Some recovery conditions for basis learning by L1-minimization

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 Institute (ITS),
School of Engineering, EPFL
Station 11, CH - 1015 Lausanne,Switzerland
E-mail firstname.lastname@epfl.ch

***Abstract*—Many recent works have shown that if a given signal admits a sufficiently sparse representation in a given dictionary, then this representation is recovered by several standard optimization algorithms, in particular the convex ℓ^1 minimization approach. Here we investigate the related problem of infering the dictionary from training data, with an approach where ℓ^1-minimization is used as a criterion to select a dictionary. We restrict our analysis to basis learning and identify necessary / sufficient / necessary and sufficient conditions on ideal (not necessarily very sparse) coefficients of the training data in an ideal basis to guarantee that the ideal basis is a strict local optimum of the ℓ^1-minimization criterion among (not necessarily orthogonal) bases of normalized vectors. We illustrate these conditions on deterministic as well as toy random models in dimension two and highlight the main challenges that remain open by this preliminary theoretical results.**

***Index Terms*—Sparse representation, dictionary learning, nonconvex optimization, independent component analysis.**

I. Introduction

Many signal processing tasks, such as denoising and compression, can be efficiently performed if one knows a sparse representation of the signals of interest. Moreover, a huge body of recent results on sparse representations have highlighted their impact in inverse linear problems such as (blind) source separation and compressed sampling.

All applications of sparse representations rely on a signal dictionary from which sparse linear expansions can be built to efficiently approximate the signals from a class of interest. Success heavily depends on the good fit between the data class and the dictionary. For many signal classes, good dictionaries – such as time-frequency or time-scale dictionaries – are now well known, but new data classes may require the construction of new dictionaries to fit new types of data features.

The analytic construction of dictionaries such as wavelets and curvelets stems from deep mathematical tools from Harmonic Analysis. It may however be difficult and long to develop new mathematical know-how each time a new class of data is met which requires a different type of dictionary. An alternative approach is dictionary learning, which aims at infering the dictionary from a set of training data. Dictionary learning, also known as *sparse coding*, has the potential of somehow "industrializing" sparse representation techniques for new data classes.

Learning a dictionary is a task deeply related to vector quantization and blind source separation / independent component analysis (ICA). Even though several dictionary learning algorithms [1], [2], [3] have been proposed, there is relatively few work dedicated to the theoretical aspects of the problem. This is in high contrast to the sparse signal decomposition problem with a given dictionary, which has generated a huge literature analyzing the provable performance of the main algorithms of the field, Basis Pursuit (ℓ^1-minimization) [4], [5], [6], [7] and greedy algorithms / matching pursuit.

In ICA, the dictionary is more commonly called a mixing matrix, and the identifiability (up to gain and permutations) of this matrix, in a statistical sense, is the core result of ICA. Such an identifiability is formally proved using the uniqueness of the factorization of the joint probability density function (PDF) of independent non-Gaussian random variables (called sources) as a product of their marginal PDF. In practice, given a finite number of training examples, one can only access an empirical estimate of the joint PDF, which somehow questions the relevance of such identifiability results. In the orthogonal case, Cardoso [8] discusses the effect of finite training sample size to derive lower bounds on the achievable identification performance of several families of ICA algorithms.

More recently, in the specific case of sparse sources, Georgiev, Theis and Cichocki [9] as well as Aharon and Elad [10] exhibited geometric identifiability conditions on the (sparse) coefficients of training data in an ideal dictionary – which is possibly overcomplete, that is to say with n sources and m mixtures and $m < n$. Both approaches to the identifiability problem rely on rather strong sparsity assumptions, and in particular require that for each training sample, the number of active sources is at most $m - 1$. In addition to a theoretical study of dictionary identifiability, these works also provide algorithms to perform the desired identification. Unfortunately the naive implementation of these provably good dictionary recovery algorithms seems combinatorial – limiting their applicability to low dimensional data analysis problems– and fragile to outliers (training examples with no sparse enough representation).

K. Schnass contributed to this work while visiting IRISA with support from the program INRIA - Equipe Associée "SPARS". The authors would like to thank Pierre Vandergheynst, Jean-Jacques Fuchs and Gabriel Peyré for enlightening discussions about the dictionary learning problem studied here.

978-1-4244-1688-2/08/$25.00

In this note, we study the possibility to design provably good, non-combinatorial dictionary learning algorithms that are robust to outliers. Inspired by the recent results on the provably good properties of ℓ^1-minimization for sparse signal decomposition with a given dictionary, we investigate the properties of ℓ^1-based dictionary learning [11], [12]. Our goal is to characterize properties that a set of training samples should satisfy to guarantee that an ideal dictionary is a local minimum – or, even better, a global minimum – of the ℓ^1 criterion, opening the possibility to replace combinatorial learning algorithms with provably good descent techniques. We limit here our characterization to *basis* learning, and provide examples in $\mathbb{R}^2$ to illustrate the robustness of ℓ^1-based basis learning to non-sparse outliers.

II. Sparsity and dictionary learning

In the vector space $\mathcal{H} = \mathbb{R}^d$ of d-dimensional signals, a dictionary is a collection of $K \geq d$ unit vectors φ_k, $1 \leq k \leq K$ which span the whole space. Alternatively, a dictionary can be seen as a $d \times K$ matrix $\mathbf{\Phi}$ with unit columns. For a given signal $y \in \mathcal{H}$, the sparse representation problem consists in finding a representation $y = \mathbf{\Phi} \cdot x$ where $x \in \mathbb{R}^K$ is a "sparse" vector, *i.e.* with few significantly large coefficients and most of its coefficients negligible.

