# DICTIONARY IDENTIFIABILITY FROM FEW TRAINING SAMPLES

*Rémi Gribonval*

Projet METISS
Centre de Recherche INRIA Rennes - Bretagne Atlantique
IRISA, Campus de Beaulieu
F-35042 Rennes Cedex, France
E-mail: firstname.lastname@irisa.fr

*Karin Schnass*

Signal Processing Laboratory (LTS2)
School of Engineering, EPFL
Station 11, CH - 1015 Lausanne, Switzerland
E-mail: firstname.lastname@epfl.ch

## ABSTRACT

This article treats the problem of learning a dictionary providing sparse representations for a given signal class, via $\ell^1$ minimisation. The problem is to identify a dictionary $\mathbf{\Phi}$ from a set of training samples $Y$ knowing that $Y = \mathbf{\Phi}X$ for some coefficient matrix $X$. Using a characterisation of coefficient matrices $X$ that allow to recover any orthonormal basis (ONB) as a local minimum of an $\ell^1$ minimisation problem, it is shown that certain types of sparse random coefficient matrices will ensure local identifiability of the ONB with high probability, for a number of training samples which essentially grows linearly with the signal dimension.

## 1. INTRODUCTION

In the last years sparse signals have received a lot of attention as the signal processing community started to realise their usefulness. For instance they are easy to store and to compute with and recently it has been discovered that they are also quite easy to capture, using compressed sensing [6]. The drawback is that it is actually far from easy to find sparse representations. Assuming that someone just gives you a dictionary $\mathbf{\Phi}$ of $K$ atoms $\varphi_i \in \mathbb{R}^d$, a signal $y$ and the knowledge that this signal has an $S$-sparse representation, i.e. can be written as linear combination of $S$ atoms, the only way you can generically be guaranteed to find this sparse representation is to search among all $\binom{K}{S}$ subsets of $S$ atoms for the correct one. By now there are many results showing that by making additional assumptions on the dictionary, having low cumulative coherence [7, 10, 18] or satisfying a uniform uncertainty principle [3], sub-optimal algorithms like (Orthogonal) Matching Pursuit or algorithms based on the Basis Pursuit Principle, will give you the correct answer or be very likely to. However, in any of the cited publications you will more likely than not find a statement starting with 'given a dictionary ... ' which points exactly to the remaining problem. If you have a class of signals and you would like to find sparse approximations someone has to give you the right dictionary. For many signal classes good dictionaries like time-frequency or time scale dictionaries are known and from theoretical study of your signal class you might be able to identify one that will fit well. On the other hand, if you run into a new class of signals, chances that the best fit will already be known are quite slim and it can be quite a time consuming overkill to develop a deep theory like that of wavelets every time. An attractive alternative approach is dictionary learning, where you try to infer the dictionary that will give you good sparse representations for your whole signal class from a small portion of training signals.

Considering the extensive literature available for the sparse decomposition problem surprisingly little work has been dedicated to theoretical dictionary learning so far. There exist several dictionary learning algorithms [8, 13, 1], but only recently people have started to consider also the theoretical aspects of the problem. Dictionary learning finds its roots in the field of Independent Component Analysis (ICA) [4], where many identifiability results are available, which however rely on asymptotic statistical properties, under independence assumptions. Georgiev, Theis and Cichocki [9] as well as Aharon and Elad [2] describe more geometric identifiability conditions on the (sparse) coefficients of training data in an ideal (overcomplete) dictionary. Both approaches to the identifiability problem rely on rather strong sparsity assumptions, and require a huge amount of training samples. In addition to a theoretical study of dictionary identifiability, both cited papers provide theoretical algorithms to perform the desired identification. Unfortunately the naive implementation of these provably good dictionary recovery algorithms seems combinatorial, which limits their applicability to low dimensional data analysis problems and renders them fragile to outliers, i.e. training signals without a sparse enough representation. In this article we will study the question when a dictionary can be learned via $\ell^1$-minimisation [20, 17], and thus by a non-combinatorial algorithm. First we will shortly explain the minimisation problem that we use to find the dictionary $\mathbf{\Phi}$ from a set of training signals $y^n = \mathbf{\Phi}x^n, 1 \leq n \leq N$ (or in short $Y = \mathbf{\Phi}X$) and recent results [11], giving conditions on $X$ for the pair $(\mathbf{\Phi}, X)$ to be a local minimum of the minimisation problem, in case $\mathbf{\Phi}$ is an orthonormal basis (ONB). Then we will prove that if the entries of $X$ follow a certain type of sparse distribution these conditions will be satisfied with high probability. We quantify how rapidly this probability approaches one as the number $N$ of training signals grows. Denoting $p$ the proportion of zero entries in $X$, the number of training samples $N$ needed to guarantee the local identifiability condition does not grow significantly faster than $Kp^{-\gamma}$ for an exponent $2 \leq \gamma \leq 3$, i.e, for a fixed $p$ it essentially grows linearly with the dictionary size $K = d$.

