Basis Identification from Random Sparse Samples

Rémi Gribonval, Karin Schnass

***Abstract*—This article treats the problem of learning a dictionary providing sparse representations for a given signal class, via ℓ_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 basis as a local minimum of an ℓ_1-minimisation problem, it is shown that certain types of sparse random coefficient matrices will ensure local identifiability of the basis with high probability. The necessary number of training samples grows up to a logarithmic factor linearly with the signal dimension.**

Keywords: basis identification, ℓ_1-minimisation, sparse samples

I. Introduction

Sparse signals are useful. They are easy to store and to compute with and, as has become apparent through the theory of compressed sensing, they are also easy to capture. However, finding sparse representations is far from easy and by now there exists a quite comprehensive literature on algorithms and solutions strategies, for a starting point see e.g. [16], [6], [4], [17]. In any of these publications one will more likely than not find a statement starting with 'given a dictionary $\mathbf{\Phi}$ and a signal having an S-sparse approximation/representation ...', which points exactly to the remaining problem. If one has a class of signals and would like to find sparse approximations someone still has to provide the right dictionary. For many signal classes good dictionaries like time-frequency or time-scale dictionaries are known and from theoretical study of the signal class it might be possible to identify one that will fit well. However, if one runs into a new class of signals, chances that the best fit will already be known are quite slim and it can be a time consuming overkill to develop a deep theory like that of wavelets every time. An attractive alternative approach is dictionary learning, where one tries to infer the dictionary that will provide good sparse representations for the 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 [5], [12], [1], [11], 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) [3], where many identifiability results are available, which however rely on asymptotic statistical properties under independence assumptions. Georgiev, Theis and Cichocki [7] as well as Aharon, Elad and Bruckstein [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 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 basis can be learned via ℓ_1-minimisation [18], [15], and thus by a non-combinatorial algorithm. More precisely assuming that our training signals are generated from an 'ideal' basis with random sparse components we will analyse how many of these training signals are typically necessary to recover the basis with high probability. The special case when the basis is orthogonal has already been treated in [8] but the probabilistic methods used there were not strong enough to provide analogue results for general bases. In this article we take an new approach to the problem leading to stronger probabilistic estimates.

In the next sections we will shortly describe dictionary learning via ℓ_1-minimisation and state an algebraic recovery condition. In Section IV we introduce the random coefficient model and state our main theorem about the necessary number of training signals. We then sketch the main ideas of the proof going into detail as space allows. The last section is dedicated to the discussion of future work.

II. Dictionary Learning via ℓ_1-Minimisation

The first idea when trying to find a dictionary providing sparse representations of all signals from a class is to find the dictionary allowing representations with the most zero coefficients, i.e. given N training signals $y_n \in \mathbb{R}^d$, $1 \leq n \leq N$, and a candidate dictionary $\mathbf{\Phi}$ consisting of K atoms, one can measure the global sparsity as

$$\sum_{n=1}^{N} \min_{x_n} \|x_n\|_0, \text{ such that } \mathbf{\Phi} x_n = y_n, \forall n.$$

Collecting all signals y_n (considered as column vectors) in the $d \times N$ matrix Y and all coefficients x_n (considered as column vectors in $\mathbb{R}^K$) in the $K \times N$ matrix X, the fit between a dictionary $\mathbf{\Phi}$ and the training signals Y can be measured by the cost function

$$\mathcal{C}_0(\mathbf{\Phi}, Y) := \min_{X \ \mid \ \mathbf{\Phi}X = Y} \|X\|_0,$$

Rémi Gribonval is with 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 is with the Signal Processing Laboratories (LTS2), School of Engineering, EPFL, Station 11, CH - 1015 Lausanne, Switzerland, E-mail: firstname.lastname@epfl.ch. She was partly supported by NSF grant 200021-117884/1.

