$ or $\mathbb{C}$ is called a sub-exponential random variable if the moment-generating function of $|Z|$ satisfies $\mathrm{E}[\exp(|Z|/t)] \le 2$ for some $t > 0$. The smallest t that satisfies such inequality is called the subexponential norm of Z, denoted as ν in Assumption 2, i.e.,

$$\nu = \inf\{t > 0 : \mathrm{E}[\exp(|Z|/t)] \le 2\}.$$

A wide variety of random variables can be categorized as subexponential, including not only common ones such as Gaussian and

bounded, but also heavy-tailed ones such as Laplacian, Cauchy, and chi. By allowing the nonzero entries of $\boldsymbol{S}$ to be drawn from any subexponential distribution, we greatly expand the applicability of the sample complexity analysis given in [7].

One of the nice properties of subexponential random variables is that sums of independent ones are concentrated around their mean, as shown in Bernstein's inequality:

Theorem 2 (Bernstein's inequality [8])**.** *Let $Z_1, \dots, Z_n$ be i.i.d. subexponential random variables with mean μ and subexponential norm ν. Then, for any $t > 0$, we have*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| > \epsilon\right] \le 2\exp\left(-cn\min\left(\frac{\epsilon^2}{\nu^2}, \frac{\epsilon}{\nu}\right)\right),$$

where $c > 0$ is an absolute constant.

We are now ready to present the main theorem about improved sample complexity for identifiable dictionary learning:

Theorem 3. *Suppose $\boldsymbol{S} \in \mathbb{R}^{k\times n}$ or $\mathbb{C}^{k\times n}$ is generated from the Bernoulli-Subexponential model $\mathcal{BE}(p)$, and $\widetilde{\boldsymbol{S}}$ is obtained by scaling its rows to have unit ℓ_1 norm. Then*

$$\Pr\left[\sup_{\|\boldsymbol{w}^\top\widetilde{\boldsymbol{S}}\|_1\le 1} \|\boldsymbol{w}\| > 1\right] \le 4\exp\left(k\log(3\sqrt{k}) - cn(\mu/\nu)^2 p(1-p)\right), \tag{6}$$

where μ and ν are parameters of the subexponential distribution, and c is an absolute constant (c.f. Theorem 2).

Proof. We assume that $\boldsymbol{S} \sim \mathcal{BE}(p)$, and the first thing we do is rescaling its rows to have unit ℓ_1 norms and use it for the problem (5). To simplify the analysis, we can instead directly maximize $\|\boldsymbol{w}\|^2$ subject to $\|\boldsymbol{w}^\top\boldsymbol{S}\|_1 \le 1$, and compare it with the largest ℓ_1 norm of the rows of $\boldsymbol{S}$. The complement of the intended probability can be bounded as

$$\Pr\left[\sup_{\|\boldsymbol{w}^\top\widetilde{\boldsymbol{S}}\|_1\le 1} \|\boldsymbol{w}\| \le 1\right] \ge \Pr\left[\sup_{\|\boldsymbol{w}^\top\boldsymbol{S}\|_1\le 1} \|\boldsymbol{w}\| \le \alpha \cap \max_j \|\boldsymbol{S}_{j,:}\|_1 \ge \alpha\right],$$

with arbitrary choice of α. Conversely,

$$\begin{aligned}\Pr\left[\sup_{\|\boldsymbol{w}^\top\widetilde{\boldsymbol{S}}\|_1\le 1} \|\boldsymbol{w}\| > 1\right] &\le \Pr\left[\sup_{\|\boldsymbol{w}^\top\boldsymbol{S}\|_1\le 1} \|\boldsymbol{w}\| > \alpha \cup \max_j \|\boldsymbol{S}_{j,:}\|_1 < \alpha\right] \\ &\le \Pr\left[\sup_{\|\boldsymbol{w}^\top\boldsymbol{S}\|_1\le 1} \|\boldsymbol{w}\| > \alpha\right] + \Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 < \alpha\right] \end{aligned} \tag{7}$$

where the second inequality is obtained from the union bound.