## 2. DICTIONARY LEARNING VIA $\ell^1$-MINIMISATION

The idea of learning a dictionary via $\ell^1$-minimisation is motivated by the success of the Basis Pursuit principle for finding sparse representation. So given a dictionary, i.e. a set of $K \geq d$ unit vectors or atoms $\varphi_k \in \mathbb{R}^d$, $1 \leq k \leq K$, that span the whole space $\mathbb{R}^d$ and which we collect as columns in the

K. Schnass was partly supported by NSF grant 200021-117884/1. We would like to thank Roman Vershynin for essentially solving all our hypercubic sphere problems.

$d \times K$ matrix $\mathbf{\Phi}$, and a signal $y \in \mathbb{R}^d$, finding the sparsest representation amounts to solving the problem

$$\min_x \|x\|_0, \text{ such that } \mathbf{\Phi}x = y \tag{1}$$

where $\|x\|_0$ counts the number of nonzero entries in the vector $x$. Despite not being a norm $\|\cdot\|_0$ is often referred to as the $\ell^0$-norm. However, being nonconvex and nonsmooth, (1) is hard to solve. Enter Basis Pursuit, where we replace (1) by its convex relaxation,

$$\min_x \|x\|_1, \text{ such that } \mathbf{\Phi}x = y, \tag{2}$$

and hope that the solutions coincide. That this is actually the case whenever $y$ is sufficiently sparse can be retraced in several recent papers, e.g. [10, 7, 3, 18].

The connection to dictionary learning is now easily made. Given $N$ signals $y^n \in \mathbb{R}^d$, $1 \le n \le N$, and a candidate dictionary, we need to solve $N$ minimisation problems

$$\min_{x^n} \|x^n\|_1, \text{ such that } \mathbf{\Phi}x^n = y^n, \forall n.$$

Collect all signals $y^n$ into a $d \times N$ matrix $Y$ and all coefficients $x^n$ into a $K \times N$ matrix $X$ and define $\|X\|_1 := \sum_n \|x_n\|_1 = \sum_{k,n} |x_{kn}|$. Using this notation we can write the $N$ minimisation problems compactly as:

$$\min_X \|X\|_1, \text{ such that } \mathbf{\Phi}X = Y.$$

If the minimum is attained at $X_{\mathbf{\Phi}}$ then $\|X_{\mathbf{\Phi}}\|_1$ constitutes a measure of the global sparsity that can be achieved with the dictionary $\mathbf{\Phi}$. Thus a natural criterion to select the best dictionary within a collection $\mathscr{D}$ of admissible dictionaries is,

$$(\mathbf{\Phi}, X) = \arg\min_{\mathbf{\Phi},X} \|X\|_1, \text{ such that } \mathbf{\Phi}X = Y, \ \mathbf{\Phi} \in \mathscr{D}. \tag{3}$$

The most general families of admissible dictionaries one can imagine are the ones where just the number of atoms is fixed. However, the more general $\mathscr{D}$ is, the harder it is to find a minimum simply because more dictionaries have to be considered. To simplify the search one can concentrate on more structured families such as discrete libraries of orthonormal bases (wavelet packets or cosine packets, for which fast dictionary selection is possible using tree-based searches) or structured overcomplete dictionaries such as shift-invariant dictionaries or unions of orthonormal bases. In this paper we will focus on the simplest non-overcomplete case ($K = d$) with the set $\mathscr{O}(d)$ of arbitrary *orthogonal bases*, parameterised by a unitary matrix $\mathbf{\Phi}$. Further work is needed to check how to extend our results to the set of *oblique bases*, associated to square matrices $\mathbf{\Phi}$ with linearly independent unit columns $\|\varphi_k\|_2 = 1$ , or even to overcomplete dictionaries.