A. Sparse representation by ℓ^1-minimization

For a given dictionary, selecting an "ideal" sparse representation of some data vector $y \in \mathcal{H}$ amounts to solving the problem

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

where $\|x\|_0$ ($\|\cdot\|_0$ is often referred to as the ℓ^0 " norm, although it is definitely not a norm) counts the number of nonzero entries in the vector x. However, being nonconvex and nonsmooth, (1) is hard to solve. The good news brought by a lot of recent work [4], [5], [6], [7] is that when y admits a sufficiently sparse representation, it is unique and can be recovered by solving the convex optimization problem

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

B. Dictionary learning

A related problem is that of finding the dictionary that will fit a class of signals, in the sense that it will yield an optimal tradeoff to jointly provide sparse representations of all signals from the class. Given N signals $y^n \in \mathcal{H}$, $1 \leq n \leq N$, and a candidate dictionary, one can measure the global sparsity as

$$\sum_{n=1}^{N} \min_{x^n} \|x^n\|_1, \text{ such that } \mathbf{\Phi} x^n = y^n, \forall n.$$

Collecting all signals y^n into a $d \times N$ matrix Y and all coefficients x^n into a $K \times N$ matrix X, the fit between a dictionary $\mathbf{\Phi}$ and the training signals Y can be measured by

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

To select a dictionary $\mathbf{\Phi}$ within a collection $\mathcal{D}$ of admissible dictionaries, one can therefore consider the criterion

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

Several families of admissible dictionaries can be considered 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 orhonormal bases. Here we focus on the non-overcomplete case ($K = d$) with : a) the set $\mathcal{O}(d)$ of arbitrary *orthogonal bases*, parameterized by a unitary matrix $\mathbf{\Phi}$; b) the set $\mathcal{B}(d)$ of what we will call *oblique bases*, associated to square matrices $\mathbf{\Phi}$ with linearly independent and unit columns $\|\varphi_k\|_2 = 1.$,

C. Dictionary recovery

Several algorithms have been proposed which adopt a similar approach to learning a dictionary [1], [2], [12] from training data. Here we are interested in the *dictionary identifiability problem*: assuming that the data Y were generated from an "ideal" dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$ and "ideal" coefficients X_0, we wish to determine conditions on X_0 (and possibly $\mathbf{\Phi}_0$) such that the minimization of (3) recovers $\mathbf{\Phi}_0$.

Ideally, we want to identify conditions on the sparse coefficients X_0 such that, for any dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$, if we observe $Y = \mathbf{\Phi}_0 X_0$, then solving (3) will recover $\mathbf{\Phi}_0$ (and X_0). In other words, we want to characterize coefficient matrices X_0 such that, for any $\mathbf{\Phi}_0 \in \mathcal{D}$, the global minimum of

$$\min_{\mathbf{\Phi}, X} \|X\|_1, \text{ such that } \mathbf{\Phi} X = \mathbf{\Phi}_0 X_0, \ \mathbf{\Phi} \in \mathcal{D}. \tag{4}$$

can only be found at $\mathbf{\Phi} = \mathbf{\Phi}_0$. However, the minimizers of (3) are social animals which live in herds, each herd corresponding to a large equivalence class with respect to matching column (resp. row) permutation and sign change of $\mathbf{\Phi}$ (resp. X). Therefore, we shall relax our requirement and only ask to find conditions such that the global minima of (4) are (only) at one of these equivalent solutions. Our objective is therefore similar in spirit to previous work on dictionary recovery [9], [10] which studied the uniqueness property. The main difference is that here we specify in advance which optimization criterion we wish to use to recover the dictionary (ℓ^1-minimization) and attempt to express necessary (resp. sufficient) conditions on a matrix X_0 to guarantee that this method will successfully recover a given class of dictionaries.

Unfortunately, in the study of the ℓ^1-minimization based dictionary recovery problem, several difficulties arise at once, some due to the possible overcompleteness and non-orthogonality of the dictionary, others due to the difficulty of globally characterizing the optima of a globally nonconvex problem which we know admits exponentially many solutions because of the permutation and sign indeterminacies. To keep technicalities to a minimum we restrict our study to the characterization of *local* minima of the objective ℓ^1 criterion for orthonormal (resp. oblique) bases. We will see that in specific circumstances, global minima will be kind enough to accept to be captured.

III. Local minima of the ℓ^1 learning criterion

Given an invertible matrix $\mathbf{\Phi}_0$, the constraint $\mathbf{\Phi}X = \mathbf{\Phi}_0 X_0$ (with $\mathbf{\Phi}$ an invertible matrix) is equivalent to $X = \mathbf{\Phi}^{-1}\mathbf{\Phi}_0 \cdot X_0$, hence (4) is expressed equivalently but perhaps more simply as

$$\min_{\mathbf{\Phi}\in\mathcal{D}} \|\mathbf{\Phi}^{-1}\mathbf{\Phi}_0 X_0\|_1. \tag{5}$$

