Dictionary Identification - Sparse Matrix-Factorisation via ℓ_1-Minimisation

Rémi Gribonval, Karin Schnass

Dictionary Identification - Sparse Matrix-Factorisation via ℓ_1-Minimisation

Rémi Gribonval, *Senior Member, IEEE*, 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 can also be seen as factorising a $d \times N$ matrix $Y = (y_1 \dots y_N)$, $y_n \in \mathbb{R}^d$ of training signals into a $d \times K$ dictionary matrix $\mathbf{\Phi}$ and a $K \times N$ coefficient matrix $X = (x_1 \dots x_N)$, $x_n \in \mathbb{R}^K$, which is sparse. The exact question studied here is when a dictionary coefficient pair $(\mathbf{\Phi}, X)$ can be recovered as local minimum of a (nonconvex) ℓ_1-criterion with input $Y = \mathbf{\Phi} X$. First, for general dictionaries and coefficient matrices, algebraic conditions ensuring local identifiability are derived, which are then specialised to the case when the dictionary is a basis. Finally, assuming a random Bernoulli-Gaussian sparse model on the coefficient matrix, it is shown that sufficiently incoherent bases are locally identifiable with high probability. The perhaps surprising result is that the typically sufficient number of training samples N grows up to a logarithmic factor only linearly with the signal dimension, i.e. $N \approx CK \log K$, in contrast to previous approaches requiring combinatorially many samples.

Index Terms

ℓ_1-minimisation, compressed sensing, random matrices, sparse representation, dictionary learning, dictionary identification, nonconvex optimisation, independent component analysis, blind source separation, blind source localisation.

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 has highlighted their impact on inverse linear problems such as (blind) source separation and localisation as well as compressed sampling, for a starting point see e.g. [25], [12], [9], [27].

In any of these publications, one will - more likely than not - find a statement starting with 'given a dictionary $\mathbf{\Phi}$ and a signal y having an S-sparse approximation/representation $y = \mathbf{\Phi} x \dots$', which points exactly to the remaining problem: all applications of sparse representations rely on a signal dictionary $\mathbf{\Phi}$ from which sparse linear expansions

This work was supported in part by the European Commission through the SMALL project under FET-Open grant number: 225913.

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 Johann Radon Institute for Computational and Applied Mathematics (RICAM), Altenbergerstraße 54, 4040 Linz, Austria, E-mail: firstname.lastname@oeaw.ac.at

can be built that 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 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 time consuming to develop complex mathematical theory each time a new class of data, which requires a different type of dictionary, is met. An alternative approach is dictionary learning, which aims at infering the dictionary $\mathbf{\Phi}$ from a set of training data y_n. Dictionary learning, also known as *sparse coding*, has the potential of 'industrialising' sparse representation techniques for new data classes.

This article treats the theoretical dictionary learning problem, expressed as a factorisation problem which consists of identifying a $d \times K$ matrix $\mathbf{\Phi}$ from a set of N observed training vectors $y_n \in \mathbb{R}^d$, knowing that $y_n = \mathbf{\Phi} x_n$, $1 \leq n \leq N$ for some unknown collection of coefficient vectors $x_n \in \mathbb{R}^K$ with certain statistical properties.

Considering the extensive literature available for the sparse decomposition problem after the early work in [10], [14], [9], [4], [26], surprisingly little work has been dedicated to theoretical dictionary learning so far. There exist several dictionary learning algorithms (see e.g. [11], [16], [1], [15]), but only recently people have started to consider also the theoretical aspects of the problem. The origins of research into what is now called dictionary learning can be found in the field of Independent Component Analysis (ICA) [7], [5]. There, many identifiability results are available, which, however, rely on *asymptotic* statistical properties under *statistical independence* and *non-Gaussianity* assumptions.

In contrast, Georgiev, Theis and Cichocki, [13], as well as Aharon, Elad and Bruckstein, [2], described more geometric identifiability conditions on the sparse coefficients of training data in an ideal (overcomplete) dictionary. Yet, for these conditions to hold, the size N of the training set seems to be required to grow exponentially fast with the number of atoms K, and the provably good identification algorithms are combinatorial. Moreover, the algorithms and the identifiability analysis are not robust to 'outliers', i.e., training samples y_n where x_n fails to be sufficiently sparse. For applications, on the other hand, we are concerned with relatively large-dimensional data (e.g. $d = 30$, or even $d = 1000$) but limited availability of training data (N is not much larger than say $1000 \cdot d$) as well as limited computational resources.

In this article, we study the possibility of designing provably good, non-combinatorial dictionary learning algorithms that are robust to outliers and to the limited availability of training samples. Inspired by recent proofs of good properties of ℓ_1-minimisation for sparse signal decomposition with a given dictionary, we investigate the properties of ℓ_1-based dictionary learning, [29], [23]. Our ultimate goal, described in details in Section II, is to characterise properties that a set of training samples $y_n, 1 \leq n \leq N$ should satisfy to guarantee that an ideal dictionary is the only local minimum of the ℓ_1-criterion, opening up the possibility of replacing combinatorial learning algorithms with efficient numerical descent techniques. As a first step, we investigate conditions under which an ideal dictionary is a local minimum of the ℓ_1-criterion.

Main results. First, we describe the proposed setting in Section II and characterise the local minima of the ℓ_1-cost

function in Section III. We discuss the geometrical interpretation of this characterisation in Section IV. Then, using concentration of measure, we prove in Section V the perhaps surprising result that when

$$N \geq CK \log K,$$

if the samples $x_n, 1 \leq n \leq N$, are a typical draw from a Bernoulli-Gaussian random distribution (which can generate a large proportion of *outliers*), then any sufficiently incoherent basis matrix $\mathbf{\Phi}$, i.e. $K = d$, is a local minimum of the cost function and is therefore 'locally identifiable'. The constant C depends on a parameter of the Bernoulli-Gaussian distribution which drives the sparsity of the training set.

This number of training samples is surprisingly small considering that N training samples provide $N \times K \geq CK^2 \log K$ real parameters, while the basis matrix $\mathbf{\Phi}$ is essentially parameterised by $O(K^2)$ independent real parameters.

In the considered matrix identification setting, it should be noted that ℓ_1 is *not* a convex cost function. It admits *several local minima* hence local identifiability only implies that, upon good initial conditions, numerical optimisation schemes performing the ℓ_1-optimisation will recover the desired matrix $\mathbf{\Phi}$. However, empirical experiments in low dimension ($d = 2$), shown in Section VI, indicate that for typical draws of Bernoulli-Gaussian training samples x_n, the matrix $\mathbf{\Phi}$ is in fact the *only* local minimum of the criterion (up to natural indeterminacies of the problem such as column permutation). If this empirical observation could be turned into a theorem for general dimension K under the Bernoulli-Gaussian sparse model, this would imply that typically: a) ℓ_1-minimisation is a good *identification principle*; b) any decent ℓ_1-descent algorithm is a good *identification algorithm* .

II. Setting

In the vector space $\mathcal{H} = \mathbb{R}^d$ of d-dimensional signals, a dictionary is a collection of $K \geq d$ vectors φ_k, $1 \leq k \leq K$, and it is said to be *complete* if its columns span the whole space. Alternatively, a dictionary can be seen as a $d \times K$ matrix $\mathbf{\Phi}$. For a given signal $y \in \mathcal{H}$, the sparse representation problem consists of 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-Minimisation, with a Known Dictionary

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 the ℓ_0 pseudo-norm $\|x\|_0$ counts the number of nonzero entries in the vector x. However, being nonconvex and nonsmooth, (1) is hard to solve and has indeed been shown to be an NP-hard problem [8], [18]. As a result people turned to non optimal strategies like greedy algorithms or the Basis Pursuit Principle. There the problem above is replaced by its convex relaxation

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

The good news is that when y admits a sufficiently sparse representation the solution of the relaxed problem coincides with the solution of the original one, compare [14], [9], [4], [26].

B. Dictionary Learning from a Collection of Training Samples

A related problem is that of finding the dictionary that will fit a class of signals, in the sense that it will provide sparse representations for all signals of the class. The first idea is to find the dictionary allowing representations with the most zero coefficients, i.e. given N signals $y_n \in \mathcal{H}$, $1 \leq n \leq N$, and a candidate dictionary $\mathbf{\Phi}$, 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 $\mathbb{R}^d$) into a $d \times N$ matrix Y and all coefficients x_n (considered as column vectors in $\mathbb{R}^K$) into a $K \times N$ matrix X, the fit between a dictionary $\mathbf{\Phi}$ and the training signals Y can be measured by the cost function

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

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

$$\min_{\mathbf{\Phi} \in \mathcal{D}} \mathcal{C}_0(\mathbf{\Phi}|Y). \tag{P0}$$

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 (P0) 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 representations. There the ℓ_0-pseudo norm was 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

$$\mathcal{C}_1(\mathbf{\Phi}|Y) := \min_{X \mid \mathbf{\Phi} X = Y} \|X\|_1, \tag{3}$$

where $\|X\|_1 := \sum_n \|x_n\|_1$. Several authors, [29], [16], [22], [19], [23], [28], [24], have proposed to consider the corresponding minimisation problem

$$\min_{\mathbf{\Phi} \in \mathcal{D}} \mathcal{C}_1(\mathbf{\Phi}|Y). \tag{P1}$$

Unlike for the sparse representation problem, where this change meant a convex relaxation, the dictionary learning

problem (P1) is still *not convex* and cannot be immediately addressed with generic convex programming algorithms[1]. However, it seems better behaved than the original problem (P0) 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.

C. Constraints on the Dictionary

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 [6]). 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 in (P1) 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 (P1) within a bounded domain $\mathcal{D}$.

We propose to concentrate on inequality constraints of the form[2] $\max_k \|\varphi_k\|_2 \leq C$. Because of the homogeneity of the criterion with respect to scaling, we can assume without loss of generality that $C = 1$. We also let the reader check that the optimum of (P1) with the considered inequality constraints is indeed achieved when there is equality, see also [16], [28]. Hence we define the following constraint manifold

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

Let us turn now to the special aspect of dictionary learning treated in this paper.

D. Dictionary Recovery: the Identification Problem

Several algorithms have been proposed which adopt an ℓ_1 minimisation approach to learning a dictionary, [11], [16], [23], from training data. Their empirical behaviour has been explored, showing their ability to often recover with good precision the underlying dictionary.

Here we are interested in the more theoretical problem of *dictionary identification* by ℓ_1-minimisation: assuming that the data Y were generated from an 'ideal' dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$ and 'ideal' coefficients X_0 as $Y = \mathbf{\Phi}_0 X_0$, we want to determine conditions on X_0 (and to a lesser extent on $\mathbf{\Phi}_0$) such that the minimisation of (P1) recovers $\mathbf{\Phi}_0$. Our objective is therefore similar in spirit to previous work on dictionary recovery [13], [2] which studied the uniqueness of overcomplete dictionaries for sparse component analysis. The main difference here is that we

[1]The problem investigated here should not be confused with the problem of sparse channel estimation considered by Pfander, Rauhut and Tanner in [20]. There the goal is to identify a transmission channel $\mathbf{\Phi}$ by an appropriate choice of input sequence x and the observation of $y = \mathbf{\Phi}x$. The approach is to model $\mathbf{\Phi} = \sum_\ell \alpha_\ell \mathbf{\Phi}_\ell$ with sparse coefficients α in a *known* dictionary of "atomic channels", and to solve the convex problem $\min \|\alpha\|_1$ subject to $y = \sum_\ell \alpha_\ell (\mathbf{\Phi}_\ell x)$. Here, we do not have the freedom to choose x nor do we know the channel dictionary, and the problem we consider is no longer convex.

[2]Other constraints which replace the norm $\|\varphi_k\|_2$ with, e.g., a norm $\|\varphi_k\|_1$, would also be interesting to study when it is desirable to obtain sparse atoms and not only sparse coefficients.

specify in advance which optimisation criterion we want to use to recover the dictionary (ℓ_1-minimisation) and attempt to express conditions on a matrix X_0 to guarantee that this method will successfully recover a given class of dictionaries.

Permutation and sign ambiguity. The first problem we face consists of the ambiguities, which have been well known since the development of ICA. Because of the normalisation constraint we are assuming on the dictionary, the usual scaling ambiguity is avoided, but there remains a permutation and a sign ambiguity: for any permutation matrix $\mathbf{P}$ and $\mathbf{D}$ any diagonal matrix with unit diagonal entries we have $\mathbf{\Phi}X = (\mathbf{\Phi P^{-1}D^{-1}})(\mathbf{DP}X)$. Hence Problem (P1) has not just one but a whole equivalence class of minimisers, each of them corresponding to a matching column resp. row permutation and sign change of $\mathbf{\Phi}$ resp. X. Therefore, we have to relax our requirement and only ask to find conditions such that minimising (P1) recovers $\mathbf{\Phi}_0$ up to permutation and sign change. The notation $\mathbf{\Phi} \sim \mathbf{\Phi}_0$ will indicate this indeterminacy, meaning that $\mathbf{\Phi} = \mathbf{\Phi}_0\mathbf{PD}$ for some permutation matrix $\mathbf{P}$ and diagonal matrix $\mathbf{D}$ with unit diagonal entries.