where $\|X\|_0 := \sum_n \|x_n\|_0$ counts the total number of nonzero entries in the $K \times N$ matrix X. Thus to get the dictionary providing the most zero coefficients out of a prescribed collection $\mathcal{D}$ of admissible dictionaries, we should consider the criterion

$$\min_{\mathbf{\Phi} \in \mathcal{D}} \mathcal{C}_0(\mathbf{\Phi}, Y). \tag{1}$$

The problem is that already finding the representation with minimal non-zero coefficients for one signal in a given dictionary is np-hard, which makes trying to solve (1) indeed a daunting task. Fortunately the problem above is not only daunting but also rather uninteresting, since it is not stable with respect to noise or suited to handle signals that are only compressible. Thus the idea of learning a dictionary via ℓ_1-minimisation is motivated on the one hand by the goal to have a criterion that is taking into account that the signals might be noisy or only compressible and on the other by the success of the Basis Pursuit principle for finding sparse representation, [6], [4]. There the ℓ_0-pseudo norm is replaced with the ℓ_1-norm, which also promotes sparsity but is convex and continuous. The same strategy can be applied to the dictionary learning problem and the ℓ_0-cost function can be replaced with the ℓ_1-cost function

$$\mathcal{C}_1(\mathbf{\Phi}, Y) := \min_{X \ | \ \mathbf{\Phi}X = Y} \|X\|_1, \tag{2}$$

where $\|X\|_1 := \sum_n \|x_n\|_1$. Several authors [18], [14], [13] have proposed to consider the corresponding minimisation problem

$$\min_{\mathbf{\Phi} \in \mathcal{D}} \mathcal{C}_1(\mathbf{\Phi}, Y). \tag{3}$$

Unlike for the sparse representation problem, where this change meant a convex relaxation, the dictionary learning problem (3) is still *not convex* and cannot be immediately addressed with generic convex programming algorithms. However, it seems better behaved than the original problem (1) because of the continuity of the criterion with respect to increasing amounts of noise, which makes it more amenable to numerical implementation.

Looking at the problem above we see that in order to solve it we still need to define $\mathcal{D}$, the set of admissible dictionaries. Several families of dictionaries can be considered such as discrete libraries of orthonormal bases, like wavelet packets or cosine packets. Here we focus on the 'non parametric' learning problem where the full $d \times K$ matrix $\mathbf{\Phi}$ has to be learned. Since the value of the criterion (3) can always be decreased by jointly replacing $\mathbf{\Phi}$ and X with $\alpha\mathbf{\Phi}$ and X/α, $0 < \alpha < 1$, a scaling constraint is necessary and a common approach is to only search for the optimum of (3) within a bounded domain $\mathcal{D}$. Here we choose

$$\mathcal{D} := \{\mathbf{\Phi}, \forall k, \|\varphi_k\|_2 = 1\}. \tag{4}$$

For a discussion of alternative constraint manifolds see for instance [10].

The special aspect of dictionary learning treated here is how a coefficient matrix X has to be structured such that for any 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 ℓ_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 ℓ_1-norm must converge to the true dictionary when started from a sufficiently close initial condition.

III. Local Identifiability Conditions for Basis Learning

To formulate the local identifiability condition, which is the starting point for our analysis, we introduce the following block decomposition of the matrix X (see Figure 1):

- x^k is the k-th row of X;
- Λ_k is the set indexing the nonzero entries of x^k and $\overline{\Lambda}_k$ the set indexing its zero entries;
- s^k is the row vector $\mathrm{sign}(x^k)_{\Lambda^k}$;
- 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 Λ_k (resp. $\overline{\Lambda}_k$) .

We also define $\mathrm{M} := \mathbf{\Phi}^\star\mathbf{\Phi} - I$. The k-th column of M will be denoted by m_k and the same column without the zero entry corresponding to the diagonal by $\bar{m}_k := (\langle\varphi_\ell, \varphi_k\rangle)_{1 \le \ell \le K, \ell \neq k}$.