A. Characterization of local minima

The study of local minima of (5) uses the notion of the tangent plane $d\mathcal{D}_{\mathbf{\Phi}_0}$ to the manifold $\mathcal{D}$ at the point $\mathbf{\Phi}_0$. The tangent plane is the collection of the derivatives $\mathbf{\Phi}' := \mathbf{\Phi}'(0)$ of all smooth functions $\epsilon \mapsto \mathbf{\Phi}(\epsilon)$ which satisfy $\forall\epsilon, \mathbf{\Phi}(\epsilon) \in \mathcal{D}$ and $\mathbf{\Phi}(0) = \mathbf{\Phi}_0$. Before characterizing the tangent plane to both manifolds $\mathcal{B}(d)$ and $\mathcal{D} = \mathcal{O}(d)$, we need some additional notations to state the main lemma. We denote by $\overline{\Lambda}^n$ the set indexing the zero entries of the n-th column x^n of X_0, and $\overline{\Lambda} = \{(n,k), 1 \le n \le N, k \in \overline{\Lambda}^n\}$ the set indexing all zero entries in X_0. We let $\langle A, B\rangle_F = \text{Trace}(A^T B)$ denote the natural inner product between matrices, which is associated to the Froebenius norm $\|A\|_F^2 = \langle A, A\rangle_F$, and $\text{sign}(A)$ is the sign operator applied componentwise to the matrix A (by convention $\text{sign}(0) := 0$). With this notation we are now in a position to state the main lemma. Note that the proofs of all lemmata are gathered in the Appendix.

Lemma 3.1: Assume that X_0 has at least one zero entry ($\overline{\Lambda} \neq \emptyset$), and consider a basis $\mathbf{\Phi}_0$.

1) If for all nonzero $\mathbf{\Phi}' \in d\mathcal{D}_{\mathbf{\Phi}_0}$, we have
$$\left|\langle \mathbf{\Phi}_0^{-1}\mathbf{\Phi}', \text{sign}(X_0)X_0^T\rangle_F\right| < \|(\mathbf{\Phi}_0^{-1}\mathbf{\Phi}' X_0)_{\overline{\Lambda}}\|_1 \tag{6}$$
then $\mathbf{\Phi}_0$ is a strict local minimum of (5).
2) If the reversed strict inequality holds in (6) for some $\mathbf{\Phi}' \in \mathcal{D}_{\mathbf{\Phi}_0}$, then $\mathbf{\Phi}_0$ is *not* a local minimum.

A first corollary deals with orthonormal basis learning.

Corollary 3.2 (Orthonormal bases): 1) Assume that X_0 has at least one zero entry ($\overline{\Lambda} \neq \emptyset$) and that for all nonzero skew-symmetric matrices $\mathbf{A}$
$$\left|\langle \mathbf{A}, \text{sign}(X_0)X_0^T\rangle_F\right| < \|(\mathbf{A}X_0)_{\overline{\Lambda}}\|_1, \tag{7}$$
then, for any orthonormal matrix $\mathbf{\Phi}_0$, $\mathbf{\Phi}_0$ is a strict local minimum of (5) with $\mathcal{D} = \mathcal{O}(d)$.
2) If X_0 has no zero entry, or if the reversed strict inequality holds in (7) for at least one skew-symmetric matrix, then for each orthonormal $\mathbf{\Phi}_0$, the matrix $\mathbf{\Phi}_0$ cannot be a local minimum of (5) with $\mathcal{D} = \mathcal{O}(d)$.

A second corollary is dedicated to oblique basis learning.

Corollary 3.3 (Oblique bases): 1) Assume that X_0 has at least one zero entry ($\overline{\Lambda} \neq \emptyset$), and let $\mathbf{M} := \mathbf{\Phi}_0^T\mathbf{\Phi}_0 - \mathbf{I}$ be the off-diagonal part of the Gram matrix of $\mathbf{\Phi}_0$. If for all nonzero zero-diagonal matrices $\mathbf{Z}$
$$\left|\langle \mathbf{Z}, \text{sign}(X_0)X_0^T - \mathbf{M}^T\text{diag}(\|x_k\|_1)\rangle_F\right| < \|(\mathbf{Z}X_0)_{\overline{\Lambda}}\|_1, \tag{8}$$
then $\mathbf{\Phi}_0$ is a strict local minimum of (5) with $\mathcal{D} = \mathcal{B}(d)$.
2) If the opposite strict inequality holds in (8) for at least one zero-diagonal matrix, then $\mathbf{\Phi}_0$ is not a local minimum of (5) with $\mathcal{D} = \mathcal{B}(d)$.

B. Example in $\mathbb{R}^2$

Let us consider as an example the basis learning problem in $\mathbb{R}^2$ (i.e., $d = 2$). Skew-symmetric matrices in $\mathbb{R}^2$ form a one-dimensional subspace of 2×2 matrices generated by

$$\mathbf{A} := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Up to column permutation and removal of columns made of zeroes, any $N \times 2$ matrix X_0 can be written as

$$X_0 = \begin{bmatrix} u_1 & v_1 & 0 \\ 0 & v_2 & u_2 \end{bmatrix} \tag{9}$$

where all the entries of the row vectors u, v, w, z are nonzero. Observing that

$$\mathbf{A}X_0 = \begin{bmatrix} 0 & v_2 & u_2 \\ -u_1 & -v_1 & 0 \end{bmatrix}$$

we can apply Corollary 3.2 to realize that any orthonormal basis $\mathbf{\Phi}_0$ is a strict local minimum of (5) among all orthonormal bases ($\mathcal{D} = \mathcal{O}(d)$) whenever

$$|\langle v_2, \text{sign}(v)\rangle - \langle v_1, \text{sign}(v_2)\rangle| < \|u_1\|_1 + \|u_2\|. \tag{10}$$

It is *not* a local minimum if the reversed strict inequality holds.