Global identifiability *vs* local identifiability. Ideally, we would like to characterise coefficient matrices X_0 such that, for any $\mathbf{\Phi}_0 \in \mathcal{D}$ (or at least for a reasonable subset of $\mathcal{D}$ such as, for instance, 'incoherent' dictionaries), the *global minima* of

$$\min_{\mathbf{\Phi}\in\mathcal{D}} \mathcal{C}_1(\mathbf{\Phi}|\mathbf{\Phi}_0 X_0) \tag{5}$$

can only be found at $\mathbf{\Phi} \sim \mathbf{\Phi}_0$.

An even more ambitious objective would be to characterise coefficient matrices such that the *local minima* of (5) can only be found at $\mathbf{\Phi} \sim \mathbf{\Phi}_0$, which would guarantee that numerical optimisation algorithms cannot be trapped in spurious local minima, and would converge independently of their initialisation. This objective raises two complementary questions:

a) *Local identifiability:* Which conditions on X_0 (and $\mathbf{\Phi}_0$) guarantee that $\mathbf{\Phi}_0$ is a *local minimum* of the ℓ_1-cost function?

b) *Uniqueness:* Which conditions guarantee that, when $\mathbf{\Phi}$ is a local minimum of the ℓ_1-cost function, it must match $\mathbf{\Phi}_0$ up to column permutation and sign change?

In this paper we concentrate on the first question. The characterisation of local minima of the ℓ_1 criterion that we carry out in Section III will certainly serve to address the second question in future work.

Ideally sparse training samples *vs* non-sparse outliers In contrast to previous theoretical work on dictionary uniqueness [13], [2], we wish to determine identification conditions that do not rely on the unrealistic assumption that each training sample is ideally sparse. As a first step to deal with training data which may contain training samples $y_n = \mathbf{\Phi}_0 x_n$ with non-sparse coefficients x_n, we consider in Section V a Bernoulli-Gaussian model and show that, when the number of training samples drawn according to this model is sufficiently high, incoherent bases are associated to local minima of (5).

Figure 1 illustrates a typical cloud of $N = 1000$ points $y_n = \mathbf{\Phi}_0 x_n \in \mathbb{R}^d$, $d = 2$, where x_n was generated

according to this Bernoulli-Gaussian model with parameter $p = 0.7$ (cf Section V). Here the dictionary is a basis made of two atoms $\varphi_k^\star = (\cos\theta_k^\star, \sin\theta_k^\star)^T \in \mathbb{R}^2$, $k = 0, 1$, characterised by their angle $\theta_k^\star$, and its coherence is $\mu = |\langle\varphi_0^\star, \varphi_1^\star\rangle| = |\cos(\theta_1^\star - \theta_0^\star)| = 0.05$. One can observe that, while many training samples are perfectly aligned with the lines generated by the two atoms of the dictionary, there is also a substantial proportion of "outliers" that do not have a sparse representation in the considered dictionary.

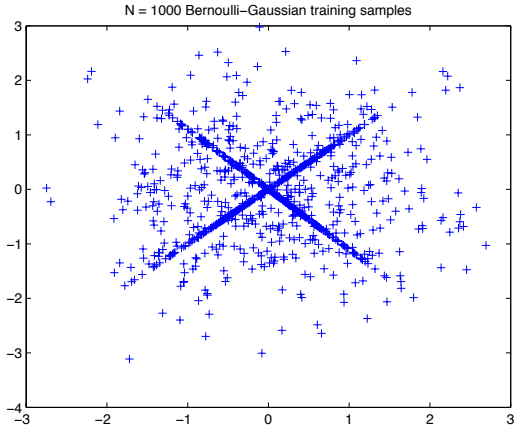

Fig. 1. A cloud of $N = 1000$ training samples in $\mathbb{R}^2$. Each point is a column y_n of the matrix $Y = \mathbf{\Phi}_0 X_0$, where X_0 was generated using the Bernoulli-Gaussian model of Section V with $p = 0.7$.

For the same point cloud shown on Figure 1, Figure 2 shows the value of the ℓ_1-cost $\mathcal{C}_1(\mathbf{\Phi}|Y)$ as a function of the angles θ_0, θ_1 which parameterise the dictionary $\mathbf{\Phi} = [\varphi_0, \varphi_1]$, where $\varphi_k = (\cos\theta_k, \sin\theta_k)^T \in \mathbb{R}^2$. One can observe that there are indeed local minima where they were expected to be located, i.e., at $(\theta_0, \theta_1) = (\theta_0^\star, \theta_1^\star)$ and $(\theta_0, \theta_1) = (\theta_1^\star, \theta_0^\star)$, which are associated to the ideal dictionary and its permuted version (the sign ambiguity is avoided by restricting the angles to the interval $[0, \pi]$). Moreover, despite the presence of many outliers in the training data, there is no other spurious local minimum. As a result, the global minima are found where they were expected, and none is missed.

For the particular case $K = d = 2$, we ran a Monte-Carlo simulation where we varied the coherence μ of the dictionary and the Bernoulli-Gaussian parameter p - which is associated to the typical sparsity of the generated training samples - repeating a hundred times the random draw of X_0. Figure 3 displays the obtained results, in terms of empirical phase transitions. For small p (associated to training data with many sparse samples), the black regions indicate that the probability of missing an expected local minimum (as well as that of finding spurious one, or an erroneous global minimum) is very low, even if the coherence of the dictionary is very high. For larger values of p, associated to training data with more non-sparse outliers in the training set, the probability of error remains very small provided that the dictionary is sufficiently incoherent. An empirical rule of thumb seems that for small p, if $\mu < 1 - p$ then the probability of learning errors is very small, provided that the number of training samples

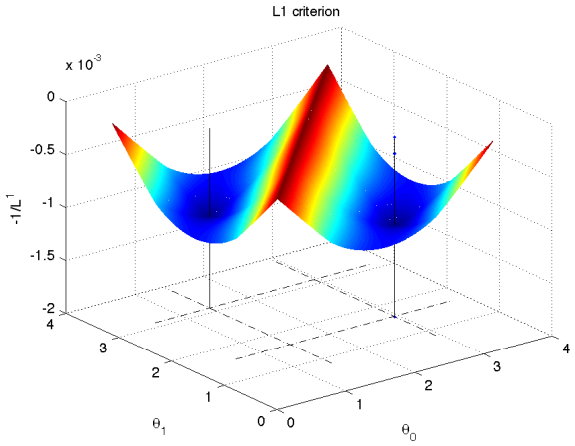

Fig. 2. The value of the cost $\mathcal{C}_1(\mathbf{\Phi}|Y)$ as a function of the angles θ_0, θ_1 which parameterise the dictionary $\mathbf{\Phi} = [\varphi_0, \varphi_1]$, $\varphi_k = (\cos\theta_k, \sin\theta_k)^T \in \mathbb{R}^2$. Because the cost function grows to infinity when $\theta_1 - \theta_0$ is close to zero, we displayed $-1/\mathcal{C}_1(\mathbf{\Phi}|Y)$ instead, which has the same minima.

is sufficiently large.

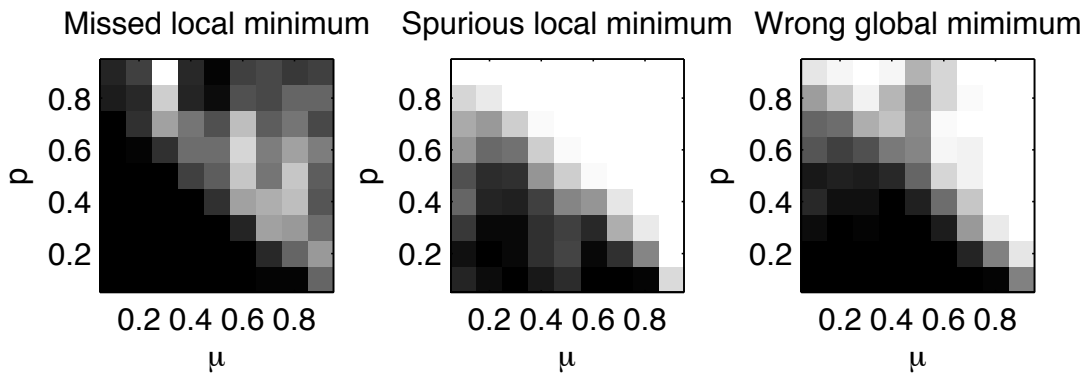

Fig. 3. Observed empirical phase transitions for dictionary identification by ℓ_1 minimisation, when $K = d = 2$ and N is large. Grey level indicates observed probability of error, from black (zero) to white (one).

Fully characterising such phase transitions for learning over-complete dictionaries is a difficult task, for several difficulties arise at once, some due to the possible overcompleteness and non-orthogonality of the dictionary, others due to the difficulty of globally characterising the optima of a globally nonconvex problem which we know admits exponentially many solutions because of the permutation and sign indeterminacies. The analytic and probabilistic machinery we set up in the next sections provides tools to significantly progress towards this ambitious goal. In particular, even though the considered Bernoulli-Gaussian model may seem simplistic (it does not account for "compressible" training samples, where x_n is not exactly sparse but only well approximated with few terms; neither

does it account for noise $y_n = \mathbf{\Phi}_0 x_n + e_n$), we believe it is a good warm up tool to understand : a) in which conditions the ℓ_1-criterion can be robust to non-sparse outliers; and b) whether dictionary identification is feasible using a limited number of samples. As we will see, fortunately, the answer to both questions is positive (but mathematically somewhat technical), under proper assumptions.

III. Local Minima

Instead of directly characterising the local mimina of the original problem (P1) we consider the related problem

$$\min_{(\mathbf{\Phi},X)|\mathbf{\Phi}\in\mathcal{D},\mathbf{\Phi}X=Y} \|X\|_1. \tag{P1'}$$

It is intimately connected to the initial problem (P1).

Remark 3.1: We let the reader check the following facts.

- When $\mathbf{\Phi}$ is a basis ($K = d$), the problem (P1') is fully equivalent to the problem (P1), in the sense that if $\mathbf{\Phi}$ is a local (resp. global) minimum of (P1), then the pair $(\mathbf{\Phi}, \mathbf{\Phi}^{-1}Y)$ is a local (resp. global) minimum of (P1'), and vice-versa.
- When $\mathbf{\Phi}$ is overcomplete ($K > d$),
  - if $\mathbf{\Phi}$ is a local (resp. global) minimum of the original problem (P1), then there is a coefficient matrix X such that the pair $(\mathbf{\Phi}, X)$ is a local (resp. global) minimum of (P1').
  - if $(\mathbf{\Phi}, X)$ is a global minimum of (P1'), then $\mathbf{\Phi}$ is a global minimum of (P1).

Just as in the representation problem (2), where the ℓ_1-cost is not a smooth function of x as soon as x has at least one zero entry, the cost in Equation (P1') is not a smooth function of $(\mathbf{\Phi}, X)$ whenever X has at least one zero entry. Therefore, one cannot fully characterise the local minima of the cost function (P1') as a subset of the zeros of a 'gradient' of the ℓ_1-cost function with respect to $(\mathbf{\Phi}, X)$, for this gradient is not even well defined in a standard sense[3].

Here, on the opposite, we want to understand the effect of the non-smooth behaviour of the cost function, and to exploit it to characterise its local minima. For that we will develop a replacement for the 'gradient' which accounts for the fact that the ℓ_1-cost function indeed admits one-sided directional derivatives everywhere. To keep the flow of the paper, we postpone most proofs and technical lemmata to the appendix.

A. Basic Notations

We denote by $\overline{\Lambda}_n$ the set indexing the zero entries of the n-th column x_n of X_0, and $\overline{\Lambda} = \{(n,k), 1 \leq n \leq N, k \in \overline{\Lambda}_n\}$ the set indexing all zero entries in X_0. The notation[4] x^k is for the k-th row of X_0, and $\overline{\Lambda}^k$ is the set indexing the columns with a zero entry in x^k.

For any $K \times N$ matrix A and index set $\Omega \subset [\![1, K]\!] \times [\![1, N]\!]$, the notation A_Ω will refer ubiquitously either to the

[3]Even the notion of Gâteaux derivatives is not applicable to this cost function, which may be a reason why a standard numerical approach [29] is to smooth it.