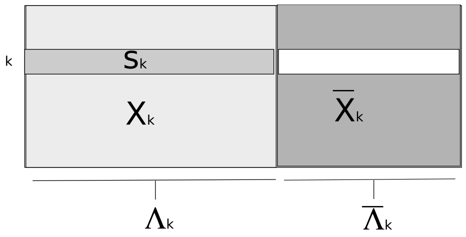

Fig. 1. Block decomposition of the matrix X_0 with respect to a given row x^k. Without loss of generality, the columns of X_0 have been permuted so that the first $|\Lambda_k|$ columns hold the nonzero entries of x^k while the last $|\overline{\Lambda}_k|$ hold its zero entries.

Theorem 3.1: Consider a $K \times N$ matrix X. If for every k there exists a vector d_k with $\max_k \|d_k\|_\infty < 1$ such that

$$\bar{X}_k d_k = X_k (s^k)^\star - \mathrm{diag}(\|x^j\|_1)_{j \neq k} \bar{m}_k. \tag{5}$$

then $(\mathbf{\Phi}, X)$ constitutes a strict local minimum of the ℓ_1-criterion.

The proof can be found in the forthcoming paper [10] or [9].

IV. Probabilistic Analysis

In this section we will derive how many training signals are typically needed to ensure that a basis constitutes a local minimum of the ℓ_1-criterion, given that the coefficients of these signals are generated by a random process.

A. The Model

We assume that the entries x_{kn} of the $K \times N$ coefficient matrix X are i.i.d. with $x_{kn} = \varepsilon_{kn} g_{kn}$, where the ε_{kn} are indicator variables taking the value one with probability p and zero with probability $1-p$, i.e. $\varepsilon \sim p\delta_1 + (1-p)\delta_0$. The variables g_{nk} follow a standard Gaussian distribution, i.e. centered with unit variance.

The important role of the indicator variables is to guarantee a strictly positive probability that the entry x_{kn} is exactly zero. The assumption that the g_{nk} are centered Gaussians with unit variance is mainly for simplicity reasons as it allows to do all proofs using only elementary probability theory. However, we believe that the same results hold for many other distributions as long as they show a certain amount of concentration, as for instance Bernoulli ± 1 with equal probability or any other subgaussian distribution.

Let us start with a geometric interpretation of the necessary recovery conditions.

B. Geometric Inspiration

We want to show that with high probability for each index k there exists a vector d_k with $\|d_k\|_\infty < 1$ such that $\bar{X}_k d_k = X_k (s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j \neq k} \bar{m}_k$. From a geometric point of view, we need to verify that the image of the unit cube $Q^{|\bar{\Lambda}_k|} = [-1,\ 1]^{|\bar{\Lambda}_k|}$ by the linear operator $\bar{X}_k$ contains the vector $u_k := X_k (s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j \neq k} \bar{m}_k$. One way to ensure this to be true is to ask that:

- the vector u_k belongs to the Euclidean ball $B_2^{K-1}(\alpha)$ of radius α, i.e., $\|u_k\|_2 \leq \alpha$;
- the image of the unit cube $Q^{|\bar{\Lambda}_k|} := [-1,1]^{|\bar{\Lambda}_k|}$ by $\bar{X}_k$ contains $B_2^{K-1}(\alpha)$.

We can see that the probability of satisfying both conditions will largely depend on the number of non zero coefficients in each row. The more zeros, the shorter the vectors s^k and x^k, thus the more likely that $\|u_k\|_2$ is small, and the higher the dimension of the unit cube, thus more chances its image covers a big ball. So we get a higher probability to recover a basis, the sparser the signals are and the more incoherent the basis is, i.e. the smaller $\|\bar{m}_k\|_2 = \|m_k\|_2$. The following theorem gives concrete estimates, derived by working out the details of the geometric sketch above.