Remark 3.1: If $\mathbf{\Phi}_0$ is orthonormal ($\mathbf{M} = 0$) but nevertheless we want to check whether it is a local minimum of (5) among *oblique* bases ($\mathcal{D} = \mathcal{B}(d)$), then we need to rely on Corollary 3.3. The space of zero-diagonal matrices is two dimensional and generated by the two matrices $\mathbf{Z}_1$ (respectively $\mathbf{Z}_2$) with zero entries everywhere but on the entry above (respectively below) the diagonal. Since

$$\begin{aligned} \langle \mathbf{Z}_1, \text{sign}(X_0)X_0^T\rangle_F &= \langle v_2, \text{sign}(v_1)\rangle, \\ \langle \mathbf{Z}_2, \text{sign}(X_0)X_0^T\rangle_F &= \langle v_1, \text{sign}(v_2)\rangle, \\ \|(\mathbf{Z}_1X_0)_{\overline{\Lambda}}\|_1 = \|u_2\|_1, \quad &\text{and} \quad \|(\mathbf{Z}_2X_0)_{\overline{\Lambda}}\|_1 = \|u_1\|_1, \end{aligned}$$

the condition (8) cannot be satisfied for these two matrices unless the following necessary conditions hold :

$$|\langle v_2, \text{sign}(v_1)\rangle| < \|u_2\|_1 \text{ and } |\langle v_1, \text{sign}(w_2)\rangle| < \|u_1\|_1. \tag{11}$$

We will see next that (11) is also sufficient to guarantee that (8) is satisfied for *all* zero diagonal matrices when $\mathbf{M} = 0$.

IV. A sufficient recovery condition

The conditions on X_0 (and $\mathbf{\Phi}_0$) expressed so far are quite implicit and involve auxiliary matrices $\mathbf{A}$ or $\mathbf{Z}$. We now turn to a more intrinsic sufficient condition on X_0 which guarantees that any orthogonal matrix $\mathbf{\Phi}_0$ is a local optimum of the learning criterion (5) among oblique bases.

The condition is expressed based on a block decomposition of the matrix X_0 as follows (see Figure 1):

- x_k is the k-th row of X_0, and we define Λ_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_0 and keeping only the columns indexed by Λ_k (resp. $\overline{\Lambda}_k$) .

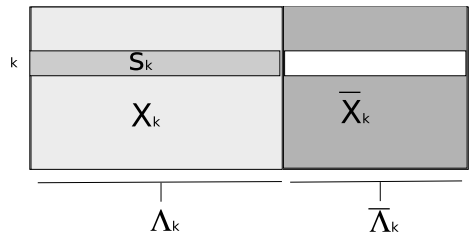

Fig. 1. Block decomposition of the matrix X_0 with respect to a given row x_k. Without loss of generality, the columns of X_0 have been permuted so that the first $|\Lambda_k|$ columns hold the nonzero entries of x_k while the last $|\overline{\Lambda}_k|$ hold its zero entries.

Theorem 4.1: Consider a $K \times N$ matrix X_0. Assume that for each k, there exists some vector d_k with

$$\bar{X}_k d_k = X_k s_k^T \text{ and } \|d_k\|_\infty < 1. \tag{12}$$

Then, for any *orthogonal* matrix $\mathbf{\Phi}_0$, the optimization problem

$$\min_{\mathbf{\Phi}, X} \|X\|_1, \text{such that } \mathbf{\Phi} X = \mathbf{\Phi}_0 X_0,$$

where $\mathbf{\Phi}$ is constrained to be a basis of unit vectors, admits a strict local minimum at $\mathbf{\Phi} = \mathbf{\Phi}_0$.

A slight modification of the proof of the Theorem shows that when condition (12) holds, $\mathbf{\Phi}_0$ is a strict local minimum even if it is "slightly" oblique, in the sense that it has low-coherence (measured through the off-diagonal part $\mathbf{M}$ of its Gram matrix). We keep for future work a precise quantification and a discussion of this robustness to obliqueness.

V. Examples

A few examples are in order to illustrate how Theorem 4.1 can be used to check the properties of a given matrix X_0.

A. A simple example: ideally sparse training data

Assume that X_0 has the following structure:

1) each column x^n is "ideally" sparse, in the sense that it has exactly one nonzero component. This means that each training sample $y^n = \mathbf{\Phi}_0 \cdot x^n$ is colinear to some dictionary vector;
2) each row x_k has at least one nonzero component, that is to say the direction of each dictionary vector is represented at least once in the training samples.

Intuitively, it is almost obvious that such properties imply that X_0 must be the (unique, up to permutation and sign change) sparsest representation of $Y = \mathbf{\Phi}_0 \cdot X_0$, but let us check how this can be turned into a proved result for orthonormal dictionaries using Theorem 4.1.

For each k, using Figure 1, we see that for each column indexed by Λ_k, since the only nonzero component is on the k-th row, we have $X_k = 0$. It follows that $X_k s_k^T = 0$ and we can take $d_k = 0$ for all k, which obviously satisfy the sufficient condition (12). It is not difficult to check that, in the simple situation considered here, $\mathbf{\Phi}_0$ will not only be local optimum of the criterion but it must indeed be its global optimum, unique up to permutation.