[4]We will generally distinguish column vectors from row vectors using subscripts *vs* superscript indices.

vector $(A_{kn})_{(k,n)\in\Omega}$ or the $K \times N$ matrix which matches A on Ω and is zero elsewhere. The cardinality of Ω is denoted $|\Omega|$.

B. Block Decomposition of the Considered Matrices

In Appendix B we provide a full characterisation of local minima (Lemma B.3) which is sharp but somewhat abstract. To make its meaning more explicit, it is useful to consider the following block decompositions of the coefficient matrix X_0 (see Figure 4):

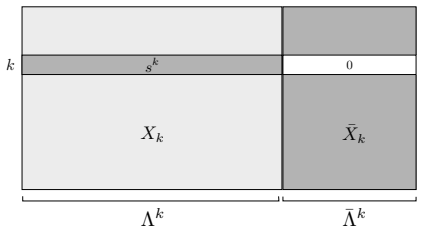

Fig. 4. Block decomposition of the matrix X_0 with respect to a given row x^k. Without loss of generality, the columns of X_0 have been permuted so that the first $|\Lambda^k|$ columns hold the nonzero entries of x^k while the last $|\overline{\Lambda}^k|$ hold its zero entries.

- x^k is the k-th row of X_0;
- 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_0 and keeping only the columns indexed by Λ^k (resp. $\overline{\Lambda}^k$) .

We also define m_k the k-th column of the off-diagonal part of the Gram matrix $\mathbf{M}_0 = \mathbf{\Phi}_0^\star \mathbf{\Phi}_0 - \mathbf{I}$ and

$$\bar{m}_k := (\langle \varphi_\ell, \varphi_k \rangle)_{1\leq \ell \leq K, \ell \neq k} \tag{6}$$

the k-th column of this matrix without the zero entry corresponding to the diagonal. Finally, we consider the vectors

$$u_k \quad := \quad X_k (s^k)^\star - \mathrm{diag}(\|x^\ell\|_1)_{1\leq \ell \leq K, \ell \neq k} \cdot \bar{m}_k. \tag{7}$$

C. A Necessary Condition, and a Sufficient Condition

Equipped with these notations, we can now state the following necessary condition.

Theorem 3.1 (Necessary condition): Consider a complete dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$, and a coefficient matrix X_0 such that $\mathbf{\Phi}_0 X_0 = Y$. Assume that X_0 is the minimum ℓ_1 norm representation of Y. With the above defined notations:

a) if $(\mathbf{\Phi}_0, X_0)$ is a local minimum of (P1'); or

b) if $\mathbf{\Phi}_0$ is a global minimum of (P1);

then we have

$$\max_k \sup_{z\neq 0} \frac{|\langle u_k, z\rangle|}{\|\bar{X}_k^\star z\|_1} \leq 1. \tag{NC}$$

As a matter of fact, condition (NC) is almost sufficient to ensure that we have a local minimum, at least in the restricted case where $\mathbf{\Phi}_0$ is a *basis*, i.e., $K = d$.

Theorem 3.2 (Sufficient condition, case of a basis, $K = d$): Consider a *basis* matrix $\mathbf{\Phi}_0$ with unit columns and a coefficient matrix X_0 such that $\mathbf{\Phi}_0 X_0 = Y$. Assume that

$$\max_k \sup_{z\neq 0} \frac{|\langle u_k, z\rangle|}{\|\bar{X}_k^\star z\|_1} < 1. \tag{SC}$$

Then $(\mathbf{\Phi}_0, X_0)$ is a strict local minimum of (P1').

It remains an open question whether this type of condition is also sufficient in the case of overcomplete dictionaries. We conjecture that the answer is positive when the constant 1 on the right hand side of (SC) is replaced by a sufficiently smaller value, under some additional assumptions relating the sparsity of X_0 and the null space of $\mathbf{\Phi}_0$. This will be the object of further studies. For the time being, we wish to obtain a more explicit understanding of the meaning of conditions (NC)-(SC), and to characterize nontrivial collections X_0 for which they are satisfied for reasonable dictionaries. In the next section we discuss the geometric interpretation of (NC)-(SC).

IV. Geometric interpretation

Using a duality argument (Lemma B.5 in the Appendix) we first observe that for any vector $v \in \mathbb{R}^{K-1}$, we have

$$\sup_{z\neq 0} \frac{|\langle v, z\rangle|}{\|\bar{X}_k^\star z\|_1} \leq 1 \tag{8}$$

if, and only if, there exists a vector d with $\|d\|_\infty \leq 1$ such that $v = \bar{X}_k d$. In other words, condition (8) holds if the vector $v \in \mathbb{R}^{K-1}$ belongs to the convex polytope obtained by projecting the high-dimensional unit hypercube[5] $Q := \{d, \|d\|_\infty \leq 1\}$ using the matrix $\bar{X}_k$.

The second observation is that the first summand in the definition of the vector u_k (cf Eq. (7)), which is the vector

$$v_k := X_k (s_k)^\star, \tag{9}$$

is a simple weighted sum of colums of X_k. Indeed, denoting X_k^+ (resp. X_k^-) the matrix made of the columns of X_k for which $x_n(k)$ is positive (resp. negative), the vector v_k is the difference between the sum of the columns of X_k^+ and the sum of those of X_k^-.

[5]We chose to denote the hypercube Q while, technically, it depends on the considered dimension $|\overline{\Lambda}^k|$ and will be denoted $Q^{|\overline{\Lambda}^k|}$ when needed.

A. Orthonormal Dictionaries

Assume for a moment that the reference dictionary $\mathbf{\Phi}_0$ is an orthonormal basis. Then, we have $\mathbf{M}_0 = 0$ and therefore $\bar{m}_k = 0$ and $u_k = v_k$ for all k. The necessary condition (NC) then simply reads: *for each k, the vector v_k must lie within the convex polytope $\bar{X}_k Q$*. This is illustrated on Figures 5 and 6, in dimension $K = 3$, so that the vector v_k as well as all the columns of X_k and $\bar{X}_k$ live in $\mathbb{R}^2$. Both figures were obtained using training data drawn according to the Bernoulli-Gaussian model described in Section V. Figure 5 corresponds to relatively sparse

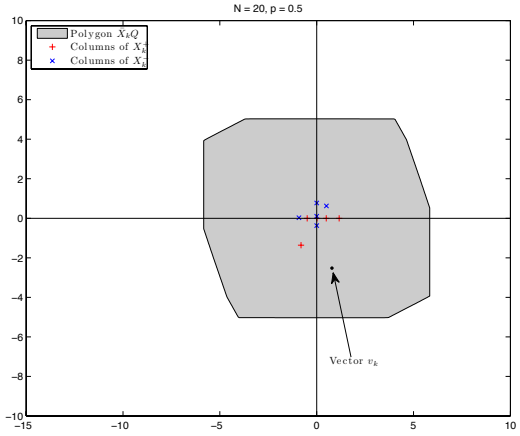

Fig. 5. Geometric depiction, when $K = 3$, of the condition (NC). The data was drawn according to the Bernoulli-Gaussian model described in Section V, with $p = 0.5$ and $N = 20$.

data (the parameter of the Bernoulli-Gaussian model is $p = 0.5$) and we can observe that despite the relatively low number of training samples ($N = 20$) the vector v_k does belong to the polygon $\bar{X}_k Q$: the necessary condition (NC) is satisfied for the considered index k, and on the same data we checked that it is also satisfied for the other two indexes. Since the vectors are indeed *strictly* inside the considered polygons, the sufficient condition (SC) is also satisfied.

On the contrary, Figure 6 corresponds to data with many non-sparse outliers ($p = 0.9$) and one can observe that despite the larger number of training samples ($N = 100$), the vector v_k does not belong to the polygon $\bar{X}_k Q$: the necessary condition (NC) is not satisfied.

B. Robustness to Dictionary Coherence

One can observe on Figure (5) that the vector v_k is well inside the convex polytope $\bar{X}_k Q$. If we choose some $1 \leq q \leq \infty$, one way to quantify this fact is to say that v_k has a small ℓ_q-norm $\|v_k\|_q$ compared to the radius of the largest ℓ_q-ball that is included in $\bar{X}_k Q$. From the definition of the vector u_k (cf Eq. (7)), it follows that if the vector

$$\mathrm{diag}(\|x^\ell\|_1)_{1 \leq \ell \leq K, \ell \neq k} \cdot \bar{m}_k$$

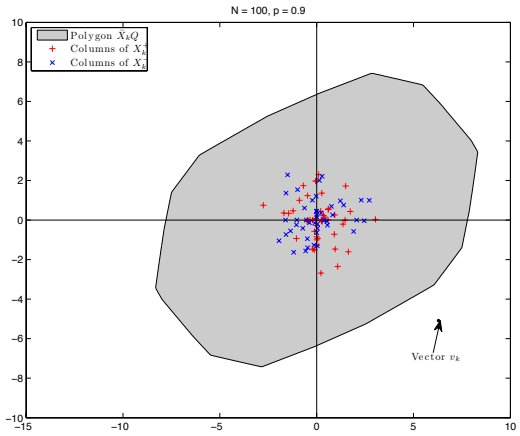

Fig. 6. Geometric depiction, when $K = 3$, of the condition (NC). The data was drawn according to the Bernoulli-Gaussian model described in Section V, with $p = 0.9$ and $N = 100$.

also has a small ℓ_q-norm (which is the case when $\mathbf{\Phi}_0$ is not necessarily orthogonal but sufficiently "incoherent"), then u_k is close to v_k, hence u_k also lies in the polytope $\bar{X}_k Q$. We then conclude that conditions (NC)-(SC) hold true. In other words, these conditions are robust to a certain level of dictionary coherence provided that:

a) each polytope $\bar{X}_k Q$ contains a "large" ℓ_q-ball;

b) each vector v_k has "small" ℓ_q-norm;

c) each row x^k of X_0 has "small" ℓ_1-norm.

Lemma B.6 in the appendix states that the radius of the largest ℓ_q-ball included in all $\bar{X}_k Q$ is given by

$$\alpha_q(X_0) := \min_k \inf_{z \neq 0} \frac{\|\bar{X}_k^\star z\|_1}{\|z\|_{q'}}, \tag{10}$$

where $1 \leq q' \leq \infty$ satisfies $1/q + 1/q' = 1$. We also define

$$\beta_q(X_0) := \max_k \|v_k\|_q, \tag{11}$$

$$\gamma(X_0) := \max_k \|x^k\|_1. \tag{12}$$

We can now state the following theorem.

Theorem 4.1: Consider $1 \leq q \leq \infty$ and a $K \times N$ matrix X_0. The conditions (NC)-(SC) are satisfied provided that the dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$ is "incoherent", in the sense that

$$\mu_q(\mathbf{\Phi}_0) := \max_k \|\bar{m}_k\|_q < \frac{\alpha_q(X_0) - \beta_q(X_0)}{\gamma(X_0)} \tag{13}$$

In particular, if $\mathbf{\Phi}_0$ is an incoherent basis ($K = d$), then the optimisation problem (P1') with $Y := \mathbf{\Phi}_0 X_0$ admits a strict local minimum at $(\mathbf{\Phi}, X) = (\mathbf{\Phi}_0, X_0)$.

Compared to Theorems 3.1 and 3.2, the above Theorem now decouples the assumptions on the coefficient matrix X_0 from those on the dictionary $\mathbf{\Phi}_0$. This will considerably simplify the analysis since we now "only" need to estimate the three quantities $\alpha_q(X_0)$, $\beta_q(X_0)$ and $\gamma(X_0)$. While the last two quantities are explicit and easy to compute for a given X_0, $\alpha_q(X_0)$ is a bit more difficult to compute for a specific X_0. In Section V, we show how to estimate its typical value when X_0 is drawn according to a Bernoulli-Gaussian model.

C. Discussion: Choice of q.

Notice that Theorem 4.1 involves a parameter $1 \leq q \leq \infty$. One may obtain coherence conditions that may be either very restrictive on the dictionary or quite weak, depending on the choice of q. As we illustrate below with a few examples, the nature of the training data can have a substantial influence on the "right" choice of q.

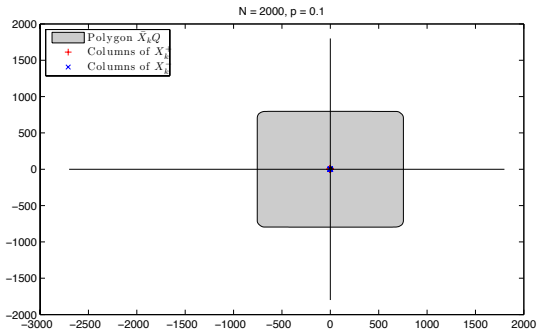

Fig. 7. Shape of the polytope $\bar{X}_k Q$, $K = 3$, $p = 0.1$ and $N = 2000$. The data was drawn according to the Bernoulli-Gaussian model described in Section V, and is highly sparse. The shape is close to a cube.