C. Main Theorem

Theorem 4.1: Denote the event 'the original basis is not a local minimum of the ℓ_1-criterion' shortly by '$\frownie$'. If for a basis $\mathbf{\Phi}$ we have $\max_k \|m_k\|_2 < \frac{1-2p}{20}$ and the number of randomly generated training signals exceeds $N > \frac{600(K-1)}{(1-2p)^2}$ where $p < 1/2$, the probability of '$\frownie$' decays as

$$\begin{aligned}\mathbb{P}(\frownie) \leq 2K \Bigg[&\exp\left((K-1)\log(61\sqrt{\tfrac{K-1}{p}}) - \tfrac{(1-2p)pN}{13}\right) \\ &+ \exp\left(\tfrac{-(1-2p)^2 pN}{800}\right) + (K-1)\exp\left(\tfrac{-pN}{4}\right) \\ &+ \exp\left(-2p^2 N\right)\Bigg] \qquad (6)\end{aligned}$$

The crucial probabilities in the bound above are the first because of the term $\mathcal{O}(K \log K)$ and the second because of the big constant. The third is dominated by the first and for $p > 1/1603$ the last is dominated by the second one. Thus in this case we can get the cruder but more readable bound.

$$\begin{aligned}\mathbb{P}(\frownie) \leq &4K \exp\left(K \log(61\sqrt{\tfrac{K}{p}}) - \frac{(1-2p)pN}{13}\right) \\ &+ 4K \exp\left(-\frac{(1-2p)^2 pN}{800}\right).\end{aligned}$$

We can see that the general behaviour as predicted by the bound above is, that to have a good chance of recovering the dictionary we need the number of training signals N to grow faster than $K \log K$ or $d \log d$ (for a basis the number of atoms equals the signal dimension). This is only a log-factor larger than the absolute minimum of the $K+1$ training signals necessary for learning a dictionary of K elements.[1] So, as a practical example, for learning a basis for images of size 256×256 pixels, we would need around 727000 images. While this is a huge number for the more common approach of learning a basis of patches of size 100×100 we would only need around 93000 patches, which is still reasonable.

To state the theorem in a concrete form, we had to crudely bounding some intermediate probabilities. The next subsection gives a skeleton of the proof, indicating where these bounds are, so in case all parameters are precisely known, it is easy to retrace the steps and get the optimal bounds. In the course of that we will also prove the following simple but totally abstract theorem.

Theorem 4.2: If for a basis $\mathbf{\Phi}$ we have $\max_k \|m_k\|_2 < (1-p)$ then there exist constants $b > 0$ and $a, c < \infty$, depending only on p, such that for $N > c \cdot d$ we have

$$\mathbb{P}(\frownie) \leq \exp(a \cdot d \log d - b \cdot N). \qquad (7)$$

D. Skeleton of the Proof - Probability Split

To estimate the overall probability that the original basis is not a local minimum of the ℓ_1-criterion, we have a look at all aspects of the sufficient condition in (5) that could possibly go wrong and bound their probabilities individually. First, we can take the union bound over every row index k,

$$\begin{aligned}\mathbb{P}(\frownie) &\leq \mathbb{P}(\exists k,\ \text{s.t. } \nexists d_k,\ \text{s.t. } \|d_k\|_\infty < 1 \text{ and } \bar{X}_k d_k = u_k) \\ &\leq \sum_{k=1}^{K} \underbrace{\mathbb{P}(\nexists d_k,\ \text{s.t. } \|d_k\|_\infty < 1 \text{ and } \bar{X}_k d_k = u_k)}_{:=\mathbb{P}(\frownie_k)}.\end{aligned}$$

We further split by conditioning on the number of zero coefficients in each row.

$$\begin{aligned}\mathbb{P}(\frownie_k) = &\sum_{M=0}^{N} \mathbb{P}(\frownie_k \,|\, |\bar{\Lambda}_k| = M) \cdot \mathbb{P}(|\bar{\Lambda}_k| = M) \\ &\leq \max_{M_l \leq M \leq M_u} \mathbb{P}(\frownie_k \,|\, |\bar{\Lambda}_k| = M) + \mathbb{P}(|\bar{\Lambda}_k| \notin [M_l, M_u]).\end{aligned}$$

To bound the probability of the first term in the expression above, we use the geometric inspiration from Subsection IV-B.