B. Recovering an orthonormal basis of $\mathbb{R}^2$

We now come back to the orthonormal dictionary learning problem in $\mathbb{R}^2$. For X_0 as in (9) one can readily compute

$$\bar{X}_1 = u_2, \quad X_1 = [0, v_2], \quad s_1 = [\text{sign}(u_1), \text{sign}(v_1)]$$

hence the equality $\bar{X}_1 d = X_1 s_1^T$ is equivalent to $\langle u_2, d\rangle = \langle v_2, \text{sign}(v_1)\rangle$. If $\|u_2\|_1 \leq |\langle v_2, \text{sign}(v_1)\rangle|$ then for all "admissible" d we have

$$\|u_2\|_1 \leq |\langle v_2, \text{sign}(v_1)\rangle| = |\langle u_2, d\rangle| \leq \|u_2\|_1 \cdot \|d\|_\infty$$

which implies $\|d\|_\infty \geq 1$. In the opposite case, $d = \text{sign}(u_2) \cdot \langle v_2, \text{sign}(v_1)\rangle / \|u_2\|_1$ is "admissible" and satisfies $\|d\|_\infty < 1$. A similar analysis with $\bar{X}_2$, X_2 and s_2 shows that the sufficient recovery condition expressed in Theorem 4.1 is explicitly: $\|u_2\|_1 > |\langle v_2, \text{sign}(v_1)\rangle|$ and $\|u_1\|_1 > |\langle v, \text{sign}(v_2)\rangle|$. We have therefore proved that the condition (11), which was necessary for an orthonormal basis $\mathbf{\Phi}_0$ to be a strict local minimum of (5) among oblique bases ($\mathcal{D} = \mathcal{B}(d)$), is indeed also sufficient. In this particular two-dimensional setting, Theorem 4.1 happens to be sharp. It is unlikely that the situation remains as favorable in higher dimension, since an explicit characterization of the optimum vectors d_k is not as straightforward.

C. A toy random model in $\mathbb{R}^2$

While Theorem 4.1 gives a deterministic condition for of set of coefficient X_0 to correspond to a strict local minimum of the ℓ^1 based dictionary learning criterion, it is interesting to understand how likely this condition is to be satisfied when X_0 is drawn according to a random sparse distribution.

Assume that the entries of X_0 are realizations of i.i.d random variables which can be written $x_{nk} = z_{nk} a_{nk}$ where $a_{nk} \sim \mathcal{N}(0, \sigma^2)$ and z_{nk} is a binary indicator variable with $p = P(Z = 0)$. Informally, concentration inequalities could be used to show that, for a large enough number of columns N of X_0, the number of columns with exactly one zero on the first (resp. the second) row is about $p(1-p)N$, while the number of columns with two zeros is roughly $p^2 N$ and the number of columns with two nonzero entries is about $(1-p)^2 N$. As a result, with high probability $\|z\|_1 \approx \|u\|_1 \approx Cp(1-p)N\sigma$ for some constant C, while $|\langle w, \text{sign}(v)\rangle| \approx |\langle v, \text{sign}(w)\rangle| \approx \sqrt{(1-p)^2 N}\sigma$. Therefore, condition (11) is satisfied with high probability provided that $(1-p)\sqrt{N} < Cp(1-p)N$. In other words, the probability that the "right" orthonormal basis is a strict local mininum of the ℓ^1 criterion is close to one if the number of training samples N is sufficiently large given the probability of observing a zero entry: $\sqrt{N} > C/p$.

Figure 2 displays two clouds of $N = 2500$ random points x_n obtained with the above model with $p = 0.04$ (top) and $p = 0.20$ (bottom) and $\sigma = 1$, as well as the curves $\theta \mapsto \|X_\theta\|_1$ where $X_\theta := \mathbf{\Phi}_\theta^{-1} X_0$ and

$$\mathbf{\Phi}_\theta := \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

On both curves, deep local minima of the ℓ^1 criterion are found at $\theta = 0$ (resp. $\theta = \pi/2$). They correspond to the

ideal basis up to permutation and sign change. At these local minima (indicated by circles on the curves), condition (11) is satisfied by X_θ, so one could not even decrease the ℓ^1 criterion by relaxing the orthogonality constraint to perturb the ideal bases. For small p (top curve), several spurious local minima are also found around $\theta \approx \pi/4$ and $\theta \approx 3\pi/4$, indicated with crosses. The corresponding X_θ do not satisfy condition (11): a gradient descent without the orthogonality constraint would escape from these apparent local minima. For larger p (bottom curve), only two local minima are found. They are the global optima and correspond exactly to the true underlying basis.

Are there other local minima than the global ones? We conjecture that, under this toy random model, the probability that the ideal basis is (up to standard indeterminacies) is the *only local minimum* of the ℓ^1 criterion (and therefore its global minimum) converges fast to one when the number of training samples N is sufficiently large. Indeed, for any pair of training examples x^n and $x^{n'}$ with no zero entry, the probability that x^n and $x^{n'}$ are aligned is zero, hence for $\theta \notin \{0, \pi/2\}$, X_θ will have at most one zero entry, making it impossible to satisfy the condition (11), and implying that $\max(\|u'\|_1, \|z'\|_1) \lessapprox C\sigma$ where u', z', v', w' stand for a decomposition of X_θ similar to (9). Moreover, for $\theta \notin \{0, \pi/2\}$ one roughly estimates $|\langle w', \text{sign}(v')\rangle - \langle v', \text{sign}(w')\rangle| \approx \sqrt{2N}\sigma$ by making th crude approximation that w' and v' are almost Gaussian and independent vectors. As a result, it is highly unlikely that (10) can be satisfied for $\theta \notin \{0, \pi/2\}$, which means that no rotation except the ideal ones can yield a local minimum. This is only a crude sketch, and a formal proof will require a significantly more subtle analysis of the random model, but the conjectured result would imply that, with high probability, a simple gradient descent is bound to converge to a globally optimal basis.