1) Highly sparse training data: For a Bernoulli-Gaussian coefficient matrix X_0 associated to small p (highly sparse data with few non-sparse outliers), as illustrated on Figure 7, the polytope $\bar{X}_k Q$ seems to be roughly shaped (when the number N of training samples is large) as a cube in $\mathbb{R}^{K-1}$. Therefore, the radius of the largest included ℓ_q-ball is almost independent of q, i.e., $\alpha_q(X_0)$ is almost constant.

Note that $\alpha_q(X_0)$, $\beta_q(X_0)$ and $\mu_q(X_0)$ are always non-increasing functions of q. If $\alpha_q(X_0)$ were actually constant, choosing $q = \infty$ in Eq.(13) would lead to the weakest possible incoherence condition which would read in terms of the well known *coherence* of the dictionary

$$\mu_\infty(\mathbf{\Phi}_0) := \max_{k \neq \ell} |\langle \varphi_k, \varphi_\ell \rangle| < \frac{\alpha_\infty(X_0) - \beta_\infty(X_0)}{\gamma(X_0)}.$$

2) Almost not sparse training data: However, the behaviour of $\alpha_q(X_0)$ as a function of X_0 heavily depends on the nature of the training data, which determines the size and shape of the polytopes $\bar{X}_k Q$. Indeed, for Bernoulli-Gaussian data associated to a large p (data with many non-sparse outliers), $\bar{X}_k Q$ seems rather shaped (when N is

large) as a Euclidean ball in $\mathbb{R}^{K-1}$, as illustrated on Figure 8. Therefore, for such data we expect that

$$\alpha_q(X_0) \approx \begin{cases} \alpha_2, & q \leq 2 \\ \alpha_2 \cdot (K-1)^{-(1/2-1/q)}, & q \geq 2. \end{cases}$$

As a result, $q = 2$ is essentially the best choice among $1 \leq q \leq 2$, but all choices $2 \leq q \leq \infty$ remain *a priori* possible, depending on the behaviour of $\beta_q(X_0)$.

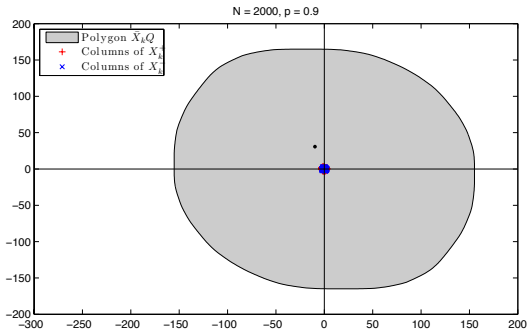

Fig. 8. Shape of the polytope $\bar{X}_k Q$, $K = 3$, $p = 0.9$ and $N = 2000$. The data was drawn according to the Bernoulli-Gaussian model described in Section V, and is almost not sparse. The shape is close to a Euclidean ball. Note the axis coordinate which indicates that the *size* of the ball is somewhat smaller than in Figure 7, for the same number of training samples but $p = 0.1$.

V. Probabilistic Analysis

In this section we will derive how many training signals are typically needed to ensure that a sufficiently incoherent basis constitutes a local minimum of the ℓ_1-criterion, given that the coefficients of these signals are drawn from a certain probability distribution.

From a Bayesian perspective, it would seem natural to consider the Laplacian distribution: minimising the ℓ_1-cost function corresponds to maximising the likelihood of $\mathbf{\Phi}$ under a Laplacian prior. However, when drawing coefficients from a Laplacian distribution, the probability of observing a zero entry is zero. Therefore, under the Laplacian prior, the minimum of the ℓ_1-cost function might be close to $\mathbf{\Phi}_0$ but cannot be *exactly* located at $\mathbf{\Phi}_0$, no matter how many training samples are drawn. For this reason, we choose to consider coefficients drawn according to a Bernoulli-Gaussian distribution, which ensures a nonzero probability $1 - p > 0$ of observing zero entries. In a sense, the setting we consider is similar to the hypotheses of the first papers on Compressed Sensing and sparse recovery [10], [14], [9], where ill-posed linear inverse problems are solved by ℓ_1-minimisation under an exact sparsity assumption. The difference here is that the model we consider also allows a certain proportion of non-sparse "outliers" in the training samples, as previously illustrated in Figure 1.

A. The Bernoulli-Gaussian Model

We assume that the entries x_{kn} of the $K \times N$ coefficient matrix X are i.i.d. with $x_{kn} = \xi_{kn} g_{kn}$, where the ξ_{kn} are indicator variables taking the value one with probability p and zero with probability $1-p$, i.e. $\xi \sim p\delta_1 + (1-p)\delta_0$.

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 made mainly for simplicity reasons as it allows us 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.

B. Asymptotic Coherence Condition

From Theorem 4.1 we know that we have to determine α, β and γ so that with high probability

a) for all k, the image $\bar{X}_k Q^{|\overline{\Lambda}^k|}$ of the unit cube by the linear map $\bar{X}_k$ contains a large ℓ_q-ball:

$$\alpha_q(X_0) \geq \alpha$$

b) for all k, the vector $X_k(s^k)^\star$ has small ℓ_q norm:

$$\beta_q(X_0) \leq \beta,$$

c) for all k, the k-th row x^k has small ℓ_1 norm

$$\gamma(X_0) \leq \gamma.$$

In Appendix C-D we derive estimates for α, β, γ and the associated probabilities using an ℓ_2-ball, i.e. $q = 2$. Our main tools are concentration of measure results to bound the probability that a random variable deviates significantly from its expected value. We obtain probability bounds exponentially small in N using

$$\begin{aligned} \alpha &\approx Np(1-p)\sqrt{\tfrac{2}{\pi}} \\ \beta &\approx \sqrt{NK}p \\ \gamma &\approx Np\sqrt{\tfrac{2}{\pi}} \end{aligned}$$

yielding, in the asymptotic regime of large N, coherence constraints of the type

$$\mu_2(\mathbf{\Phi}_0) < 1 - p.$$

C. Non-Asymptotic Result - Required Number of Training Samples

More specifically, we wish to quantify which number N of training samples guarantees, with high probability, that a basis is locally identifiable by ℓ_1 minimisation. The following theorem, whose proof can be found in Appendix E, provides an answer to this question.

Theorem 5.1: Let X be an $K \times N$ matrix drawn according to the Bernoulli-Gaussian model described in Section V-A with parameter $p < 4/5$. Assume that $N > \frac{\pi K}{2(1-p)^2}$ and that $\mathbf{\Phi}_0$ is an incoherent basis such that

$$\mu_2(\mathbf{\Phi}_0) < 1 - p - \sqrt{\frac{\pi}{2}\frac{K}{N}}. \tag{14}$$

Then $\mathbf{\Phi}_0$ is locally identifiable from $Y := \mathbf{\Phi}_0 X$ by ℓ_1-minimisation, except with probability at most

$$4K \exp\left(\frac{K}{2}\log\left(\frac{9K}{\varepsilon^2 p}\right) - Np(1-p)\frac{\varepsilon^2(1-2\varepsilon)}{2}\right), \tag{15}$$

where $0 < \varepsilon < 1/5$ is chosen as large as possible under the constraint

$$\begin{aligned}\mu_2(\mathbf{\Phi}_0) \leq & (1-p)\cdot(1-5\varepsilon) \\ & - \sqrt{\frac{\pi}{2}\left(\frac{K}{N}+\varepsilon\right)\left(1+\frac{\varepsilon}{p}\right)}.\end{aligned} \tag{16}$$

Note that we only require $p < 4/5$ to give a simple probability bound. Similar estimates also hold for $p \geq 4/5$, see proof in Appendix E.

In the theorem above, note that we need $Np(1-p)\varepsilon^2 > K$ to have failure probability smaller than one in (15). The failure probability will rapidly approach zero as soon as the number of training signals N is larger than a constant times

$$\frac{K \log K}{p(1-p)\varepsilon^2}.$$

Considering that, in order not to have a trivial sparse solution, where the columns of $\mathbf{\Phi}$ are scaled versions of the training samples y_n, we need at least $K+1$ training samples, this is not a large requirement.

Example: consider $\mathbf{\Phi}_0$ a basis of $\mathbb{R}^K$ made of $1 \leq \ell \leq K/2$ (resp. $K-\ell$) vectors from an orthonormal basis $\mathbf{\Phi}_1$ (resp. $\mathbf{\Phi}_2$) where $\mathbf{\Phi}_2$ is maximally incoherent with $\mathbf{\Phi}_1$ [10], [14]. It is easy to check that $\mu_2(\mathbf{\Phi}_0) = 1 - \ell/K < 1$, hence $\mathbf{\Phi}_0$ is, with high probability, a local minimum of the ℓ_1-criterion with $Y = \mathbf{\Phi}_0 X_0$ when X_0 is drawn according to the Bernoulli Gaussian model with $p < \ell/K < 1/2$.

VI. Discussion

We have developed necessary and sufficient algebraic conditions on a dictionary coefficient pair to constitute a local minimum of the ℓ_1-dictionary learning criterion. In case the dictionary is an incoherent basis 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 a helpful result for practical purposes 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, as shown in Figure 2 where the only two local minima are at the original dictionary $\mathbf{\Phi}_0$ and at the dictionary corresponding to $\mathbf{\Phi}_0$ with permuted columns.

It would be also 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. Again, this is empiricaly the case as shown in Figure 3. To answer this question from a theoretical perspective, it will first be necessary to investigate for which q the ℓ_q-ball most resembles the image of the unit

cube under $\bar{X}_k$. In the proof here we used $q = 2$ but there are some indications that $q = \infty$ is the more appropriate choice, which could also lead to a sharper version of the current result. Ideally we could then show that, as soon as a basis has coherence $\max_k \|m_k\|_q$ higher than $(1 - p)$, it is extremely unlikely to be a local minimum.

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 necessary and sufficient conditions of Lemma B.3 to explicit sufficient conditions as in Theorem 3.2. In the noisy case, even 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.

APPENDIX A
NOTATIONS

To state the main lemmata we need to introduce the following notation conventions.

Froebenius norm and inner product.

For any matrix, $A^\star$ denotes the transpose of A. We let $\langle A, B\rangle_F = \mathrm{trace}(A^\star B)$ denote the natural inner product between matrices, which is associated to the Froebenius norm $\|A\|_F^2 = \langle A, A\rangle_F$, and $\mathrm{sign}(A)$ is the sign operator applied componentwise to the matrix A (by convention $\mathrm{sign}(0) := 0$). All proofs will rely extensively on the fact that

$$\langle AB, C\rangle_F = \mathrm{trace}(B^\star A^\star C) = \mathrm{trace}(A^\star C B^\star) = \langle A, CB^\star\rangle_F \tag{17}$$

and similar relations such as

$$\langle \mathrm{diag}(A), B\rangle_F = \langle A, \mathrm{diag}(B)\rangle_F. \tag{18}$$

Zero-diagonal & diagonal decomposition.

We will use the following simple lemma.

Lemma A.1: Consider $\mathbf{A}$,$\mathbf{B}$ two matrices and let $\mathbf{A} = \mathbf{Z}_1 + \mathbf{\Delta}_1$, $\mathbf{B} = \mathbf{Z}_2 + \mathbf{\Delta}_2$ be their unique decomposition into a sum of a zero-diagonal and a diagonal matrix. Then

$$\mathrm{diag}(\mathbf{AB}) = \mathbf{\Delta}_1\mathbf{\Delta}_2 + \mathrm{diag}(\mathbf{Z}_1\mathbf{Z}_2).$$

Proof: The product of a zero-diagonal matrix with a diagonal matrix is zero-diagonal, hence $\mathbf{Z}_1\mathbf{\Delta}_2$ and $\mathbf{\Delta}_1\mathbf{Z}_2$ are zero-diagonal and

$$\begin{aligned}
\mathrm{diag}(\mathbf{AB}) &= \mathrm{diag}\left((\mathbf{Z}_1 + \mathbf{\Delta}_1)(\mathbf{Z}_2 + \mathbf{\Delta}_2)\right) \\
&= \mathrm{diag}\left(\mathbf{Z}_1\mathbf{Z}_2 + \mathbf{\Delta}_1\mathbf{Z}_2 + \mathbf{Z}_1\mathbf{\Delta}_2 + \mathbf{\Delta}_1\mathbf{\Delta}_2\right) \\
&= \mathrm{diag}(\mathbf{Z}_1\mathbf{Z}_2) + \mathbf{\Delta}_1\mathbf{\Delta}_2.
\end{aligned}$$

□

For any dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$, we will consider in particular the decomposition of the Gram matrix $\mathbf{\Phi}_0^\star \mathbf{\Phi}_0$ into a diagonal part and a zero-diagonal part:

$$\mathbf{\Delta}_0 := \mathrm{diag}(\mathbf{\Phi}_0^\star \mathbf{\Phi}_0) = \mathrm{diag}(\|\varphi_k\|_2^2) = \mathbf{I}, \quad (19)$$

$$\mathbf{M}_0 := \mathbf{\Phi}_0^\star \mathbf{\Phi}_0 - \mathbf{I}. \quad (20)$$

Null space

We denote by $\mathcal{N}(\mathbf{\Phi})$ the null space of the dictionary $\mathbf{\Phi}$, i.e. the linear subspace made up of all column vectors $v \in \mathbb{R}^K$ such that $\mathbf{\Phi} v = 0$. By abuse of notation, we will also denote $\mathcal{N}(\mathbf{\Phi})$ the linear space of $K \times N$ matrices $\mathbf{V}$ such that $\mathbf{\Phi}\mathbf{V} = 0$.