$$\begin{aligned}&\mathbb{P}(\nexists d_k,\ \text{s.t. } \|d_k\|_\infty < 1 \text{ and } \bar{X}_k d_k = u_k \,|\, |\bar{\Lambda}_k| = M) \\ &\leq \mathbb{P}(\bar{X}_k(Q^M) \not\supseteq B_2^{K-1}(\alpha_M)) + \mathbb{P}(\|u_k\|_2 > \alpha_M \,|\, |\bar{\Lambda}_k| = M).\end{aligned}$$

[1] Given only K training signals the dictionary giving the sparsest representation is the set of training signals itself.

Retracing our steps we can thus bound the overall probability of failure as

$$\mathbb{P}(☹) \leq \sum_{k=1}^{K} \max_{M_l \leq M \leq M_u} \big[\mathbb{P}(\bar{X}_k(Q^M) \not\supseteq B_2^{K-1}(\alpha_M)) + \mathbb{P}\big(\|u_k\|_2 > \alpha_M\big)\big] + \sum_{k=1}^{K} \mathbb{P}(|\bar{\Lambda}_k| \notin [M_l, M_u]). \tag{8}$$

From (8) it becomes clear how important it is to carefully choose the parameters M_l, M_u and α_M to keep the sum of all probabilities small. However, to make this choice we first need to estimate the magnitude of the probabilities involved.

E. Estimating the Individual Probabilities

All estimates are based on concentration of measure results to bound the probability that a random variable deviates a lot from its expected value. For conciseness we will skip most proofs which can be found in [10].

The easiest estimate, the probability of the number of zero coefficients in each row being below M_l or above M_u, is a consequence of Hoeffding's inequality.

Theorem 4.3: Let $Y_1 \dots Y_N$ be independent, almost surely bounded random variables, i.e. $\mathbb{P}(Y_n \in [a_n, b_n]) = 1$. Then, for the sum $S = Y_1 + \ldots + Y_N$ and $t > 0$ we have

$$\mathbb{P}(S - \mathbb{E}(S) \geq Nt) \leq \exp\Big(-\frac{2N^2t^2}{\sum_{n=1}^N (b_n - a_n)^2}\Big).$$

Applying this for $Y_n = \varepsilon_{kn}$ with $t = (1-p)\varepsilon_\Lambda$ we get

$$\mathbb{P}(|\bar{\Lambda}_k| \leq N(1-p)(1-\varepsilon_\Lambda)) \leq \exp(-2N(1-p)^2\varepsilon_\Lambda^2).$$

The converse inequality we get in the same way for $Y_n = 1 - \varepsilon_{kn}$. Choosing $M_l = N(1-p)(1-\varepsilon_\Lambda)$ and $M_u = N(1-p)(1+\varepsilon_\Lambda)$ leads to

$$\mathbb{P}\big(|\bar{\Lambda}_k| \notin [M_l, M_u]\big) \leq 2\exp(-2N(1-p)^2\varepsilon_\Lambda^2).$$

Next we will estimate the typical size of the largest ball we can inscribe into the image of the unit cube $Q^{|\bar{\Lambda}_k|}$ by $\bar{X}_k$ when $|\bar{\Lambda}_k| = M$. We start with some geometrical observations.

Lemma 4.4: Let A be a matrix of size $d \times M$. The image of the unit cube Q^M by A contains a Euklidean ball of size α if and only if for all x with $\|x\|_2 = 1$ there exists a $v \in Q^M$, i.e. $\|v\|_\infty \leq 1$ such that $|\langle Av, x\rangle| \geq \alpha$.

Lemma 4.5: If there exists an $\varepsilon_\mathcal{N}$-net $\mathcal{N}$ for the unit sphere in $\mathbb{R}^d$ such that for all $x_i \in \mathcal{N}$ we have a $v_i \in Q^M$ such that $|\langle Av_i, x_i\rangle| \geq \alpha$ and $\|A\|_{2,\infty} \leq \beta$ then $A(Q^M) \supseteq B_2^d(\alpha - \beta\varepsilon_\mathcal{N})$.