a) Bounding the second term in (7)*:* Let $\boldsymbol{s} = (s_1, \dots, s_n)$ be generated from $\mathcal{BG}(p)$, i.e., each $s_i = b_i g_i$ with b_i Bernoulli with probability p and g_i subexponential with norm ν, we will use Bernstein's inequality with $Z_i = b_i|g_i|$. If $\mathrm{E}\,|g_i| = \mu$, then $\mathrm{E}\,Z_i = p\mu$, and Z_i is also subexponential with norm $p\nu$. Therefore

$$\begin{aligned}\Pr\left[\|\boldsymbol{s}\|_1 < n(p\mu-\epsilon)\right] &\le \Pr\left[\left|\sum_{i=1}^{n}(Z_i - \mathrm{E}[Z_i])\right| > n\epsilon\right] \\ &\le 2\exp\left(-cn\min\left(\frac{\epsilon^2}{p^2\nu^2}, \frac{\epsilon}{p\nu}\right)\right)\end{aligned}$$

This puts a bound on the probability that one row of $\boldsymbol{S}$ has bounded ℓ_1 norm. The second term in (7) requires all k rows to be bounded, which clearly has an even smaller probability, therefore

$$\Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 < \alpha\right] \le 2\exp\left(-cn\min\left(\frac{\epsilon^2}{p^2\nu^2}, \frac{\epsilon}{p\nu}\right)\right). \tag{8}$$

b) Bounding the first term in (7)*:* First we note the following equivalence:

$$\Pr\left[\sup_{\|\boldsymbol{w}^\top\boldsymbol{S}\|_1\le 1} \|\boldsymbol{w}\| > \alpha\right] = \Pr\left[\inf_{\|\boldsymbol{w}\|=1} \|\boldsymbol{S}^\top\boldsymbol{w}\|_1 < 1/\alpha\right] \tag{9}$$

We are also going to use the following notion of δ-cover from convex geometry [17] that holds for all ℓ_p norm balls, but we are only to instantiate the Euclidean ball:

Lemma 1 (δ-cover)**.** *A finite δ-cover of the unit sphere in $\mathbb{R}^k$ is a finite set C_δ of points with unit ℓ_2 norm such that any point on the unit sphere is within ϵ away from an element in C_δ, i.e.*

$$\min_{\boldsymbol{w}_i\in C_\delta} \|\boldsymbol{w} - \boldsymbol{w}_i\| < \delta, \quad \forall\|\boldsymbol{w}\| = 1.$$

For $\delta \in (0,1)$ there always exists an δ-cover C_ϵ with cardinality $|C_\delta| < (3/\delta)^k$. In the complex case, the ℓ_2 norm essentially treats the real and imaginary parts separately, therefore the δ-cover is $|C_\delta| < (3/\delta)^{2k}$

Let $C_\delta = \{\boldsymbol{w}_i\}$ be an δ-cover for the sphere in $\mathbb{R}^k$ or $\mathbb{C}^k$. Assume that we have both the lowerbound

$$\|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}_i\|_1 \ge \beta, \forall \boldsymbol{w}_i \in C_\delta$$

and the upperbound

$$\|\boldsymbol{S}^{\mathsf{H}}\|_1 = \sup_{\|\boldsymbol{w}\|_1\le 1} \|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}\|_1 \le \gamma.$$

Then

$$\begin{aligned}\|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}\|_1 &\ge \|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}_i\|_1 - \|\boldsymbol{S}^{\mathsf{H}}(\boldsymbol{w}-\boldsymbol{w}_i)\|_1 \\ &\ge \beta - \|\boldsymbol{S}^{\mathsf{H}}\|_1\|\boldsymbol{w}-\boldsymbol{w}_i\|_1 \\ &\ge \beta - \|\boldsymbol{S}^{\mathsf{H}}\|_1\|\boldsymbol{w}-\boldsymbol{w}_i\|\sqrt{k} \ge \beta - \gamma\delta\sqrt{k}.\end{aligned}$$