ε-cover

A finite ε-cover of the unit ℓ^q-sphere in $\mathbb{R}^n$ is a finite set $\mathcal{X}$ of points with unit ℓ^q-norm such that for all points in the sphere, i.e. $\|x\|_q = 1$, we have

$$\min_{x_i \in \mathcal{X}} \|x - x_i\|_q < \varepsilon.$$

From Lemma 4.10 in [21] we know that for $\varepsilon \in (0,1)$ there always exists an ε-cover $\mathcal{X}$ with cardinality $|\mathcal{X}| < (3/\varepsilon)^n$.

APPENDIX B
TANGENT SPACES AND LOCAL MINIMA

To characterise whether $(\mathbf{\Phi}_0, X_0)$ is a local minimum of (P1'), we will use the notion of the tangent space $T_{(\mathbf{\Phi}_0, X_0)}\mathcal{M}(Y)$ to the constraint manifold

$$\mathcal{M}(Y) := \{(\mathbf{\Phi}, X), \mathbf{\Phi} \in \mathcal{D}, \mathbf{\Phi} X = Y\} \quad (21)$$

at the point $(\mathbf{\Phi}_0, X_0)$. We characterise this tangent space before providing the characterisation of the local minima.

A. The Tangent Space $T_{(\mathbf{\Phi}_0, X_0)}\mathcal{M}(Y)$

The tangent space $T_{(\mathbf{\Phi}_0, X_0)}\mathcal{M}(Y)$ to the constraint manifold $\mathcal{M}(Y)$ at the point $(\mathbf{\Phi}_0, X_0)$ is the collection of the derivatives $(\mathbf{\Phi}', X') := (\mathbf{\Phi}'(0), X'(0))$ of all smooth functions $\epsilon \mapsto (\mathbf{\Phi}(\epsilon), X(\epsilon))$ which satisfy $\forall \epsilon, (\mathbf{\Phi}(\epsilon), X(\epsilon)) \in \mathcal{M}(Y)$ and $(\mathbf{\Phi}(0), X(0)) = (\mathbf{\Phi}_0, X_0)$.

Below we characterise the tangent spaces $T_{\mathbf{\Phi}_0}\mathcal{D}$ and $T_{(\mathbf{\Phi}_0, X_0)}\mathcal{M}(Y)$. The characterisations use the decomposition $\mathbf{\Phi}_0^\star \mathbf{\Phi}_0 = \mathbf{I} + \mathbf{M}_0$ introduced in Equations (19)-(20), through the notion of *admissible* matrices: a square $K \times K$ matrix C is said to be admissible if $\mathbf{\Phi}' := \mathbf{\Phi}_0 \cdot C \in T_{\mathbf{\Phi}_0}\mathcal{D}$.

Lemma B.1: Let $\mathbf{\Phi}_0 \in \mathcal{D}$ be a complete dictionary.

- Any matrix $\mathbf{\Phi}' \in T_{\mathbf{\Phi}_0}\mathcal{D}$ can be written as $\mathbf{\Phi}' = \mathbf{\Phi}_0 \cdot C$ for some admissible C.
- The matrix C is admissible if, and only if there exists a zero-diagonal matrix $\mathbf{Z}$ such that

$$C = \mathbf{Z} - \operatorname{diag}(\mathbf{M}_0\mathbf{Z}) \tag{22}$$

Proof: The first claim is a trivial consequence of the completeness of $\mathbf{\Phi}_0$, which shows that any matrix can be written as $\mathbf{\Phi}_0 \cdot C$, and the definition of an admissible matrix.
The constraint in (4) can be rewritten as $\operatorname{diag}(\mathbf{\Phi}^\star\mathbf{\Phi}) = \mathbf{I}$. Taking the derivative, it follows that $\mathbf{\Phi}' \in T_{\mathbf{\Phi}_0}\mathcal{D}$ if, and only if, $\operatorname{diag}(\mathbf{\Phi}_0^\star\mathbf{\Phi}') = 0$. Writing $\mathbf{\Phi}' = \mathbf{\Phi}_0 \cdot C$ and decomposing $C = \mathbf{Z} + \mathbf{\Delta}$ into a zero-diagonal and a diagonal matrix, we obtain from Lemma A.1

$$\begin{aligned} \operatorname{diag}(\mathbf{\Phi}_0^\star\mathbf{\Phi}') &= \operatorname{diag}(\mathbf{\Phi}_0^\star\mathbf{\Phi}_0 \cdot C) = \operatorname{diag}\left((\mathbf{M}_0 + \mathbf{I})(\mathbf{Z} + \mathbf{\Delta})\right) \\ &= \mathbf{\Delta} + \operatorname{diag}(\mathbf{M}_0\mathbf{Z}). \end{aligned}$$

Hence $\mathbf{\Phi}_0 \cdot C \in T_{\mathbf{\Phi}_0}\mathcal{D}$ if, and only if, $\mathbf{\Delta} = -\operatorname{diag}(\mathbf{M}_0\mathbf{Z})$, i.e. if $C = \mathbf{Z} - \operatorname{diag}(\mathbf{M}_0\mathbf{Z})$.

□

Lemma B.2: The pair $(\mathbf{\Phi}', X')$ is in the tangent space $T_{(\mathbf{\Phi}_0, X_0)}\mathcal{M}(Y)$ if, and only if, there exists an arbitrary admissible matrix C and an arbitrary element $\mathbf{V}$ of $\mathcal{N}(\mathbf{\Phi}_0)$ such that

$$\mathbf{\Phi}' = \mathbf{\Phi}_0 \cdot C \tag{23}$$

$$X' = -CX_0 + \mathbf{V}. \tag{24}$$

Proof: Given the nature of the constraint manifold $\mathcal{M}(Y)$, its tangent space at $(\mathbf{\Phi}_0, X_0)$ is made up of all the pairs $(\mathbf{\Phi}', X')$ such that $\mathbf{\Phi}' \in T_{\mathbf{\Phi}_0}\mathcal{D}$ and $\mathbf{\Phi}'X_0 + \mathbf{\Phi}_0X' = 0$, meaning $\mathbf{\Phi}' = \mathbf{\Phi}_0 \cdot C$ with some admissible C, and $\mathbf{\Phi}_0(CX_0 + X') = 0$. The latter is equivalent to $CX_0 + X' \in \mathcal{N}(\mathbf{\Phi}_0)$. □

B. Characterisation of Local Minima

Lemma B.3: Consider a complete dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$, and a coefficient matrix X_0 such that $\mathbf{\Phi}_0X_0 = Y$. Define the $K \times K$ matrix

$$\mathbf{U} := \operatorname{sign}(X_0)X_0^\star - \mathbf{M}_0^\star \operatorname{diag}(\|x^k\|_1). \tag{25}$$

a) If for every zero-diagonal $\mathbf{Z}$ and $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$ such that $\mathbf{Z}X_0 + \mathbf{V} \neq 0$ we have

$$|\langle \mathbf{Z}, \mathbf{U}\rangle_F + \langle \mathbf{V}, \operatorname{sign}(X_0)\rangle_F| < \|(\mathbf{Z}X_0 + \mathbf{V})_{\overline{\Lambda}}\|_1, \tag{26}$$

then $(\mathbf{\Phi}_0, X_0)$ is a strict local minimum of (P1').

b) If the reversed strict inequality holds in (26) for some zero-diagonal $\mathbf{Z}$ and some $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$ such that $\mathbf{Z}X_0 + \mathbf{V} \neq 0$, then $(\mathbf{\Phi}_0, X_0)$ is *not* a local minimum of (P1').

Proof: Denote $a(\epsilon) \doteq b(\epsilon)$ when $\lim_{\epsilon\to 0} \|a(\epsilon)-b(\epsilon)\|/|\epsilon| = 0$. Consider any smooth function $\epsilon \mapsto (\mathbf{\Phi}(\epsilon), X(\epsilon)) \in \mathcal{M}(Y)$. By definition we have $X(\epsilon) \doteq X_0 + \epsilon X'$, and for small ϵ, the sign of $X(\epsilon)$ matches that of $X_0 = X(0)$ on the support Λ of X_0, hence we may write

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

As a result, the one-sided derivatives of the ℓ_1-criterion in the tangent direction $(\mathbf{\Phi}', X')$ are

$$\begin{aligned}
\nabla^+_{\mathbf{\Phi}', X'}\|X\|_1 &:= \lim_{\epsilon\to 0, \epsilon>0} \frac{\|X(\epsilon)\|_1 - \|X_0\|_1}{\epsilon} \\
&= +\|(X')_{\overline{\Lambda}}\|_1 + \langle X', \mathrm{sign}(X_0)\rangle_F \\
\nabla^-_{\mathbf{\Phi}', X'}\|X\|_1 &:= \lim_{\epsilon\to 0, \epsilon<0} \frac{\|X(\epsilon)\|_1 - \|X_0\|_1}{\epsilon} \\
&= -\|(X')_{\overline{\Lambda}}\|_1 + \langle X', \mathrm{sign}(X_0)\rangle_F,
\end{aligned}$$

and the ℓ_1-criterion admits a local minimum at $(\mathbf{\Phi}_0, X_0)$ if for all $(\mathbf{\Phi}', X')$ in the tangent space $T_{(\mathbf{\Phi}_0, X_0)}\mathcal{M}(Y)$ with $X' \neq 0$ we have

$$|\langle X', \mathrm{sign}(X_0)\rangle_F| < \|(X')_{\overline{\Lambda}}\|_1.$$

Vice-versa, the ℓ_1-criterion does not admit a local minimum at $(\mathbf{\Phi}_0, X_0)$ if there exists some $(\mathbf{\Phi}', X')$ in the tangent space $T_{(\mathbf{\Phi}_0, X_0)}\mathcal{M}(Y)$ yielding the reversed strict inequality.

Using Lemma B.2 we get that the ℓ_1-criterion admits a local minimum at $(\mathbf{\Phi}_0, X_0)$ if for all admissible C and all $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$ such that $\mathbf{V} \neq CX_0$ we have

$$|\langle CX_0 + \mathbf{V}, \mathrm{sign}(X_0)\rangle_F| < \|(CX_0 + \mathbf{V})_{\overline{\Lambda}}\|_1. \tag{27}$$

The rest of the proof consists in rewriting (27) using Lemma B.1 and the properties (17) and (18).

First, using (17), the inequality in (27) is equivalent to

$$|\langle C, \mathrm{sign}(X_0)X_0^\star\rangle_F + \langle \mathbf{V}, \mathrm{sign}(X_0)\rangle_F| < \|(CX_0 + \mathbf{V})_{\overline{\Lambda}}\|_1.$$

Second, by Lemma B.1, the admissible matrices are exactly the matrices $C = \mathbf{Z} - \mathrm{diag}(\mathbf{M}_0\mathbf{Z})$, with $\mathbf{Z}$ an arbitrary zero-diagonal matrix. Since $(\mathbf{\Delta}\cdot X_0)_{\overline{\Lambda}} = 0$ for any diagonal matrix $\mathbf{\Delta}$, we get $(CX_0)_{\overline{\Lambda}} = (\mathbf{Z}X_0)_{\overline{\Lambda}}$ for any admissible matrix. The inequality is therefore equivalent to

$$\begin{aligned}
&|\langle \mathbf{Z} - \mathrm{diag}(\mathbf{M}_0\mathbf{Z}), \mathrm{sign}(X_0)X_0^\star\rangle_F + \langle \mathbf{V}, \mathrm{sign}(X_0)\rangle_F| \\
&< \|(\mathbf{Z}X_0 + \mathbf{V})_{\overline{\Lambda}}\|_1,
\end{aligned} \tag{28}$$

with arbitrary zero-diagonal $\mathbf{Z}$ and $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$.