The special aspect of dictionary learning treated here is how a coefficient matrix $X$ has to be structured such that for any orthonormal basis $\mathbf{\Phi}$ the pair $(\mathbf{\Phi}, X)$ will constitute a global minimum of (3) with input $Y = \mathbf{\Phi}X$. In other words when can a dictionary be uniquely identified from $N$ sparse training signals $y^n$ by $\ell^1$ minimisation. However since the minimisers of (3) are only unique up to matching column (resp. row) permutation and sign change of $\mathbf{\Phi}$ (resp. $X$), and also because it is generally hard to find global minima, we will reduce our ambition to finding conditions such that $(\mathbf{\Phi}, X)$ constitutes a *local* minimum, which we will call *local identifiability conditions*. They guarantee that algorithms which decrease the $\ell^1$ norm must converge to the true dictionary when started from a sufficiently close initial condition.

## 3. LOCAL IDENTIFIABILITY CONDITION

As starting point for our analysis that certain random sparse matrices will have the required structure, we use the result developed in [11]. The local identifiability condition is expressed based on a block decomposition of the coefficient matrix $X$ as follows (see Figure 1):

- $x_k$ is the $k$-th row of $X$, and we define $\Lambda_k$ the set indexing its nonzero entries and $\overline{\Lambda}_k$ the set indexing its zero entries;
- $s_k$ is the row vector $\text{sign}(x_k)_{\Lambda_k}$;
- $X_k$ (resp. $\bar{X}_k$) is the matrix obtained by removing the $k$-th row of $X$ and keeping only the columns indexed by $\Lambda_k$ (resp. $\overline{\Lambda}_k$) .

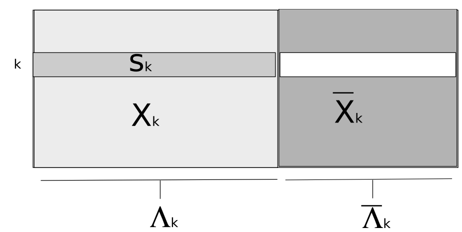


Figure 1: Block decomposition of the matrix $X$ with respect to a given row $x_k$. Without loss of generality, the columns of $X$ have been permuted so that the first $\sharp\Lambda_k$ columns hold the nonzero entries of $x_k$ while the last $\sharp\overline{\Lambda}_k$ hold its zero entries.

**Theorem 3.1 ([11])** *Consider a $K \times N$ matrix $X$. Assume that for each $k$, there exists a vector $d_k$ with*

$$\bar{X}_k d_k = X_k s_k^T \textit{ and } \|d_k\|_\infty < 1. \tag{4}$$

*Then, for any* orthogonal *matrix $\mathbf{\Phi}$, the optimisation problem*

$$\min_{\mathbf{\Phi}', X'} \|X'\|_1, \textit{such that } \mathbf{\Phi}'X' = \mathbf{\Phi}X,$$

*where $\mathbf{\Phi}'$ is constrained to be any basis of unit vectors (i.e., not necessarily orthonormal but* oblique*), admits a strict local minimum at $\mathbf{\Phi}' = \mathbf{\Phi}$.*

Note that an ONB $\mathbf{\Phi}$ in combination with any $X$ exhibiting the above property, will be a local minimum not only among all pairs of ONBs and coefficients but among all pairs of oblique bases and coefficients.

## 4. RANDOM SPARSE MODEL ON $X$

We now detail the random sparse model on $X$ and outline the proof that, when the number of training samples $N$ is large, the local identifiability condition of Theorem 3.1 is satisfied with high probability. We will merely sketch the estimation of the small probability that the condition is not satisfied.

### 4.1 The model

We assume that the entries $x_{kn}$ of the $K \times N$ matrix $X$ are i.i.d with $x_{kn} = z_{kn} w_{kn}$ where $z_{kn}$ are i.i.d indicator variables taking the value zero with probability $0 < p < 1$, i.e. $z \sim (1-p)\delta_1 + p\delta_0$, and $w_{kn}$ are i.i.d., centered, of unit variance.