Therefore

$$\inf_{\|\boldsymbol{w}\|\le 1} \|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}\|_1 \ge \beta - \gamma\delta\sqrt{k}.$$

As a result, we have

$$\begin{aligned}&\Pr\left[\inf_{\|\boldsymbol{w}\|\le 1} \|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}\|_1 < \beta - \gamma\delta\sqrt{k}\right] \\ &\quad\le \sum_{\boldsymbol{w}_i\in C_\delta} \Pr\left[\|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}_i\|_1 < \beta\right] + \Pr\left[\|\boldsymbol{S}^{\mathsf{H}}\|_1 > \gamma\right],\end{aligned} \tag{10}$$

where C_δ is a δ-cover of the unit sphere with cardinality $|C_\delta| < (3/\delta)^k$ in $\mathbb{R}^k$ or $|C_\delta| < (3/\delta)^{2k}$ in $\mathbb{C}^k$.

The bound to the first term in (10) is almost identical to (7). Dropping the subscript of $\boldsymbol{w}_i$, we write

$$\|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}\|_1 = \sum_{i=1}^{n}\left|\sum_{j=1}^{k} b_{ij}g_{ij}w_j\right| := \sum_{i=1}^{n} Z_i.$$

Each Z_i is the absolute value of the weighted sum of k subexponentials with norm $p\nu$, so it is also subexponential with norm less than $p\nu$, Denote $\mu_{\boldsymbol{w}} = \mathrm{E}[Z_i]$, then

$$\begin{aligned}\Pr\left[\|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}\|_1 < n(\mu_{\boldsymbol{w}}-\epsilon)\right] &\le \Pr\left[\left|\sum_{i=1}^{n}(Z_i - \mathrm{E}[Z_i])\right| > n\epsilon\right] \\ &\le 2\exp\left(-cn\min\left(\frac{\epsilon^2}{p^2\nu^2}, \frac{\epsilon}{p\nu}\right)\right).\end{aligned} \tag{11}$$

To bound $\|\boldsymbol{S}^{\mathsf{H}}\|_1$, we recall that this is the ℓ_1 induced norm for matrix $\boldsymbol{S}^{\mathsf{H}}$, which is shown to be the maximum of the ℓ_1 norms of

the columns of $\boldsymbol{S}^{\mathsf{H}}$. This means we can use similar arguments used in (8) (but applied to the other direction) to have

$$\begin{aligned}\Pr\left[\|\boldsymbol{S}^{\mathsf{H}}\|_1 > n(p\mu+\epsilon)\right] &= \Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 > n(p\mu+\epsilon)\right] \\ &\le \Pr\left[\|\boldsymbol{S}_{j,:}\|_1 > n(p\mu+\epsilon)\right] \\ &\le \Pr\left[\left|\sum_{i=1}^{n}(Z_i - \mathrm{E}[Z_i])\right| > n\epsilon\right] \\ &\le 2\exp\left(-cn\min\left(\frac{\epsilon^2}{p^2v^2}, \frac{\epsilon}{pv}\right)\right), \end{aligned} \tag{12}$$

where we pick an arbitrary $j \in [k]$ in the second line since this event implies that the maximum ℓ_1 norm of the rows are lowerbounded.

Combining (11), (11), and (12) with $\beta = n(\mu_{\boldsymbol{w}} - \epsilon)$, $\gamma = n(p\mu+\epsilon)$, and

$$\delta = \frac{n^2(\mu_{\boldsymbol{w}} - \epsilon)(p\mu - \epsilon) - 1}{n^2(p\mu+\epsilon)(p\mu-\epsilon)\sqrt{k}},$$

which satisfies $0 < \delta < 1$ for small enough ϵ, then we have

$$\begin{aligned}&\Pr\left[\inf_{\|\boldsymbol{w}\|=1} \|\boldsymbol{S}^{\mathsf{T}}\boldsymbol{w}\|_1 < 1/(n(p\mu-\epsilon))\right] \\ &\le ((3/\delta)^k + 1)2\exp\left(-cn\min\left(\frac{\epsilon^2}{p^2v^2}, \frac{\epsilon}{pv}\right)\right). \end{aligned} \tag{13}$$