Third, since $\mathrm{diag}(\mathrm{sign}(X_0)X_0^\star) = \mathrm{diag}(\|x^k\|_1)$, we observe using (17) and (18) that

$$\begin{aligned}\langle \mathrm{diag}(\mathbf{M}_0\mathbf{Z}), \mathrm{sign}(X_0)X_0^\star\rangle_F &= \langle \mathbf{M}_0\mathbf{Z}, \mathrm{diag}(\mathrm{sign}(X_0)X_0^\star)\rangle_F \\ &= \langle \mathbf{Z}, \mathbf{M}_0^\star \,\mathrm{diag}(\|x^k\|_1)\rangle_F.\end{aligned} \tag{29}$$

Hence the inequality in (28) is equivalent to

$$\begin{aligned}\big|\langle \mathbf{Z}, \mathrm{sign}(X_0)X_0^\star - \mathbf{M}_0^\star \,\mathrm{diag}(\|x^k\|_1)\rangle_F + \langle \mathbf{V}, \mathrm{sign}(X_0)\rangle_F\big| \\ < \|(\mathbf{Z}X_0 + \mathbf{V})_{\overline{\Lambda}}\|_1.\end{aligned}$$

□

C. Proof of Theorems 3.1 and 3.2

Lemma B.4: Using the notations of Section III we have

$$\sup_{\mathbf{Z}\neq 0} \frac{\big|\langle \mathbf{Z}, \mathbf{U}\rangle\big|}{\big\|(\mathbf{Z}X_0)_{\overline{\Lambda}}\big\|_1} = \max_k \sup_{z\in\mathbb{R}^{K-1}\setminus\{0\}}, \frac{|\langle u_k, z\rangle|}{\|\bar{X}_k^\star z\|_1}. \tag{30}$$

Proof: Denote z^k the k-th row of the zero diagonal matrix $\mathbf{Z}$: it is a row vector in $\mathbb{R}^K$ with a zero entry at the k-th coordinate, and we denote $\bar{z}^k$ the row vector in $\mathbb{R}^{K-1}$ obtained by removing this zero entry. Observe that 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. As a consequence the denominator in Eq. (30) is decomposed into the sum

$$\begin{aligned}\|(\mathbf{Z}X_0)_{\overline{\Lambda}}\|_1 &= \sum_k \|(z^kX_0)_{\overline{\Lambda}^k}\|_1 = \sum_k \|(\bar{z}^kX_0^k)_{\overline{\Lambda}^k}\|_1 \\ &= \sum_k \|\bar{z}^k(X_0^k)_{\overline{\Lambda}^k}\|_1 = \sum_k \|\bar{z}^k\bar{X}_k\|_1.\end{aligned} \tag{31}$$

Now we decompose the numerator into a similar sum. First, we observe that

$$\begin{aligned}\langle \mathbf{Z}, \mathbf{M}_0^\star \,\mathrm{diag}(\|x^k\|_1)\rangle_F &= \sum_k \langle z^k, m_k^\star \,\mathbf{diag}(\|x^\ell\|_1)_{1\le\ell\le K}\rangle \\ &= \sum_k \cdot\langle \bar{z}^k, \bar{m}_k^\star \,\mathbf{diag}(\|x^\ell\|_1)_{1\le\ell\le K,\ell\neq k}\rangle,\end{aligned}$$

$$\begin{aligned}\langle \mathbf{Z}, \mathrm{sign}(X_0)X_0^\star\rangle_F &= \langle \mathbf{Z}X_0, \mathrm{sign}(X_0)\rangle_F \\ &= \sum_k \langle z^kX_0, \mathrm{sign}(x^k)\rangle \\ &= \sum_k \langle \bar{z}^kX_0^k, \mathrm{sign}(x^k)\rangle.\end{aligned}$$

Then, by matching column permutations of X_0^k and $\mathrm{sign}(x^k)$ we get

$$\begin{aligned}\langle \bar{z}^kX_0^k, \mathrm{sign}(x^k)\rangle = \langle \bar{z}^k[X_k; \bar{X}_k], [s^k; 0]\rangle &= \langle \bar{z}^kX_k, s^k\rangle \\ &= \langle \bar{z}^k, s^kX_k^\star\rangle,\end{aligned}$$

and conclude that the numerator is

$$\big|\langle \mathbf{Z}, \mathbf{U}\rangle\big| = \big|\sum_k \langle \bar{z}^k, u_k^\star\rangle\big|. \tag{32}$$

The conclusion is then straightforward. □

Proof: [Proof of Theorem 3.1] Using Lemma B.3 and Remark 3.1 we know that if $\mathbf{\Phi}_0$ is a local minimum of (P1') or a global minimum of 5, then for any zero-diagonal matrix $\mathbf{Z}$ and any $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$ such that $\mathbf{Z}X_0 + \mathbf{V} \neq 0$ we have $\big|\langle \mathbf{Z}, \mathbf{U}\rangle + \langle \mathbf{V}, \mathrm{sign}(X_0)\rangle\big| \leq \big\|(\mathbf{Z}X_0 + \mathbf{V})_{\overline{\Lambda}}\big\|_1$. In particular, for any $\mathbf{Z} \neq 0$ and $\mathbf{V} = 0$, we have $\big|\langle \mathbf{Z}, \mathbf{U}\rangle\big| \leq \big\|(\mathbf{Z}X_0)_{\overline{\Lambda}}\big\|_1$. We conclude using Lemma B.4. □

Proof: [Proof of Theorem 3.2] When $\mathbf{\Phi}_0$ is a basis, the null space is $\mathcal{N}(\mathbf{\Phi}_0) = \{0\}$, and Condition (26) is satisfied for all nonzero zero-diagonal matrices $\mathbf{Z}$ and $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$ such that $\mathbf{Z}X_0 + \mathbf{V} \neq 0$ if, and only if, for all nonzero zero-diagonal matrix $\mathbf{Z}$ we have $|\langle \mathbf{Z}, \mathbf{U}\rangle_F| < \|(\mathbf{Z}X_0)_{\overline{\Lambda}}\|_1$. Again, we conclude thanks to Lemma B.4. □

D. Duality Analysis

The next lemma exploits duality to understand the geometric meaning of conditions in (NC)-(SC). The following Lemma is used with the matrix $A = \bar{X}_k$ to obtain the equivalent characterization of (8) used in Section IV.

Lemma B.5: Let A be an $n \times M$ matrix with rank n. For any vector v define

$$\|v\|_A \quad := \quad \sup_{z \neq 0} \frac{\langle v, z\rangle}{\|A^\star z\|_1}. \tag{33}$$

We have the equivalent characterisation

$$\|v\|_A = \min \|d\|_\infty, \text{ under the constraint } Ad = v. \tag{34}$$

Proof: We will just prove that

$$\|v\|_A \leq \min \|d\|_\infty, \text{ under the constraint } Ad = v.$$

The reversed inequality is more technical but only requires casting both norm characterisations (33)-(34) to a pair of linear programs in primal and dual form, and using the strong duality theorem to show that both programs, which are bounded and feasible, have the same value of the optimum. To check the easy inequality, take any d such that $Ad = u$. Since A has rank n, we have $\|A^\star z\|_1 \neq 0$ whenever $z \neq 0$. Thus, for any $z \neq 0$ we have $\langle v, z\rangle = \langle Ad, z\rangle = \langle d, A^\star z\rangle \leq \|d\|_\infty \cdot \|A^\star z\|_1$, hence $\|v\|_A \leq \|d\|_\infty$. □

Lemma B.6: Consider A an $n \times M$ matrix and $1 \leq q, q' \leq \infty$ with $1/q + 1/q' = 1$. The radius of the largest ℓ_q ball included in AQ^M is

$$R_q(A) := \inf_{z \neq 0} \frac{\|A^\star z\|_1}{\|z\|_{q'}}. \tag{35}$$

Proof: If A is not of rank n we let the reader check that $R_q(A) = 0$ is also the radius of the largest ball included in AQ^M. Otherwise, from Lemma B.5 we know that $v \in AQ^M$ if and only if $\sup_{z \neq 0} \frac{|\langle v, z\rangle|}{\|A^\star z\|_1} \leq 1$. The inclusion

of an ℓ_q ball of radius α in AQ^N is therefore equivalent to

$$\sup_{\|v\|_q \leq \alpha} \sup_{z \neq 0} \frac{|\langle v, z \rangle|}{\|A^\star z\|_1} \leq 1.$$

Conclude by rewriting the left hand side:

$$\alpha \sup_{\|v'\|_q \leq 1, z \neq 0} \frac{|\langle v', z \rangle|}{\|A^\star z\|_1} = \alpha \sup_{z \neq 0} \frac{\|z\|_{q'}}{\|A^\star z\|_1}.$$

□

E. Proof of Theorem 4.1

Using the definition of u_k, $\beta_q(X_0)$, $\gamma(X_0)$ and $\mu_q(\mathbf{\Phi}_0)$ (cf Eqs. (7), (11), (12) and (13)) and the assumption on $\mu_q(\mathbf{\Phi}_0)$ (Eq. (13)) we have for all k

$$\begin{aligned} \|u_k\|_q &\leq \|v_k\|_q + \gamma(X_0) \cdot \mu_q(\mathbf{\Phi}_0) \\ &\leq \beta_q(X_0) + \gamma(X_0) \cdot \mu_q(\mathbf{\Phi}_0) < \alpha_q(X_0). \end{aligned}$$

Hence, by definition of $\alpha_q(X_0)$ the vector u_k belongs to $\bar{X}_k Q$ for all k, and we conclude using Lemma B.5 that the condition (SC) is satisfied. In particular, if $\mathbf{\Phi}_0$ is a basis then we conclude using Theorem 3.2 that $(\mathbf{\Phi}_0, X_0)$ is a local minimum of (P1').

APPENDIX C
PROBABILITY ESTIMATES

A. Typical Size of $\|x^k\|_1$

The typical size of $\gamma(X_0) = \max_k \|x^k\|_1$ can be directly derived from the following concentration of measure result.

Theorem C.1: Let x be a vector of length N, whose entries follow the distribution described in Subsection V-A, $x_n = \xi_n g_n$, $n = 1 \dots N$. Then for any $\varepsilon > 0$

$$\mathbb{P}\big(\|x\|_1 > Np(\sqrt{\tfrac{2}{\pi}} + \varepsilon)\big) \leq 2 \cdot \exp\left(-\frac{Np \cdot \varepsilon^2}{2 + \sqrt{2} \cdot \varepsilon}\right).$$

It follows immediately, using a union bound, that with

$$\gamma := Np(\sqrt{\tfrac{2}{\pi}} + \varepsilon), \tag{36}$$

we have

$$\mathbb{P}\big(\gamma(X_0) > \gamma\big) \leq 2K \cdot \exp\left(-\frac{pN\varepsilon^2}{2 + \sqrt{2} \cdot \varepsilon}\right). \tag{37}$$

B. General Approach to Estimating α and β

Now we will estimate the probability that for one index k either a) or b) fails. Denote Ω_k the event

$$\Omega_k := \{R_q(\bar{X}_k) < \alpha\} \cup \{\|X_k(s^k)^\star\|_q > \beta\},$$

i.e. either a) or b) fails for row k. Then $\Omega = \cup_k \Omega_k$ is the undesired event $\{\alpha_q(X_0) < \alpha\} \cup \{\beta_q(X_0) > \beta\}$. Using a union bound over the row indices k and conditioning on the size of the set of zero entries $|\overline{\Lambda}^k|$ we get,

$$\begin{aligned}\mathbb{P}\big(\Omega\big) &\leq \sum_{k,M} \mathbb{P}\big(\Omega_k \;\big|\; |\overline{\Lambda}^k| = M\big) \cdot \mathbb{P}\big(|\overline{\Lambda}^k| = M\big) \\ &\leq K \cdot \max_{M \in [M_l, M_u]} \mathbb{P}\big(\Omega^k \;\big|\; |\overline{\Lambda}^k| = M\big) \\ &\qquad + K \cdot \mathbb{P}\big(|\overline{\Lambda}^k| \notin [M_l, M_u]\big). \end{aligned} \tag{38}$$

We start with the estimate of the second term in the sum above, the probability of the number of zero coefficients in a given row being below M_l or above M_u.