VI. Conclusion

We characterized the local minima of the ℓ^1 criterion for basis learning in terms of conditions on training samples which guarantee that an ideal orthonormal basis is a local optimum of the learning criterion. The main novelty compared to previous related work is that the resulting conditions do not require all training samples to be highly sparse (i.e., perfectly aligned on a few hyperplanes), therefore showing some robustness of the ℓ^1 learning criterion to non-sparse "outliers".

Yet, the assumption that a sufficient amount of the training data is perfectly aligned on a few hyperplanes is unlikely for real data. A thread for future researchis the study of the learning paradigm based on quadratic programming (ℓ^2-approximation penalized by ℓ^1-sparsity), which is known to correspond to a form of thresholding mapping to each hyperplane exactly all points that are "sufficiently close" to it.

A second challenge is to go beyond the non-overcomplete setting studied here and obtain similar results even when the dictionary $\mathbf{\Phi}_0$ is overcomplete, and we expect much from a deeper analysis of the oblique case initiated here.

Last, but not least, we want to investigate how to go beyond local minimum analysis and actually identify appropriate conditions under which the only local minima are at the global minima. In conjunction with a proof that the global minima are all equivalent up to sign and permutations, this would provide conditions under which simple gradient descent methods would be proved to converge to the global optimum, independently of the initialization.

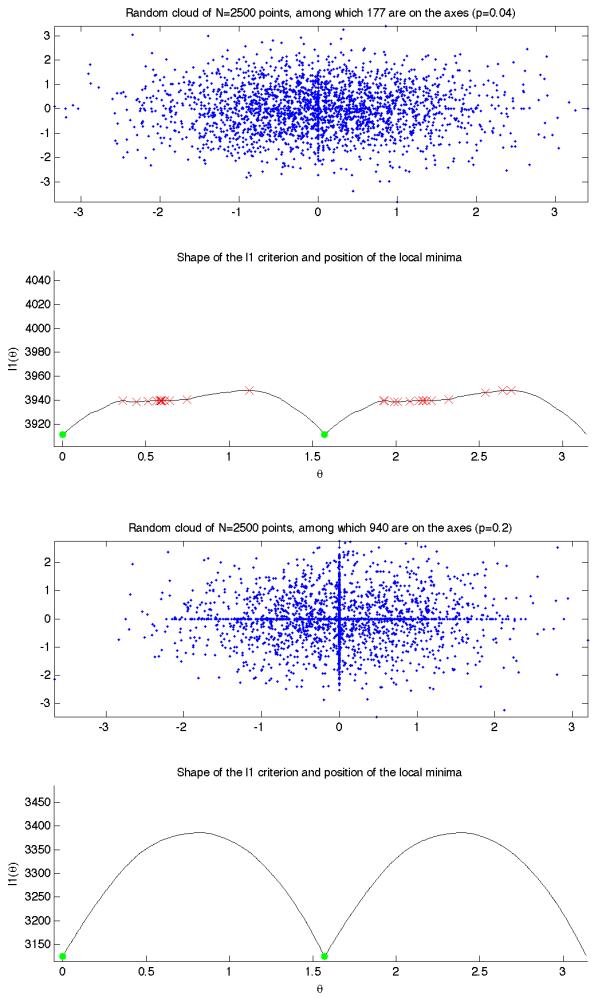

Fig. 2. Two examples of random clouds X_0 and the corresponding values of $\|\mathbf{\Phi}_\theta^{-1} X_0\|_1$ as a function of θ. Top: X_0 with few zero entries; Bottom: X_0 with "sufficiently many" zero entries (see text).

Appendix

A. Proof of Lemma 3.1

All proofs will rely extensively on the fact that

$$\langle AB, C\rangle_F = \text{Tr}(B^T A^T C) = \text{Tr}(A^T C B^T) = \langle A, CB^T\rangle_F \tag{13}$$

and similar relations. First, for small enough ϵ, $\text{sign}(X)_\Lambda = \text{sign}(X_0)_\Lambda$ and we may write

$$\|X\|_1 = \langle X, \text{sign}(X)\rangle_F = \langle X, \text{sign}(X_0)\rangle_F + \|(X - X_0)_{\overline{\Lambda}}\|_1.$$