Combining (13) and (8) with $\alpha = n(p\mu - \epsilon)$ into (7) gives us

$$\Pr\left[\sup_{\|\boldsymbol{w}^{\mathsf{H}}\tilde{\boldsymbol{S}}\|_1 \le 1} \|\boldsymbol{w}\| > 1\right] \tag{14}$$

$$\le \Pr\left[\inf_{\|\boldsymbol{w}\|=1} \|\boldsymbol{S}^{\mathsf{H}}\boldsymbol{w}\|_1 < 1/\alpha\right] + \Pr\left[\max_j \|\boldsymbol{S}_{j,:}\|_1 < \alpha\right]$$

$$\le 2\left(\left(\frac{3}{\delta}\right)^k + 2\right)\exp\left(-cn\min\left(\frac{\epsilon^2}{p^2v^2}, \frac{\epsilon}{pv}\right)\right). \tag{15}$$

For any $0 < \epsilon < p\mu$ we have $\delta > 1/\sqrt{k}$; if we further require $\epsilon^2/p^2v^2 < \epsilon/pv$, i.e. $\epsilon < pv$, then (14) becomes

$$\begin{aligned}\Pr\left[\sup_{\|\boldsymbol{w}^{\mathsf{H}}\tilde{\boldsymbol{S}}\|_1 \le 1} \|\boldsymbol{w}\| > 1\right] &\le 2\left(\left(3\sqrt{k}\right)^k + 2\right)\exp\left(-\frac{cn\epsilon^2}{p^2v^2}\right) \\ &\le 4\left(3\sqrt{k}\right)^k \exp\left(-\frac{cn\epsilon^2}{p^2v^2}\right) \\ &\le 4\exp\left(k\log(3\sqrt{k}) - cn(\mu/v)^2p(1-p)\right). \end{aligned}$$

□

Theorem 3 shows that if $\boldsymbol{S}$ is generated from a Bernoulli-Subexponential model, then the probability that it does not satisfy the sufficiently scattered condition can be upperbounded by (6), which approaches zero at an exponential rate as $n \gg k\log(2\sqrt{k})/p(1-p)$. As a result, we have the following corollary that directly bounds the probability that such a Bernoulli-Subexponential $\boldsymbol{S}$ can be uniquely identified via complex dictionary learning.

Corollary 1. *Consider the generative model $\boldsymbol{X} = \boldsymbol{A}_\natural\boldsymbol{S}_\natural$, where $\boldsymbol{A}_\natural \in \mathbb{C}^{k\times k}$ is the groundtruth dictionary and $\boldsymbol{S}$ is the groundtruth sparse coefficients. If $\mathrm{rank}(\boldsymbol{A}_\natural) = k$ and the matrix $\boldsymbol{S}_\natural \in \mathbb{C}^{k\times n}$ is generated from the Bernoulli-Subexponential model $\mathcal{BE}(p)$, then $(\boldsymbol{A}_\natural, \boldsymbol{S}_\natural)$ are globally identifiable via optimizing* (1) *with probability at least*

$$1 - 4\exp\left(k\log(3\sqrt{k}) - cn(\mu/v)^2p(1-p)\right).$$

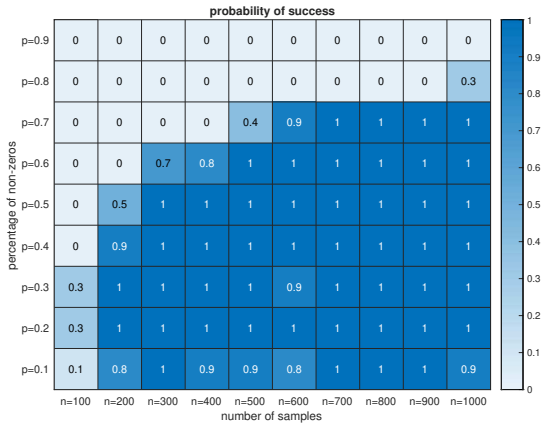

Fig. 1. Groundtruth $\boldsymbol{S}_\natural$ is generated from a Bernoulli-Laplacian model. The probability of exactly recovering the dictionary for various sample size n and sparsity p (probability of nonzeros in $\boldsymbol{S}_\natural$). The dictionary size is fixed with $k = 20$.