This leads to the following probabilistic estimate.

Corollary 4.6: Choose an $\varepsilon_\mathcal{N}$-net $\mathcal{N}$ for the unit sphere in $\mathbb{R}^d$ with $|\mathcal{N}| \leq (\frac{6}{\varepsilon_\mathcal{N}})^d$. For a 'random' $d \times M$ matrix $A = (A_1 \dots A_M)$ we can bound the probability that $A(Q^M)$ covers a ball of radius $\alpha - \beta\varepsilon_\mathcal{N}$ as

$$\mathbb{P}\big(A(Q^M) \supseteq B_2^d(\alpha - \beta\varepsilon_\mathcal{N}\big) \geq 1 - \sum_{x_i \in \mathcal{N}} P\big(\|A^\star x_i\|_1 \leq \alpha\big) - \mathbb{P}\big(\sum_i \|A_i\|_2 \geq \beta\big).$$

To finally get a quantitative estimate, we need the following two concentration of measure inequalities.

Theorem 4.7: Let $A = (A_1 \dots A_M)$ be a $d \times M$ matrix, with entries as described in Subsection IV-A, $A_{ij} = \varepsilon_{ij}g_{ij}$, $i = 1 \dots d$, $j = 1 \dots M$, and $x \in \mathbb{R}^d$ be a unit vector. Then

$$a)\ \mathbb{P}\big(\|A^\star x\|_1 \leq Mp(\sqrt{\tfrac{2}{\pi}} - \varepsilon_\alpha)\big) \leq 2\exp\Big(\tfrac{-\varepsilon_\alpha^2 Mp}{2+\sqrt{2}\varepsilon_\alpha}\Big),$$

$$b)\ \mathbb{P}\big(\sum_{j=1}^M \|A_j\|_2 \geq M\sqrt{pd}(1+\varepsilon_\beta)\big) \leq 2\exp\Big(\tfrac{-\varepsilon_\beta^2 M\sqrt{p}}{2\sqrt{p}+\sqrt{2}\varepsilon_\beta}\Big).$$

The first equation tells us that we need $\alpha < \sqrt{\frac{2}{\pi}}Mp$. Indeed, since also the converse bound exists, the probability of finding a unit vector violating the condition in Lemma 4.4 rapidly approaches 1, meaning that the radius of the maximal ball cannot exceed $\sqrt{\frac{2}{\pi}}Mp$.

Choosing $\varepsilon_\alpha = \sqrt{2/\pi} - 1/3$, $\varepsilon_\beta = 1/3$ and $\varepsilon_\mathcal{N} = 10^{-1}\sqrt{p/d}$ and taking into account that $p \leq \frac{1}{2}$, we get using Corollary 4.6 and some simplifications that

$$\mathbb{P}\big(A(Q^M) \not\supseteq B_2^d(\frac{Mp}{5})\big) \leq 2\exp\left(d\log(61\sqrt{\frac{d}{p}}) - \frac{Mp}{13}\right).$$

To estimate the probability that the vector $u_k = X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k$ is not contained in the Euklidean ball of radius $\alpha = Mp/5$, we will split it into its two components and use a union bound for the second term, i.e.

$$\mathbb{P}(\|u_k|_2 > \alpha) \leq \mathbb{P}(\|X_k(s^k)^\star\|_2 > q\alpha) + \sum_{k\neq j} \mathbb{P}(\|x^j\|_1\|m_k\|_2 > (1-q)\alpha),$$

for any $q \in [0, 1]$. The optimal choice for the parameter q depends on the magnitude of $\|m_k\|_2$ measuring the coherence of the basis. So in case the basis is orthogonal we have $\|m_k\|_2 = 0$ and can set $q = 1$. For further bounds we need another two concentration of measure results.