Lemma C.2: Consider $0 < \varepsilon < 1$. Setting $M_l = N(1-p)(1-\varepsilon)$ and $M_u = N(1-p)(1+\varepsilon)$ we get that

$$\mathbb{P}\big(|\overline{\Lambda}^k| \notin [M_l, M_u]\big) \leq 2\exp(-2N(1-p)^2\varepsilon^2). \tag{39}$$

We will estimate the first term in (38) by splitting it into two terms that we will estimate separately

$$\begin{aligned}\mathbb{P}\big(\Omega_k \;\big|\; |\overline{\Lambda}^k| = M\big) &\leq \mathbb{P}\big(R_q(\bar{X}_k) < \alpha \;\big|\; |\overline{\Lambda}^k| = M\big) \\ &\qquad + \mathbb{P}\big(\|X_k(s^k)^\star\|_q > \beta \;\big|\; |\overline{\Lambda}^k| = M\big). \end{aligned} \tag{40}$$

C. Typical Size of $\alpha_q(X_0)$

Now we estimate the typical size of the largest ℓ_q ball we can inscribe into the image of the unit cube $Q^{|\overline{\Lambda}^k|}$ by $\bar{X}_k$ when $|\overline{\Lambda}^k| = M$. For simplicity we write L for $K-1$, and we denote $A = \bar{X}_k$. From Lemma B.6 we know that we need to estimate the value of $\|A^\star z\|_1$ and compare it to $\|z\|_1$. We begin with some geometrical observations.

Lemma C.3: Let $\mathcal{X} = \{z_i\}$ be a finite $\varepsilon_{\mathcal{X}}$-cover for the unit $\ell_{q'}$ sphere in $\mathbb{R}^L$. Assume that we have both the lower bound

$$\|A^\star z_i\|_1 \geq \alpha, \;\; \forall z_i \in \mathcal{X};$$

and the upper bound

$$\|A^\star\|_{q' \to 1} = \sup_{\|v\|_{q'} \leq 1} \|A^\star v\|_1 \leq \delta.$$

Then $R_\infty(A) \geq \alpha - \delta\varepsilon_{\mathcal{X}}$.

Proof: By Lemma B.6 we only need to show that for all z with unit $\ell_{q'}$ norm we have $\|A^\star z\|_1 \geq \alpha - \delta\varepsilon_{\mathcal{X}}$. By definition of an $\varepsilon_{\mathcal{X}}$-cover, for all z with unit $\ell_{q'}$ norm we can find $z_i \in \mathcal{X}$ with $\|z - z_i\|_{q'} \leq \varepsilon_{\mathcal{X}}$. We then have

$$\begin{aligned}\|A^\star z\|_1 &\geq \|A^\star z_i\|_1 - \|A^\star(z - z_i)\|_1 \\ &\geq \alpha - \|A^\star\|_{q' \to 1} \cdot \|z - z_i\|_{q'} \geq \alpha - \delta\varepsilon_{\mathcal{X}}. \end{aligned}$$

□

We will therefore estimate a (typical) lower bound for the norm $\|A^\star z_i\|_1$, and an upper bound on the operator norm $\|A^\star\|_{q'\to 1}$. We specialize to the case $q = 2$, but other bounds could be derived for other values of q.

Lemma C.4: Let $A = (A_1 \dots A_M)$ be a random matrix of size $L \times M$, whose entries follow the distribution described in Subsection V-A, $A_{ij} = \xi_{ij} g_{ij}$, $i = 1 \dots L$, $j = 1 \dots M$. Let $z \in \mathbb{R}^L$ be a vector with $\|z\|_2 = 1$. We have the concentration bounds, for $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}\big(\|A^\star\|_{2\to 1} > M\sqrt{pL}(1+\varepsilon)\big) &\le 2\exp\left(-\frac{Mp\cdot\varepsilon^2}{2+\sqrt{2}\cdot\varepsilon}\right). \\ \mathbb{P}\big(\|A^\star z\|_1 \le Mp(\sqrt{\tfrac{2}{\pi}}-\varepsilon)\big) &\le 2\exp\left(-\frac{Mp\cdot\varepsilon^2}{2+\sqrt{2}\cdot\varepsilon}\right). \end{aligned} \tag{41}$$

Combining the above estimates we obtain

Corollary C.5: Let $0 < \varepsilon < 1$ and define

$$\alpha := Np(1-p)(1-\varepsilon)(\sqrt{\tfrac{2}{\pi}} - 2\varepsilon - \varepsilon^2) \tag{42}$$

Then, for all $M \in [M_l \ M_u]$ we have

$$\begin{aligned} &\mathbb{P}\big(R_2(\bar{X}_k) < \alpha \ \big|\ |\overline{\Lambda}^k| = M\big) \\ \le& 2\cdot\left[\left(\frac{3}{\varepsilon}\sqrt{\frac{K}{p}}\right)^K + 1\right]\cdot\exp\left(-\frac{Np(1-p)(1-\varepsilon)\varepsilon^2}{2+\sqrt{2}\varepsilon}\right) \end{aligned} \tag{43}$$

Proof: Given $\varepsilon_{\mathcal{X}} \in (0,1)$, we can choose an $\varepsilon_{\mathcal{X}}$-cover $\mathcal{X} = \{z_i\}$ for the unit ℓ_2 sphere in $\mathbb{R}^L$ with $|\mathcal{X}| \le (3/\varepsilon_{\mathcal{X}})^L$. For a random $L \times M$ matrix $A = (A_1 \dots A_M)$ distributed as in Lemma C.4 we have, combining Lemma C.3 with Lemma C.4 and using $\tilde{\alpha} := Mp(\sqrt{\frac{2}{\pi}} - \varepsilon)$ and $\delta := M\sqrt{pL}(1+\varepsilon)$,

$$\begin{aligned} &\mathbb{P}\big(R_2(A) < \tilde{\alpha} - \delta\varepsilon_{\mathcal{X}}\big) \\ &\qquad\le \sum_{z_i\in\mathcal{X}} P\big(\|A^\star z_i\|_1 \le \alpha\big) + \mathbb{P}\big(\|A^\star\|_{2\to 1} \ge \delta\big). \\ &\qquad\le [(3/\varepsilon_{\mathcal{X}})^L + 1]\cdot 2\exp\left(-\frac{Mp\cdot\varepsilon^2}{2+\sqrt{2}\cdot\varepsilon}\right) \end{aligned}$$

Setting $\varepsilon_{\mathcal{X}} = \varepsilon\sqrt{p/L}$ yields $\tilde{\alpha} - \delta\varepsilon_{\mathcal{X}} = Mp(\sqrt{\frac{2}{\pi}} - 2\varepsilon - \varepsilon^2)$. According to the probability split in (38), we need to find the maximum of the above expression for $M \in [M_l, M_u]$ which is achieved at $M = M_l = N(1-p)(1-\varepsilon)$.

□

D. Typical Size of $\|X_k(s^k)^\star\|_q$

We now estimate the size of $\|X_k(s^k)^\star\|_q$. We need the following theorem.

Theorem C.6: Let B be a random matrix of size $L \times n$, whose entries follow the distribution described in Subsection V-A, $B_{ij} = \xi_{ij} g_{ij}$, $i = 1 \dots L$, $j = 1 \dots n$, and s be a vector of length n with entries $s_j = \pm 1$,

$j = 1 \dots n$. Then for $\varepsilon' > 0$

$$\mathbb{P}\big(\|Bs\|_2^2 \geq Lnp(1+\varepsilon')\big) \leq 2\exp\left(-\frac{Lp(\varepsilon')^2}{6+2\varepsilon'}\right).$$

Applying this to the situation at hand, inserting $L = K-1$ and the worst case value for $n = N-M_l = N(p+\varepsilon-\varepsilon p)$ and setting $\varepsilon' = (N/L)\varepsilon$ we get:

Lemma C.7: Define

$$\beta := Np\sqrt{(\tfrac{K-1}{N}+\varepsilon)(1+\tfrac{\varepsilon}{p}-\varepsilon)}, \tag{44}$$

For any $M \in [M_l, M_u]$ we have

$$\begin{aligned}&\mathbb{P}\big(\|X_k(s^k)^\star\|_2 > \beta \big| |\overline{\Lambda}^k| = M\big)\\ \leq& 2\cdot\exp\left(-\frac{Np\varepsilon^2}{6\frac{K-1}{N}+2\varepsilon}\right).\end{aligned} \tag{45}$$

APPENDIX D
CONCENTRATION INEQUALITIES

Here we will sketch the proofs of the concentration inequalities used in the previous section. They are based on a special version of Bernstein's inequality, see e.g. [3].

Theorem D.1: Let Y_i, $i = 1 \dots M$, be independent random variables with

$$\mathbb{E}(Y_i^2) \leq v^2 \qquad \text{and} \qquad \mathbb{E}(|Y_i|^k) \leq \frac{1}{2}k!\, v^2 c^{k-2},\ k > 2. \tag{46}$$

Then

$$\mathbb{P}\Big(|\sum_{i=1}^{M}(Y_i - \mathbb{E}(Y_i))| > \varepsilon\Big) \leq 2\exp\Big(-\frac{\varepsilon^2}{2(Mv^2+c\varepsilon)}\Big).$$

We will also use Hoeffding's inequality.

Theorem D.2 (Hoeffding's inequality): Let $Y_1 \dots Y_N$ be independent random variables. Assume that the Y_n are almost surely bounded, meaning for $1 \leq i \leq N$ we have $\mathbb{P}(Y_n \in [a_n, b_n]) = 1$. Then, for the sum of these variables $S = Y_1 + \ldots + Y_N$ we have the inequality

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

which is valid for positive values of t. $\mathbb{E}(S)$ is the expected value of S.

A. Proof of Lemma C.2

In each row of X, the number of zero coefficients $|\overline{\Lambda}^k|$ is N minus the number of non-zero coefficients $|\Lambda^k|$, which is the sum of the indicator variables $\sum_n \xi_{kn}$. The ξ_{nk} are taking only the values zero and one, so we can use Hoeffding's inequality with $a_i = 0$, $b_i = 1$ and $\mathbb{E}(\sum_n \xi_{kn}) = pN$, leading to

$$\mathbb{P}(|\Lambda^k| - pN \geq Nt) \leq \exp(-2Nt^2).$$

Choosing $t=(1-p)\varepsilon$ and using $|\overline{\Lambda}^k| = N - |\Lambda^k|$ we get

$$\mathbb{P}(|\overline{\Lambda}^k| \leq N(1-p)(1-\varepsilon)) \leq \exp(-2N(1-p)^2\varepsilon^2).$$

To bound the converse probability that $|\overline{\Lambda}^k|$ is very large, we set $Y_n = 1-\xi_{kn}$ and again $t=(1-p)\varepsilon$ to get directly to

$$\mathbb{P}(|\overline{\Lambda}^k| \geq N(1-p)(1+\varepsilon)) \leq \exp(-2N(1-p)^2\varepsilon^2).$$

B. Proof of Theorem C.1

Since $\|x\|_1 = \sum_{i=1}^N \xi_i|g_i|$, we will use the Bernstein inequality with $Y_i = \xi_i \cdot |g_i|$. The moments of ξ_i are constant equal to p. The random variable $|g_i|$ follows a Chi-distribution of degree 1 so its moments are

$$\mathbb{E}(|g_i|^k) = 2^{\frac{k}{2}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{1}{2})} \tag{47}$$

Especially, we have $\mathbb{E}(Y_i) = p\sqrt{\frac{2}{\pi}}$ and $\mathbb{E}(|Y_i|^2) = p$, and using the recurrence relation for the Gamma function $\Gamma(t+1) = t\Gamma(t)$ and $\sqrt{2}/\Gamma(\frac{1}{2}) = \sqrt{\frac{2}{\pi}} < 1$ we can bound by induction the moments of Y_i for $k \geq 2$ as

$$\mathbb{E}(|Y_i|^k) \leq p \cdot \frac{k!}{2^{k/2}},\ k \geq 2, \tag{48}$$

so the moments suffice Condition (46) with $c = 1/\sqrt{2}$ and we get

$$\mathbb{P}(\|x\|_1 > Np\sqrt{\tfrac{2}{\pi}} + \varepsilon) \leq 2 \cdot \exp\left(-\frac{\varepsilon^2}{2(Mp + \varepsilon/\sqrt{2})}\right).$$

Setting $\varepsilon = Mp \cdot \varepsilon'$ yields the result.