IV. Experiments

We justify the improved sample complexity analysis of identifiable dictionary learning in the following numerical experiment conducted in MATLAB. The algorithm to approximately solve formulation (1) is the L-ADMM algorithm proposed in [7]. Different from the simulations done in [7], the groundtruth dictionary $\boldsymbol{S}_\natural$ is generated from a Bernoulli-Laplacian model; the Laplacian distribution has a much heavier tail distribution compared to Gaussian, but is still in the category of subexponential distributions, and we demonstrate that similar identifiability results still hold here, as confirmed by Theorem 3.

We fix the dictionary size $k = 20$, and change the sample size n from 100 to 1000 and the probability p in the Bernoulli-Subexponential model from 0.1 to 0.9. To ensure definitively that the dictionary is recovered, we will normalize the column and use the Hungarian algorithm [18] to find the best column matching, and then calculate the estimation error. We declare success if the estimation error is smaller than `1e-5`. In each setting the experiments are repeated 10 times and the percentage of exact recovery are recorded. The results are shown in Figure 1, which agrees with the bound in Theorem 3.

V. Conclusion

In this work we greatly improve the identifiability result of complete dictionary learning in [7] in two ways. First, the result is generalized to the complex domain, while [7] was only shown in the real domain. Second, we show that the same sample complexity analysis could be extended to a wider class of generative model in which the nonzeros of the sparse coefficients can be drawn from any subexponential distributions, as opposed to only Gaussian models in the previous work. We demonstrate the recovery performance in Monte-Carlo simulation where the nonzeros of the sparse coefficients are drawn from the Laplacian distributions, and the results show that the identifiability agrees with the developed theory.

References

[1] I. Tošić and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.

[2] M. Aharon, M. Elad, and A. M. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear algebra and its applications*, vol. 416, no. 1, pp. 48–67, 2006.

[3] C. J. Hillar and F. T. Sommer, "When can dictionary learning uniquely recover sparse data from subsamples?" *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6290–6297, 2015.
[4] R. Gribonval and K. Schnass, "Dictionary identification—–sparse matrix-factorization via ℓ_1-minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.
[5] S. Wu and B. Yu, "Local identifiability of l1-minimization dictionary learning: a sufficient and almost necessary condition," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6121–6176, 2017.
[6] Y. Wang, S. Wu, and B. Yu, "Unique sharp local minimum in l1-minimization complete dictionary learning," *Journal of Machine Learning Research*, vol. 21, no. 63, pp. 1–52, 2020.
[7] J. Hu and K. Huang, "Global identifiability of ℓ_1-based dictionary learning via matrix volume optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.
[8] R. Vershynin, *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge university press, 2018, vol. 47.
[9] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 211–224, 2013.
[10] K. Huang, X. Fu, and N. D. Sidiropoulos, "Anchor-free correlated topic modeling: Identifiability and algorithm," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
[11] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 328–332, 2018.
[12] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Transactions on Signal Processing*, vol. 63, no. 9, pp. 2306–2320, 2015.
[13] K. Huang and X. Fu, "Detecting overlapping and correlated communities without pure nodes: Identifiability and algorithm," in *International Conference on Machine Learning*, 2019, pp. 2859–2868.
[14] G. Tatli and A. T. Erdogan, "Polytopic matrix factorization: Determinant maximization based criterion and identifiability," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5431–5447, 2021.
[15] J. Hu and K. Huang, "Identifiable bounded component analysis via minimum volume enclosing parallelotope," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.
[16] ——, "Complex bounded component analysis: Identifiability and algorithm," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.
[17] G. Pisier, *The volume of convex bodies and Banach space geometry*. Cambridge University Press, 1999, vol. 94.
[18] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.