The important role of the indicator variables is to guarantee a strictly positive probability that $x_{kn}$ is exactly zero. The distribution of $w_{kn}$ seems to play a less important role. Here we will assume that this distribution is "subgaussian with parameter $\beta$", in the sense that

$$\mathbb{P}_w(|w| > u) \leq \exp(1 - u^2/\beta^2),\ \forall u > 0. \tag{5}$$

Examples of distributions which fit this model are when $w_{kn}$ is Gaussian, or Bernoulli $\pm 1$ with equal probability. The subgaussian assumption will be used as a technical assumption in the analysis carried below, but we believe that similar results can also be achieved with other distributions such as the Laplacian distribution, which is not subgaussian and seems more natural in the $\ell^1$ minimisation framework.

### 4.2 Geometric insight

For each index $k$ we need to check if there is a vector $d_k$ with $\|d_k\|_\infty < 1$ such that $\bar{X}_k d_k = X_k s_k^T$. Geometrically speaking, we need to verify if the vector $u_k := X_k s_k^T$ lies in the image by the linear operator $\bar{X}_k$ of the unit cube $Q^{\bar{\Lambda}_k} = [-1,\ 1]^{\bar{\Lambda}_k}$. This will be true whenever we have simultaneously that:

- the vector $u_k$ belongs to the Euclidean ball $B(0,r)$ of radius $r$, i.e., $\|u_k\|_2 \leq r$;
- the image of the unit cube $Q^{\bar{\Lambda}_k}$ by $\bar{X}_k$ contains $B(0,r)$.

### 4.3 Outline of the approach

To achieve our goal, we will prove that:

**P1** with high probability $1 - P_1$, the matrix $X_k$ has roughly $(1-p) \times N$ columns, and $\bar{X}_k$ roughly $p \times N$ columns.

**P2** with probability $1 - P_2(\alpha)$, we have $\|u_k\|_2^2 \leq \alpha(K-1)N$. This can be seen from the fact that $u_k$ is a sum of at most $N$ i.i.d zero-mean vectors (the columns of $X_k$ multiplied by independent random signs), each with expected squared norm $(K-1) \times (1-p) \leq K-1$. The probability $P_2(\alpha)$ decays exponentially fast to zero with $\alpha$. For technical reasons we choose

$$\alpha = \alpha(p,K,N) = 4 \log \frac{p^{2+\frac{1}{K-1}} N}{K-1}.$$

**P3** with high probability $1 - P_3$, we have the inclusion

$$B(0, \sqrt{\alpha \cdot (K-1) \cdot N}) \subset \bar{X}_k Q^{\bar{\Lambda}_k}.$$

In the appendix we provide the main ideas indicating why the three steps **P1-3** are valid.

## 5. QUALITATIVE BEHAVIOUR

The overall probability that the coefficient matrixs $X$ satisfies the local identifiability condition of Theorem 3.1 is driven by $P_1$, $P_2$, $P_3$. The sketches of the proofs provided in the appendix indicate that while $P_1$ decreases exponentially fast with $N$, $P_2$ and $P_3$ do not decay as fast with $N$, and $P_2$ is dominated by $P_3$. Globally, the order of magnitude of the overall probability decays is at least as fast as

$$[4 \log f(p,K,N)]^K \cdot f(p,K,N)^{-(K-1)}$$

with

$$f(p,K,N) := p^{2+\frac{1}{K-1}} N/(K-1).$$

In other words, there is a constant $C$ such that whenever the number of training samples satisfies

$$N \geq (K-1) \cdot p^{-2-\frac{1}{K-1}} \cdot A$$

for some value $A$, the probability that $X$ does not satisfy the local identifiability condition does not exceed $C(4 \log A)^K A^{-(K-1)}$.

This behaviour is good news for two reasons. Firstly, for a given proportion $p$ of zero entries in $X$, the number $N$ of training samples that is sufficient to guarantee with high probability local stability of the $\ell^1$ learning criterion only grows linearly with the ambient signal dimension $K$. Secondly, even for small $p$ - i.e. for not really sparse matrices $X$ having relatively few zero coefficients - local identifiability with the $\ell^1$ minimisation criterion does not require exponentially many training samples, but rather. Indeed, for the smallest possible dimension of a dictionary learning problem, $K = 2$, $N \geq p^{-3}A$, for large $A$, is sufficient. For dictionary learning problems in higher dimensions, $K \gg 2$, the number of training samples only needs to grow like $N/(K-1) \geq p^{-2}A$. If $p^2 N/(K-1) \gg 1$ the probability that $X$ yields a local minimum of the $\ell^1$ criterion rapidly approaches one.