Theorem 4.8: a) Let B be a matrix of size $d \times L$, whose entries follow the distribution described in Subsection IV-A, $B_{ij} = \varepsilon_{ij}g_{ij}$, $i = 1 \dots d$, $j = 1 \dots L$, and s be a vector of length L with entries $s_j = \pm 1$, $j = 1 \dots L$. Then for $\varepsilon_s > 0$

$$\mathbb{P}\big(\|Bs\|_2^2 \geq dLp(1+\varepsilon_s)\big) \leq 2\exp\left(\frac{-dp\varepsilon_s^2}{6+2\varepsilon_s}\right). \tag{9}$$

b) Let x be a vector of length N, whose entries follow the distribution described in Subsection IV-A, $x_i = \varepsilon_i g_i$, $i = 1 \dots N$. Then for $\varepsilon_m > 0$

$$\mathbb{P}\big(\|x\|_1 \geq L(\sqrt{\tfrac{2}{\pi}} + \varepsilon_m)\big) \leq 2\exp\left(\frac{-pN\varepsilon_m^2}{2 + \varepsilon_m/\sqrt{2}}\right). \tag{10}$$

We apply the theorem to the matrix X_k, the vector s^k and the vector x_k. Write shortly $d = K - 1$ and set $\varepsilon_s = \frac{(q\alpha)^2}{dLp} - 1$ and $\varepsilon_m = \frac{(1-q)\alpha}{pN\|m_k\|_2} - \sqrt{\frac{2}{\pi}}$ to get

$$\mathbb{P}(\|u_k\|_2 > \alpha) \leq 2\exp\Big(\tfrac{-(q\alpha)^2}{2L}c_s\Big) + 2d\exp\Big(\tfrac{-(1-q)\alpha\sqrt{2}}{\|m_k\|}c_m\Big)$$

with $c_s = \dfrac{(1 - \frac{dLp}{(q\alpha)^2})^2}{1 + 2\frac{dLp}{(q\alpha)^2}}$, $c_m = \dfrac{(1 - \sqrt{\frac{2}{\pi}}\frac{pN\|m_k\|}{(1-q)\alpha})^2}{1 + \frac{pN\|m_k\|}{(1-q)\alpha}(2\sqrt{2} - \sqrt{\frac{2}{\pi}})}$.

Let us investigate the conditions that $c_s, c_m > 0$ in more detail. Inserting the expected values for $\alpha, M, L = N - M$ shows that $c_s > 0$ will always be satisfied as soon as the number of signals N is large enough.

The condition on c_m is more interesting as in the worst case for M it is equivalent to $\|m_k\|_2 < \sqrt{\frac{\pi}{2}}\frac{(1-p)}{5}$. Looking back at the estimate of the radius of the maximal ball we see that α necessarily has to be smaller than $\sqrt{\frac{2}{\pi}}Mp$, leading to $\|m_k\|_2 < 1 - p$. This means that as soon as $\|m_k\|_2 \geq (1-p)$ the size of the vector u_k grows faster than the size of the maximal ball, and recovery can no longer be guaranteed.

However, let's assume that $\|m_k\|_2 < \frac{M}{20N}$ and choose $q = 1/\sqrt{3}$. If $M^2 > 300dL/p$ a long calculation shows that we have

$$\mathbb{P}\big(\|u_k\|_2 > \frac{Mp}{5}\big) \leq 2\exp\left(-\frac{M^2p^2}{400L}\right) + 2d\exp\left(-\frac{Np}{4}\right).$$

To get the statement of the main theorem we need to combine all the estimates and insert the worst case values for M, L with $\varepsilon_\Lambda = p/(1-p)$.