Moreover, we have $X = \mathbf{\Phi}^{-1}(\epsilon)\mathbf{\Phi}_0 X_0 \doteq X_0 + \epsilon \mathbf{\Psi} X_0$ with

$$\mathbf{\Psi} := \frac{d(\mathbf{\Phi}^{-1}(\epsilon)\mathbf{\Phi}_0)}{d\epsilon} = -(\mathbf{\Phi}_0^{-1}\mathbf{\Phi}'\mathbf{\Phi}_0^{-1})\mathbf{\Phi}_0 = -\mathbf{\Phi}_0^{-1}\mathbf{\Phi}'$$

where the notation $a(\epsilon) \doteq b(\epsilon)$ means $\lim_{\epsilon\to 0} |(a(\epsilon) - b(\epsilon))|/|\epsilon| = 0$. Therefore, using (13)

$$\begin{aligned}\|X\|_1 - \|X_0\|_1 &= \langle X - X_0, \mathrm{sign}(X_0)\rangle_F + \|(X-X_0)_{\overline{\Lambda}}\|_1 \\ &\doteq \epsilon\langle \mathbf{\Psi}, \mathrm{sign}(X_0)X_0^T\rangle_F + |\epsilon|\cdot\|(\mathbf{\Psi}X_0)_{\overline{\Lambda}}\|_1\end{aligned}$$

Since we assume that $\overline{\Lambda} \neq \emptyset$, if for any (nonzero) admissible matrix $\mathbf{\Psi} = -\mathbf{\Phi}_0^{-1}\mathbf{\Phi}'$ we have $\left|\langle \mathbf{\Psi}, \mathrm{sign}(X_0)X_0^T\rangle_F\right| < \|(\mathbf{\Psi}X_0)_{\overline{\Lambda}}\|_1$ then $\mathbf{\Phi}_0$ is a strict local minimum of (5). Conversely, if there exists some admissible matrix $\mathbf{\Psi}$ such that $\left|\langle \mathbf{\Psi}, \mathrm{sign}(X_0)X_0^T\rangle_F\right| > \|(\mathbf{\Psi}X_0)_{\overline{\Lambda}}\|_1$ then $\mathbf{\Phi}_0$ is *not* a local minimum of (5).

B. Proof of Corollary 3.2

Since the connected component of $\mathcal{D} = \mathcal{O}(d)$ which contains the identity matrix is the set $\{\exp(\mathbf{A}), \mathbf{A}+\mathbf{A}^T = 0\}$, small orthonormal perturbations of $\mathbf{\Phi}_0$ are exactly expressed as $\mathbf{\Phi} := \mathbf{\Phi}_0\exp(\mathbf{A})$ where $\mathbf{A}$ is skew-symmetric and "small". As a result, the tangent plane to $\mathcal{O}(d)$ at $\mathbf{\Phi}_0$ is exactly the set $\mathbf{\Phi}' = \mathbf{\Phi}_0\mathbf{A}$ with $\mathbf{A}$ a skew-symmetric matrix, and most of the corollary follows directly from Lemma 3.1. We now prove the last part of the Corollary, namely that if X_0 has no zero entry then for each orthonormal basis $\mathbf{\Phi}_0$, $\mathbf{\Phi}_0$ cannot be a local minimum of (5) with $\mathcal{D} = \mathcal{O}(d)$. If X_0 has no zero entry, then for any skew-symmetric matrix $\mathbf{\Psi}$ the function $\epsilon \mapsto f_{\mathbf{\Psi}}(\epsilon) := \|\exp(\epsilon\mathbf{\Psi})\mathbf{\Phi}_0^{-1}\mathbf{\Phi}_0X_0\|_1$ is smooth in the neighborhood of $\epsilon = 0$, its value is $\langle\exp(\epsilon\mathbf{\Psi}), \mathrm{sign}(X_0)X_0^T\rangle_F$ and its derivatives at zero are $f_{\mathbf{\Psi}}^{(n)}(0) = \langle\mathbf{\Psi}^n, \mathrm{sign}(X_0)X_0^T\rangle_F$. If $\mathrm{sign}(X_0)X_0^T$ is symmetric, it is Froebenius-orthogonal to all skew-symmetric matrices, and the first derivative $f_{\mathbf{\Psi}}^{(1)}(0)$ is zero for all skew-symmetric $\mathbf{\Psi}$, hence $\mathbf{\Phi}_0$ is a stationary point of (5) with $\mathcal{D} = \mathcal{O}(d)$. However, for $\mathbf{\Psi} = (\Psi_{ij})$ the skew-symmetric matrix made of zeros everywhere except $\Psi_{12} = +1$ and $\Psi_{21} = -1$, we get that the matrix $\mathbf{\Psi}^2$ is zero everywhere except the first two diagonal terms $(\Psi^2)_{11} = (\Psi^2)_{22} = -1$, hence the second derivative is $f_{\mathbf{\Psi}}^{(2)}(0) = -\|x_1\|_1 - \|x_2\|_1 < 0$ where x_k is the k-th row of X_0, therefore this stationary point is at best a saddle point, or even worse a local maximum. Now, if $\mathrm{sign}(X_0)X_0^T$ is not symmetric, there is at least one skew-symmetric matrix $\mathbf{\Psi}$ such that $f_{\mathbf{\Psi}}^{(1)}(0) \neq 0$, hence $\mathbf{\Phi}_0$ cannot even be a stationary point of (5) with $\mathcal{D} = \mathcal{O}(d)$.