## 6. CONCLUSION

We have shown that coefficient matrices with entries following a sparsely scaled Gaussian distribution make it possible to identify an arbitrary orthonormal basis from $N$ training signals as a local minimum of an $\ell^1$ minimisation problem (3). This holds with probability rapidly approaching one as the number of training signals is growing large compared to their dimensionality and their expected sparsity, i.e.

$$N \gg K p^{-\gamma}, \quad \text{with } 2 \leq \gamma \leq 3$$

Since the dependence is linear in $K$ and inversely sub-cubic in $p$ we need far less training samples to have good recovery chances than suggested for instance by Aharon et al. in [2], and local identifiability can also be guaranteed even though many training samples have no sparse representation.

However with this result we have barely started to scratch the surface of the theoretical aspects of the dictionary learning problem with finitely many training samples, and much work will have to be invested into its extension. First of all we need to investigate in more depth for which distributions on $w_{kn}$ the current type of analysis is valid. This is deeply connected with the properties of random projections of the high-dimensional unit cube under $X$. The next step is dealing with oblique bases. Results in [11] indicate that this implies taking into account the coherence of $\mathbf{\Phi}$ in the concentration of measure arguments sketched here. In order to make the result more practically applicable to dictionary learning we need to analyse the probability that spurious local minima of the $\ell_1$ criterion exist. If they do not exist, descent algorithms are bound to converge to the optimal dictionary for any initial condition. This is similar in spirit to the work of Vrins et al [19]. Also it would be desirable to extend Theorem 3.1 to redundant dictionaries, and analyse dictionary learning with criteria which mix an $\ell^1$ term with an quadratic approximation error to account for noise in the model. Finally we want to explore whether the recovery condition of Theorem 3.1 can be used to prove the optimality of other, more greedy, dictionary learning algorithms. Preliminary results indicate that this is the case for the "deflation" approach [5].

## A. PROOF SKETCHES

**Proof:** (sketch of **P1**). The first statement **P1** amounts to measuring the probability that the sum $\sharp\Lambda_k = \sum_{n=1}^N z_{kn}$ of $N$ i.i.d variables $z_{kn} \sim (1-p)\delta_1 + p\delta_0$ deviates from its expected value $(1-p)N$ by more than a given factor. Using Hoeffding's inequality we get for $0 < \varepsilon_1 < 1/2$

$$\mathbb{P}(\sharp\Lambda_k \geq (1+\varepsilon_1)(1-p)N) \leq \exp(-(\frac{\varepsilon_1^2}{4}(1-p)N) \quad (6)$$

$$\mathbb{P}(\sharp\bar{\Lambda}_k \leq (1-\varepsilon_1)pN) \leq \exp(-(\frac{\varepsilon_1^2}{4}(1-p)N). \quad (7)$$

For a fixed $p < 1$ this probability $P_1$ decays exponentially fast with $N$, and will be negligible compared to $P_2$ and $P_3$.

**Proof:** (sketch of **P2**). In case the entries of the coefficient matrix follow a scaled Gaussian distribution the second statement **P2** essentially corresponds to bounding the tail of a $\chi^2$-distribution. Let $A$ be an $L \times M$ matrix with entries $a_{lm} = w_{lm} z_{lm}$ where $w_{lm}$ are i.i.d normally distributed and $z_{lm}$ are i.i.d indicator variables, as described in the signal model, and $s$ an $M$-dimensional vector with independent Bernoulli entries ($\pm 1$), which is independent of $A$. We want to bound the tail of the random variable

$$\|As\|_2^2 = \sum_{l=1}^L (\sum_{m=1}^M w_{lm} z_{lm} s_m)^2 =: \sum_{l=1}^L Y_l^2.$$