C. Proof of Lemma C.4 – first part

To bound $\|A^\star\|_{2\to 1}$ we begin by using the crude bound $\|A^\star\|_{2\to 1} = \|A\|_{1\to 2} \leq \sum_{i=1}^M \|A_i\|_2$. We set $Y_i = \|A_i\|_2 = (\sum_{j=1}^L \xi_{ij}^2 g_{ij}^2)^{\frac{1}{2}}$. All Y_i are identically distributed so for the analysis we can drop the subscript i. We can calculate directly

$$\mathbb{E}(Y^2) = \mathbb{E}(\sum_{j=1}^L \xi_j^2 g_j^2) = pL.$$

For the higher order moments $k > 2$ we use a little trick to separate the expectation over ξ and g,

$$\mathbb{E}Y^k = \mathbb{E}_g\mathbb{E}_\xi\big(\sum_{j=1}^n \xi_j^2 g_j^2\big)^{\frac{k}{2}} = \mathbb{E}_g\Big(\big(\sum g_j^2\big)^{\frac{k}{2}}\mathbb{E}_\xi\big(\frac{\sum \xi_j^2 g_j^2}{\sum g_j^2}\big)^{\frac{k}{2}}\Big).$$

The fraction in the last expression is always smaller than 1 so for $k > 2$ we have

$$\mathbb{E}Y^k \leq \mathbb{E}_g\Big(\big(\sum g_j^2\big)^{\frac{k}{2}}\mathbb{E}_\xi\big(\frac{\sum \xi_j^2 g_j^2}{\sum g_j^2}\big)\Big) = p \cdot \mathbb{E}_g\left(\big(\sum g_j^2\big)^{\frac{k}{2}}\right).$$

The random variable $\tilde{Y} = \big(\sum_{j=1}^n g_j^2\big)^{\frac{1}{2}}$ follows a Chi-distribution of degree L so for its k-th moments we have the formula

$$\mathbb{E}(\tilde{Y}^k) = 2^{\frac{k}{2}} \frac{\Gamma(\frac{k+L}{2})}{\Gamma(\frac{L}{2})}.$$

A long and tedious calculation involving the recurrence formula for the Gamma function, Stirling's formula and treating both cases, k is even respectively odd, yields the bound $\mathbb{E}(\tilde{Y}^k) \leq (\frac{L}{2})^{k/2} k!$. This leads to $\mathbb{E}(Y^k) \leq p\frac{L}{2}^{k/2} k$, meaning that the higher order moments follow the decay condition in (46) for $c = \sqrt{L/2}$. Together with the following bound for the first order moment,

$$\mathbb{E}(Y) \leq \mathbb{E}(Y^2)^{\frac{1}{2}} = \sqrt{pL},$$

we get

$$\mathbb{P}\big(\|A^\star\|_{2\to 1} > M\sqrt{pL} + \varepsilon\big) \leq 2\exp\Big(-\frac{\varepsilon^2}{2(MpL + \varepsilon\sqrt{L/2})}\Big).$$

To get the version of the formula used in Section V simply set $\varepsilon = M\sqrt{pL}\cdot\varepsilon'$ and observe that since $p < 1$

$$\frac{\varepsilon^2}{2(MpL + \varepsilon\sqrt{L/2})} = \frac{M\sqrt{p}(\varepsilon')^2}{2\sqrt{p} + \sqrt{2}\varepsilon'} \geq \frac{Mp(\varepsilon')^2}{2 + \sqrt{2}\varepsilon'}$$

D. Proof of Lemma C.4 – second part

To lower bound $\|A^\star z\|_1$ we expand it as

$$\|A^\star z\|_1 = \sum_{i=1}^{M} |\langle A_i, z\rangle| = \sum_{i=1}^{M} |\sum_{j=1}^{n} \xi_{ij} g_{ij} z_j| := \sum_{i=1}^{M} Y_i.$$

The random variables Y_i all follow the same distribution so it suffices to calculate the moments of $Y = |\sum_{j=1}^{n} \xi_j g_j z_j|$. Define $\tilde{Y} = \sum_{j=1}^{n} \xi_j g_j z_j$. Since the g_k are i.i.d. zero mean Gaussians with variance $\sigma^2 = 1$, $\tilde{Y}$ is zero mean Gaussian with variance $\tilde{\sigma}^2 = \sum_{j=1}^{n} z_j^2 \xi_j^2 := \|z\xi\|_2^2$ and we get

$$\mathbb{E}(|Y|^k) = \mathbb{E}(|\tilde{Y}|^k|) = \mathbb{E}(|\|z\xi\|_2 \cdot g_1|^k) = \mathbb{E}_\xi(\|z\xi\|_2^k) \cdot \mathbb{E}_g(|g_1|^k) \tag{49}$$

Since $\|z\xi\|_2 \leq \|z\|_2 = 1$, we have for $k \geq 2$

$$\mathbb{E}_\xi(\|z\xi\|_2^k) \leq \mathbb{E}_\xi(\|z\xi\|_2^2) = \mathbb{E}_\xi\big(\sum_{j=1}^{n} z_j^2 \xi_j^2\big) \leq p,$$

while for $k = 1$ we get

$$\mathbb{E}_\xi(\|z\xi\|_2) = \mathbb{E}_\xi\Big(\big(\sum_{j=1}^{n} z_j^2 \xi_j^2\big)^{\frac{1}{2}}\Big) \geq \mathbb{E}_\xi\left(\sum_{j=1}^{n} z_j^2 \xi_j^2\right) = p.$$

Again, $|g_1|$ is Chi-distributed of degree 1 so its moments are given by (47) and the moments of Y_i are thus bounded by (48), which suffices the decay condition in (46) for $c = 1/\sqrt{2}$. As a result

$$\mathbb{P}\Big(\|A^\star z\|_1 < M\mathbb{E}(|Y|) - \varepsilon\Big) < 2\exp\left(-\frac{\varepsilon^2}{2(Mp + \varepsilon/\sqrt{2})}\right).$$

Together with the bound for $\mathbb{E}(|Y|) \geq p\sqrt{\frac{2}{\pi}}$, setting $\varepsilon = Mp\cdot\varepsilon'$ leads to the final form of the bound used in Section V.

E. Proof of Theorem C.6

We expand $\|Bs\|_2^2 = \sum_{i=1}^L |\langle B^i, s\rangle|^2$, where B^i denotes the i-th row of B. and set $Y_i = |\langle B^i, s\rangle|^2 = (\sum_{j=1}^n \xi_{ij} g_{ij} s_j)^2$. Since the Y_i are again identically distributed we drop the subscript i for the analysis. First we get,

$$\mathbb{E}(Y) = \mathbb{E}\Big(\big(\sum_{j=1}^n \xi_j g_j s_j \big)^2 \Big) = \mathbb{E}\Big(\sum_{j=1}^n \xi_j^2 g_j^2 s_j^2 \Big) = p \cdot n.$$

Observe that $\sum \xi_j g_j s_j$ is again Gaussian and distributed like $(\sum \xi_j^2 s_j^2)^{\frac{1}{2}} \cdot g_1 = \|\xi\|_2 \cdot g_1$. Hence,

$$\begin{aligned}\mathbb{E}(Y^k) = \mathbb{E}\Big(\big(\sum_{j=1}^n \xi_j g_j s_j \big)^{2k} \Big) &= \mathbb{E}_\xi \mathbb{E}_g(\|\xi\|_2^{2k} g_1^{2k}) \\ &= \mathbb{E}_\xi(\|\xi\|_2^{2k}) \mathbb{E}_g(g_1^{2k}).\end{aligned}$$

For the even Gaussian moments we have the formula $\mathbb{E}_g(g_1^{2k}) = \frac{(2k)!}{2^k k!}$, while the term depending on ξ can be bounded as

$$\begin{aligned}\mathbb{E}_\xi(\|\xi\|_2^{2k}) = \mathbb{E}_\xi\Big(\big(\sum_{j=1}^n \xi_j^2 \big)^k \Big) &= n^k \cdot \mathbb{E}_\xi\Big(\big(\frac{1}{n} \sum_{j=1}^n \xi_j^2 \big)^k \Big) \\ &\leq n^k \cdot \mathbb{E}_\xi\Big(\frac{1}{n} \sum_{j=1}^n \xi_j^2 \Big) = n^k \cdot p,\end{aligned}$$

leading to $E(Y^k) \leq p n^k \frac{(2k)!}{2^k k!}$. Especially for $k = 2$ we have $E(Y^2) \leq 3pn^2$ and so for $k > 2$ we can estimate

$$E(Y^k) \leq 3pn^2 \frac{1}{3} n^{k-2} \frac{(2k)!}{2^k k!} \leq \ldots \leq \frac{1}{2} \mathbb{E}(Y^2)(2n)^{k-2} k!,$$

meaning that the moments follow the decay condition in (46) with $c = 2n$ and therefore

$$\mathbb{P}\left(\|Bs\|_2^2 > Lnp + \varepsilon \right) \leq 2 \exp\left(- \frac{\varepsilon^2}{6pn^2 L + 2n\varepsilon} \right).$$

Again setting $\varepsilon = Lnp \cdot \varepsilon'$ leads to the final version.

APPENDIX E
PROOF OF MAIN THEOREM

First, we observe that if $p \leq 4/5$ and $K/N \leq 1/3$ all the appearing exponentials can be upper bounded by

$$\exp\Big(- Np(1-p) \frac{\varepsilon^2 (1 - 2\varepsilon)}{2} \Big).$$

Therefore, with the definition of α, β, γ in (42), (44) and (36) we obtain from Lemmata C.5, C.7, C.1 that we have

$$\frac{\alpha_2(X_0) - \beta_2(X_0)}{\gamma(X_0)} \geq \frac{\alpha - \beta}{\gamma}$$

except with probability at most

$$\begin{aligned}&2K \left[\left(\frac{3}{\varepsilon} \sqrt{\tfrac{K}{p}} \right)^K + 3 \right] \cdot \exp\Big(- Np(1-p) \frac{\varepsilon^2 (1-2\varepsilon)}{2} \Big) \\ &\leq 4K \left(\frac{3}{\varepsilon} \sqrt{\tfrac{K}{p}} \right)^K \cdot \exp\Big(- Np(1-p) \frac{\varepsilon^2 (1-2\varepsilon)}{2} \Big) \\ &= 4K \exp\Big(\tfrac{K}{2} \log\Big(\tfrac{9K}{\varepsilon^2 p} \Big) - Np(1-p) \tfrac{\varepsilon^2 (1-2\varepsilon)}{2} \Big). \qquad (50)\end{aligned}$$

Next, observe that for the right hand side to be smaller than 1, we need that $\varepsilon < 1/2$ and $Np(1-p)\varepsilon^2 > K$. Consequently

$$K/N < p(1-p)\varepsilon^2 < 1/16,$$

meaning that whenever $K/N > 1/3$ the probability bound is trivially true, and we only need to assume $p \leq 4/5$.

Now, from Theorem 4.1 we know that any sufficiently incoherent basis satisfying $\max_k \|\bar{m}_k\|_2 < (\alpha - \beta)/\gamma$ will therefore be locally identifiable by ℓ_1 minimization, except with probability at most equal to the right hand side in (50).

Inserting the values for α, β, γ from (42), (44) and (36) we can lower bound the maximally allowed coherence $(\alpha - \beta)/\gamma$ with

$$\begin{aligned}
&\frac{(1-p)(1-\varepsilon)(\sqrt{\frac{2}{\pi}} - 2\varepsilon - \varepsilon^2) - \sqrt{(\frac{K}{N} + \varepsilon)(1 + \frac{\varepsilon}{p} - \varepsilon)}}{(\sqrt{\frac{2}{\pi}} + \varepsilon)} \\
&\geq (1-p)\cdot(1-5\varepsilon) - \sqrt{\tfrac{\pi}{2}\left(\tfrac{K}{N} + \varepsilon\right)\left(1 + \tfrac{\varepsilon}{p}\right)}.
\end{aligned}$$

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] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, March 1962.

[4] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Math*, 59(8):1207–1223, 2005.

[5] 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.

[6] R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory*, 38(2):713–718, March 1992.

[7] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[8] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approximation*, 13:57–98, 1997. Springer-Verlag New York Inc.

[9] 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.

[10] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decompositions. *IEEE Transactions on Information Theory*, 47(7):2845–2862, November 2001.

[11] 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.

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

[13] 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.

[14] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, December 2003.

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

[16] 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.

[17] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, December 2008.

[18] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Computing*, 25(2):227–234, 1995.

[19] 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.

[20] G. Pfander, H. Rauhut, and J. Tanner. Identification of matrices having a sparse representation. *IEEE Transactions on Signal Processing*, 56(11):5376–5388, November 2008.

[21] G. Pisier, The volume of convex bodies and Banach space geometry, 2nd edition, *Cambridge University Press*, 1999.

[22] 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.

[23] 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.

[24] G. Tauböck, D. Eiwen, F. Hlawatsch and H. Rauhut, Compressive estimation of doubly selective channels: exploiting channel sparsity to improve spectral efficiency in multicarrier transmissions, *IEEE Journal of Selected Topics in Signal Processing*, to appear.

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

[26] J. Tropp. Just relax: Convex programming methods for subset selection and sparse approximation. *IEEE Transactions on Information Theory*, 51(3):1030–1051, March 2006.

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

[28] M. Yaghoobi, T. Blumensath, and M.E. Davies. Regularized dictionary learning for sparse approximation. In *Proc. EUSIPCO08*, 2008.

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