V. Discussion

We have shown that for coefficient matrices generated from a random sparse model the resulting basis coefficient pair suffices these conditions with high probability as long as the number of training signals grows like $d \log d$. These are exciting new results but since dictionary learning is a relatively young field they lead to more open questions. For the special case when the dictionary is assumed to be a basis it would be desirable to show the converse direction, i.e. if the coherence of the basis is too high and the training signals are generated by the same random sparse model, the basis coefficient pair will not be a local minimum. Ideally this breakdown coherence $\max_k \|m_k\|_2$ would be the same or close to $(1-p)$. Another helpful result would be to prove that under the random model there exists only one local minimum, which then has to be the global one, and could be found with simple descent algorithms. Numerical experiments in two dimensions support this hypothesis. Figure 2 is a plot of the ℓ_1-cost $\|\mathbf{\Phi}^{-1}Y\|_1$ for all possible two-dimensional bases, where both atoms are parametrised by their angle θ_i to the x-axis, $\theta_i \in [0, \pi]$. The $N = 500$ training signals $Y = \mathbf{\Phi}X$ were generated using the random sparse model with $p = 0.5$. As can be seen the only two local minima are at the original dictionary $\mathbf{\Phi}$ and at the dictionary corresponding to $\mathbf{\Phi}$ with permuted columns (the sign ambiguity is avoided by restricting the angles to the interval $[0, \pi]$).

Finally much harder research will have to be invested to extend the current results to the overcomplete and the noisy case. In the overcomplete case the null space has to be taken into account which prevents a straightforward generalisation from the intrinsic conditions to the explicit ones, see [10] for more information. In the noisy case already the formulation of the problem has to be changed as we cannot expect the best dictionary for the noise contaminated training data to be exactly the same as the original dictionary but only close to it.

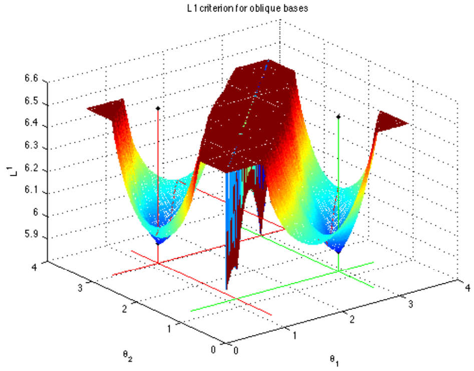

Fig. 2. ℓ_1-cost as a function of all two-dimensional bases

References

[1] M. Aharon, M. Elad, and A. M. Bruckstein. K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Trans. 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] 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.

[4] D. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Aca. Sci.*,, 100(5):2197–2202, March 2003.

[5] 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.

[6] J. J. Fuchs. Extension of the pisarenko method to sparse linear arrays. *IEEE Transactions on Signal Processing*, 45(2413-2421), October 1997.

[7] 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.

[8] R. Gribonval and K. Schnass. Dictionary identifiability from few training samples. In *Proc. EUSIPCO*, 2008.

[9] R. Gribonval and K. Schnass. Some recovery conditions for basis learning by l1-minimization. In *Proceedings ISCCSP*, March 2008.

[10] R. Gribonval and K. Schnass. Dictionary identifiability. *in preparation*, 2009.

[11] P. Jost, S. Lesage, P. Vandergheynst, and R. Gribonval. Motif: An efficient algorithm for learning translation invariant dictionaries. In *Proc. IEEE ICASSP06*, May 2006.

[12] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T. Lee, and Sejnowski T.J. Dictionary learning algorithms for sparse representation. *Neural Computations*, 15(2):349–396, 2003.

[13] B. A. Pearlmutter and R. K. Olsson. Linear program differentiation for single-channel speech separation. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2006)*, sep 2006.

[14] M. Plumbley. Geometry and homotopy for ℓ^1 sparse signal representations. In *Proc. First Workshop on Signal Processing with Sparse/Structured Representations (SPARS'05)*, pages 67–70, Rennes, France, November 2005.

[15] M.D. Plumbley. Dictionary learning for ℓ_1-exact sparse coding. In M.E. Davies, C.J. James, and S.A. Abdallah, editors, *International Conference on Independent Component Analysis and Signal Separation*, volume 4666, pages 406–413. Springer, 2007.

[16] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.

[17] J.A. Tropp. On the conditioning of random subdictionaries. *Applied Computational Harmonic Analysis*, 25(1-24), 2008.

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