For fixed indicator variables, $Y_l = \sum_{m=1}^M w_{lm} z_{lm} s_m$ is a sum of i.i.d zero mean, unit variance Gaussian random variables, hence it is again Gaussian with zero mean and variance $\|z_l\|_2^2$. Thus $Y_l/\|z_l\|_2$ is Gaussian with zero mean and unit variance and $\sum_l Y_l^2/\|z_l\|_2^2$ follows a $\chi^2$-distribution of degree L. Observing that $\|z_l\|_2^2 \leq M$ we obtain, as soon as $\alpha/\log(\alpha L) \geq 2$,

$$\begin{aligned}\mathbb{P}\Big(\sum Y_l^2 > \alpha \cdot L \cdot M\Big) &\leq \mathbb{P}\Big(\sum Y_l^2/\|z_l\|_2^2 > \alpha L\Big) \\ &= \frac{2^{-L/2}}{\Gamma(L/2)} \int_{\alpha L}^{\infty} x^{(L/2-1)} e^{-x/2} dx \\ &\leq \int_{\alpha L}^{\infty} e^{-x/4} dx = 4e^{-\alpha L/4}. \quad (8)\end{aligned}$$

The last estimate we get since for $x \geq \alpha L$ we have $x^{(L/2-1)} e^{-x/2} < e^{-x/4}$, because $x/\log x \geq \alpha L/\log(\alpha L) \geq L-2$. With $A = X_k$, $s = s_k$, we have $L = K-1$ and $M \leq N$, so with the chosen value for $\alpha$ we obtain

$$P_2 \leq C_2 \left(\frac{p^{2+\frac{1}{K-1}} N}{K-1}\right)^{-(K-1)},$$

whenever $p^{2+\frac{1}{K-1}} N/(K-1) \geq c_2$ for a universal constant $c_2$.

**Proof:** (sketch of **P3**). The third statement is strongly connected to the notion of Kashin's representations [12, 14], and its analysis is more involved. Given $M$ vectors $\{v_m\}_{m=1}^M \subset \mathbb{R}^n$, which we can collect in an $n \times M$ matrix $\mathbf{V}$, one says that the vector $a \in \mathbb{R}^M$ is a Kashin's representation of level $C$ of the vector $u \in \mathbb{R}^n$ with $\mathbf{V}$ if $u = \frac{1}{\sqrt{M}} \sum_{m=1}^M a_m v_m$, $\|a\|_\infty \leq C\|u\|_2$. The two following statements are equivalent: (a) *every vector $u \in \mathbb{R}^n$ admits a Kashin's representation of level $C$ with $\mathbf{V}$*; (b) the matrix $\mathbf{V}$ satisfies $B(0, \sqrt{M}/C) \subset \mathbf{V}Q^M$. Random matrices $\mathbf{V}$ with i.i.d. subgaussian entries satisfy the above property with high probability [16, 15, 14], which is why we introduced the subgaussian assumption on $w$ (see Eq. (5)). This immediately yields the following lemma (proved below) giving properties of the $(K-1) \times \sharp\Omega$ matrix $X_\Omega^k := (x_{\ell n})_{\ell \neq k, n \in \Omega}$:

**Lemma A.1** *Let $0 < p_0 < 1$. There are constants $\lambda_0 > 2, c_3, C_3$ with the following properties: for any $0 < p \leq p_0$ and any index set $\Omega$, if $\lambda := \sharp\Omega/(K-1) \geq \lambda_0$ then, except with probability at most*

$$\lambda^{-(K-1)} + 2\exp(-c_3 \sharp\Omega),$$

*every $u \in \mathbb{R}^{K-1}$ admits a representation $u = X_\Omega^k d$ with*

$$\|d\|_\infty \leq C_3 \cdot \frac{\|u\|_2}{\sqrt{(1-p)(K-1)\sharp\Omega}}. \quad (9)$$

This lemma is slightly too weak to be applied directly in our setting: typically we expect $\|u_k\|_2 \leq \sqrt{\alpha(K-1)N}$ and $\sharp\bar{\Lambda}_k \geq (1-\varepsilon_1)pN$, hence the lemma applied to $\Omega := \bar{\Lambda}_k$ provides a representation $u_k = \bar{X}_k d_k$ with $\|d_k\|_\infty \leq C_3' \cdot \sqrt{\alpha/p}$ with $C_3' = C_3((1-\varepsilon_1)(1-p))^{-1/2}$. Since $\alpha$ grows to infinity with $N$, this is not enough to obtain the desired result. However, if we can split $\bar{X}_k = X_\Omega^k$ into $L$ disjoint matrices $X_{\Omega_\ell}^k$ with $\sharp\Omega_\ell \geq \sharp\Omega/(2L)$ which all lead to representations $u_k = X_{\Omega_\ell}^k d_k^\ell$ satisfying (9), then it is possible to combine these representations as $u_k = \frac{1}{L}\sum_{\ell=1}^L X_{\Omega_\ell}^k d_k^\ell = \bar{X}_k d_k$ with