C. Proof of Corollary 3.3

Let $\mathbf{D}$ and $\mathbf{Z}$ be respectively an arbitrary diagonal and an arbitrary zero-diagonal matrix. We let the reader check that $\mathbf{\Phi}' := \mathbf{\Phi}_0(\mathbf{Z}+\mathbf{D})$ is in the tangent plane of $\mathcal{B}(d)$ at $\mathbf{\Phi}_0$ if, and only if, the diagonal of $\mathbf{\Phi}_0^T\mathbf{\Phi}' = (\mathbf{\Phi}_0^T\mathbf{\Phi}_0)\mathbf{\Phi}_0^{-1}\mathbf{\Phi}' = (\mathbf{I}+\mathbf{M})(\mathbf{Z}+\mathbf{D})$ is zero. Since $\mathbf{\Phi}_0^T\mathbf{\Phi}' = \mathbf{Z}+\mathbf{D}+\mathbf{M}\mathbf{D}+\mathbf{M}\mathbf{Z}$, this is equivalent to $\mathbf{D} = -\mathrm{diag}(\mathbf{M}\mathbf{Z})$, hence $\mathbf{\Phi}'$ is in the tangent plane of $\mathcal{B}(d)$ at $\mathbf{\Phi}_0$ if and only if $\mathbf{\Phi}_0^{-1}\mathbf{\Phi}' = \mathbf{Z} - \mathrm{diag}(\mathbf{M}\mathbf{Z})$ for some zero-diagonal matrix. We conclude using Lemma 3.1, after observing that $((\mathbf{Z}+\mathbf{D})X_0)_{\overline{\Lambda}} = (\mathbf{Z}X_0)_{\overline{\Lambda}}$, and

$$\begin{aligned}\langle\mathrm{diag}(\mathbf{M}\mathbf{Z}), \mathrm{sign}(X_0)X_0^T\rangle_F &= \langle\mathbf{M}\mathbf{Z}, \mathrm{diag}(\mathrm{sign}(X_0)X_0^T)\rangle_F \\ &= \langle\mathbf{Z}, \mathbf{M}^T\mathrm{diag}(\|x_k\|_1)\rangle_F.\end{aligned}$$

D. Proof of Theorem 4.1

The proof relies on Corollary 3.3. Denote z_k the k-th row of the zero diagonal matrix $\mathbf{Z}$. By definition, z_k is a K-dimensional row vector with a zero entry at the k-th entry, and we denote $\bar{z}_k$ the $(K-1)$-dimensional row vector obtained by removing this zero entry. We observe that since the k-th row of $\mathbf{Z}X_0$ is $z_kX_0 = \bar{z}_kX_0^k$ where X_0^k is X_0 with the k-th row removed, we have

$$\begin{aligned}\|(\mathbf{Z}X_0)_{\overline{\Lambda}}\|_1 &= \sum_k \|\bar{z}_k(X_0^k)_{\overline{\Lambda}_k}\|_1 = \sum_k \|\bar{z}_k\bar{X}_k\|_1 \\ \langle\mathbf{Z}X_0, \mathrm{sign}(X_0)\rangle_F &= \sum_k \langle\bar{z}_kX_0^k, \mathrm{sign}(x_k)\rangle.\end{aligned}$$

By matching column permutations of X_0^k and $\mathrm{sign}(x_k)$ we have

$$\begin{aligned}\langle\bar{z}_kX_0^k, \mathrm{sign}(x_k)\rangle &= \langle\bar{z}_k[X_k;\bar{X}_k], [s_k;0]\rangle = \langle\bar{z}_kX_k, s_k\rangle \\ &= \bar{z}_kX_ks_k^T.\end{aligned}$$

If for each k there exists d_k with $\|d_k\|_\infty < 1$ such that $X_ks_k^T = \bar{X}_kd_k$, we obtain that for any nonzero zero-diagonal matrix $\mathbf{Z}$

$$\begin{aligned}&|\langle\mathbf{Z}X_0, \mathrm{sign}(X_0)\rangle_F| \leq \sum_k |\bar{z}_kX_ks_k^T| = \sum_k |\bar{z}_k\bar{X}_kd_k| \\ =\ &\sum_k |\langle\bar{z}_k\bar{X}_k, d_k\rangle| \leq \sum_k \|\bar{z}_k\bar{X}_k\|_1\cdot\|d_k\|_\infty \\ \leq\ &\|(\mathbf{Z}X_0)_{\overline{\Lambda}}\|_1\cdot\max_k\|d_k\|_\infty < \|(\mathbf{Z}X_0)_{\overline{\Lambda}}\|_1.\end{aligned}$$

REFERENCES

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

[2] 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.*, vol. 15, no. 2, pp. 349–396, 2003.

[3] 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*, vol. 54, no. 11, pp. 4311–4322, nov 2006.

[4] Rémi Gribonval and Morten Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inform. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.

[5] D.L. Donoho and M. Elad, "Maximal sparsity representation via ℓ^1 minimization," *Proc. Nat. Aca. Sci.*, vol. 100, no. 5, pp. 2197–2202, Mar. 2003.

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

[7] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Information Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.

[8] J.-F. Cardoso, "On the performance of orthogonal source separation algorithms," in *Proc. EUSIPCO*, Edinburgh, Sept. 1994, pp. 776–779.

[9] P. Georgiev, F.J. Theis, and A. Cichocki, "Sparse component analysis and blind source separation of underdetermined mixtures," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 992–996, 2005.

[10] 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.*, vol. 416, pp. 48–67, July 2006.

[11] M. Zibulevsky and B.A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

[12] Mark D. Plumbley, "Dictionary learning for l1-exact sparse coding," in *Independent Component Analysis and Signal Separation*, Mike E. Davies, Christopher J. James, Samer A. Abdallah, and Mark D Plumbley, Eds. 2007, vol. 4666 of *LNCS*, pp. 406–413, Springer.