$$\begin{aligned}\|d_k\|_\infty &\leq \frac{\max_\ell \|d_k^\ell\|_\infty}{L} \leq \frac{C_3}{L \cdot \sqrt{\sharp\Omega_\ell}} \frac{\|u_k\|_2}{\sqrt{(1-p)(K-1)}} \\ &\leq \frac{C_3}{\sqrt{L/2}} \cdot \frac{\|u_k\|_2}{\sqrt{(1-p)(K-1)\sharp\Omega}}\end{aligned}$$

When $\|u_k\|_2$ and $\sharp\bar{\Lambda}_k$ have their typical values given by **P1-2**, we obtain

$$\|d_k\|_\infty \leq C_3' \cdot \sqrt{2\alpha/(pL)}.$$

Taking $L > 2(C_3')^2 \cdot \alpha/p$ yields $\|d_k\|_\infty < 1$ as desired. The probability that these $L$ matrices $X_{\Omega_\ell}^k$ do not simultaneously satisfy the desired Kashin's representation property is at most

$$P_3 \leq \sum_{\ell=1}^L (\lambda_\ell^{-(K-1)} + 2\exp(-c_3 \sharp\Omega_\ell)), \quad (10)$$

provided that for each block $\lambda_\ell := \sharp\Omega_\ell/(K-1) \geq \lambda_0$.

We now focus on orders of magnitude to estimate $P_3$, and assume $L \approx \alpha/p$, which means there are two constants $0 < c \leq C < \infty$ such that $cL \leq \alpha/p \leq CL$. If these constants are sufficiently large, we can choose the partition of

$\Omega$ such that all blocks have approximately the same size $\sharp\Omega_\ell \approx \sharp\Omega/L \approx pN/L \approx p^2N/\alpha$, hence $\lambda_\ell \approx p^2N/\alpha(K-1)$. The exponential term in (10) is dominated by the polynomial one, hence $P_3$ is bounded by

$$\frac{C'\alpha}{p}\left(\frac{p^2N}{\alpha(K-1)}\right)^{-(K-1)} = C'\left(\frac{p^{2+\frac{1}{K-1}}N/(K-1)}{\alpha^{1+\frac{1}{K-1}}}\right)^{-(K-1)}. \tag{11}$$

**Proof:** (Lemma A.1). Notice that since $w_{kn}$ is subgaussian with parameter $\beta$, so is $x_{kn}$. Moreover, $\mathbb{E}(x_{kn}^2) = (1-p)\mathbb{E}(w_{kn}^2) = 1-p$. First, we consider the matrix $\mathbf{\Psi} := (1-p)^{-1/2}X_\Omega^k$: its entries are independent, zero mean, subgaussian with parameter $\beta' := \beta(1-p)^{-1/2}$ and variance 1. We can therefore apply [14, Lemma 4.8] to conclude that the columns of the matrix $\mathbf{V} := \frac{1}{\sqrt{\sharp\Omega}}\mathbf{\Psi}$ form an $\varepsilon$-tight frame, except with small probability at most $2\exp(-c_3(\beta')\sharp\Omega\varepsilon^2)$, as soon as $\lambda := \sharp\Omega/(K-1) > \lambda_1 := \frac{C_3(\beta')}{\varepsilon^2}\log\frac{2}{\varepsilon}$. The dependence of $c_3(\beta')$ and $C_3(\beta')$ is polynomial, so they can be replaced with universal constants $c_3(\beta)$ and $C_3(\beta)$ independent of $p$ for $0 \le p \le p_0$. Next, we apply [14, Theorem 4.6]: provided that $\lambda \ge 2$, except with probability at most $\lambda^{-(K-1)}$ the matrix $\mathbf{V}' := \frac{1}{\sqrt{\sharp\Omega}}X_\Omega^k$ satisfies the "Uncertainty Principle with parameters $\eta = C_4\beta\sqrt{\log\lambda/\lambda}$ and $\delta = c_4/\lambda$", that is to say: $\|\mathbf{V}'d\|_2 \le \eta\|d\|_2$ for all $d \in \mathbb{R}^M$ such that $\sharp\mathrm{supp}(d) \le \delta M$. The constants $c_4$ and $C_4$ are universal. It follows that, except with probability at most $\lambda^{-(K-1)} + 2\exp(-c_3\varepsilon^2\sharp\Omega)$, the matrix $\mathbf{V} := (\sharp\Omega)^{-1/2}\mathbf{\Psi} = (1-p)^{-1/2}\mathbf{V}' = ((1-p)\sharp\Omega)^{-1/2}X_\Omega^k$ is an $\varepsilon$-tight frame *and* satisfies the UP with parameters $\eta(1-p)^{-1/2}$ and $\delta$. Obviously, there is some $\lambda_2$ such that if $\lambda > \lambda_2$, $\eta' := \sqrt{1+\varepsilon}(1-p)^{-1/2}\eta + \varepsilon < 1$, therefore if $\lambda > \max(2,\lambda_1,\lambda_2)$ we can apply [14, Theorem 3.9] to conclude that each vector $u \in \mathbb{R}^{K-1}$ admits a Kashin's representation of level $C := (1-\eta')^{-1}\delta^{-1/2}$ with $\mathbf{V}$, i.e. : $u = \frac{1}{\sqrt{\sharp\Omega}}\sum_{n\in\Omega} a_n v_n = \frac{1}{\sqrt{1-p\cdot\sharp\Omega}}X_\Omega^k a = X_\Omega^k d$ with

$$\|d\|_\infty = \frac{\|a\|_\infty}{\sqrt{1-p}\cdot\sharp\Omega} \le \frac{\|u\|_2}{(1-\eta')\cdot\sqrt{1-p}\cdot\sqrt{\delta}\cdot\sharp\Omega}$$

To conclude, we write

$$\sqrt{\delta}\cdot\sharp\Omega = \sqrt{c_4/\lambda}\cdot\lambda(K-1) = \sqrt{c_4(K-1)}\sqrt{\sharp\Omega}.$$

## REFERENCES


[1] M. Aharon, M. Elad, and A.M. Bruckstein. The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, November 2006.

[2] M. Aharon, M. Elad, and A.M. Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Journal of Linear Algebra and Applications.*, 416:48–67, July 2006.

[3] E. J. Candès, J. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math*, 59:1207–1223, 2006.

[4] J.-F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE. Special issue on blind identification and estimation*, 9(10):2009–2025, October 1998.

[5] N. Delfosse and P. Loubaton. Adaptive separation of independent sources: a deflation approach. In *Proceedings of the IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP'94)*, volume 04, pages 41–44, April 1994.

[6] David L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[7] D.L. Donoho and M. Elad. Maximal sparsity representation via $\ell^1$ minimization. *Proc. Nat. Aca. Sci.*, 100(5):2197–2202, March 2003.

[8] D.J. Field and B.A. Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[9] P. Georgiev, F.J. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks*, 16(4):992–996, 2005.

[10] Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, 49(12):3320–3325, December 2003.

[11] Rémi Gribonval and Karin Schnass. Some recovery conditions for basis learning by l1-minimization. In *3rd IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP 2008)*, St. Julians, Malta, March 2008.

[12] B. Kashin. Sections of some finite dimensional sets and classes of smooth functions. *Izv. Akad. Nauk SSSR Ser. Mat. (Russian)*, 41:334–351, 1977.

[13] 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 Comput.*, 15(2):349–396, 2003.

[14] Y. Lyubarskii and R. Vershynin. Uncertainty principles and vector quantization. *ArXiv Mathematics e-prints*, November 2006.

[15] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principles for Bernoulli and subgaussian ensembles. Technical report, submitted.

[16] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators. *Geometric and Functional Analysis*, to appear.

[17] Mark D. Plumbley. Dictionary learning for l1-exact sparse coding. In Mike E. Davies, Christopher J. James, Samer A. Abdallah, and Mark D Plumbley, editors, *Independent Component Analysis and Signal Separation*, volume 4666 of *LNCS*, pages 406–413. Springer, 2007.

[18] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Information Theory*, 52(3):1030–1051, March 2006.

[19] F. Vrins, D.-T. Pham, and M. Verleysen. Mixing and non-mixing local minima of the entropy contrast for blind source separation. *IEEE Transactions on Information Theory*, 53(3):1030–1042, March 2007.

[20] M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, 2001.