Local Identification of Overcomplete Dictionaries*

Karin Schnass karin.schnass@uibk.ac.at
University of Innsbruck
Technikerstraße 13
6020 Innsbruck, Austria

Abstract

* This is a correction of *Local Identification of Overcomplete Dictionaries*, which contains an error in the proof of Proposition 7, as pointed out by Sebastian Kaiser and Felix Krahmer. The published original together with the erratum is available from JMLR at `http://www.jmlr.org/papers/v16/schnass15a.html`.

This paper presents the first theoretical results showing that stable identification of overcomplete μ-coherent dictionaries $\Phi \in \mathbb{R}^{d\times K}$ is locally possible from training signals with sparsity levels S up to the order $O(\mu^{-2})$ and signal to noise ratios up to $O(\sqrt{d})$. In particular the dictionary is recoverable as the local maximum of a new maximisation criterion that generalises the K-means criterion. For this maximisation criterion results for asymptotic exact recovery for sparsity levels up to $O(\mu^{-1})$ and stable recovery for sparsity levels up to $O(\mu^{-2})$ as well as signal to noise ratios up to $O(\sqrt{d})$ are provided. These asymptotic results translate to finite sample size recovery results with high probability as long as the sample size N scales as $O(K^3dS\tilde{\varepsilon}^{-2})$, where the recovery precision $\tilde{\varepsilon}$ can go down to the asymptotically achievable precision. Further to actually find the local maxima of the new criterion, a very simple Iterative Thresholding& K (signed) Means algorithm (ITKM), which has complexity $O(dKN)$ in each iteration, is presented and its local efficiency is demonstrated in several experiments.

Keywords: dictionary learning, dictionary identification, sparse coding, sparse component analysis, vector quantisation, K-means, finite sample size, sample complexity, maximisation criterion, sparse representation

1. Introduction

Be it the 300 million photos uploaded to Facebook per day, the 800GB the large Hadron collider records per second or the 320.000GB per second it cannot record, it is clear that we have reached the age of big data. Indeed, in 2012, the amount of data existing worldwide is estimated to have reached 2.8 ZB = 2.800 billion GB and while 23 % of these data are expected to be useful if analysed, only 1% actually are. So how do we deal with this big data challenge? The key concept, that has driven data processing and data analysis in the past decade, is that even high-dimensional data has intrinsically low complexity, meaning that every data point y can be represented as linear combination of a sparse (small) number of elements or atoms $\phi_i \in \mathbb{R}^d$, $\|\phi\|_2 = 1$ of an overcomplete dictionary $\Phi = (\phi_1, \dots \phi_K)$, that is,

$$y \approx \Phi_I x_I = \sum_{i\in I} x(i)\varphi_i,$$

for a set I of size S, $|I| = S$, which is small compared to the ambient dimension, $S \ll d \leq K$. These sparse components do not only describe the data but the representations can also be used for a myriad of efficient sparsity based data processing schemes, ranging from denoising, [8], to compressed sensing, [7, 5]. Therefore a promising tool both for data analysis and data processing, that has emerged in the last years, is dictionary learning, also known as sparse coding or sparse component analysis. Dictionary learning addresses the fundamental question, how to automatically learn a dictionary, providing sparse representations for a given data class; that is, given N signals $y_n \in \mathbb{R}^d$, stored as columns in a matrix $Y = (y_1, \dots, y_N)$, find a decomposition

$$Y \approx \Phi X$$

into a $d \times K$ dictionary matrix Φ with unit norm columns and a $K \times N$ coefficient matrix with sparse columns.

Until recently the main research focus in dictionary learning has been on the development of algorithms. Thus by now there is an ample choice of learning algorithms, that perform well in experiments and are popular in applications, [9, 17, 18, 3, 32, 19, 27, 23]. However, slowly the interest is shifting and researchers are starting to investigate also the theoretical aspects of dictionary learning. Following the first theoretical insights, originating in the blind source separation community, [33, 11], there is now a set of generalisation bounds predicting how well a learned dictionary can be expected to sparsely approximate future data, [20, 30, 21, 14]. These results give a theoretical foundation for dictionary learning as data processing tool, for example for compression, but unfortunately do not give guarantees that an efficient algorithm will find/recover a good dictionary provided that it exists. However, in order to justify the use of dictionary learning as data analysis tool, for instance in blind source separation, it is important to provide conditions under which an algorithm or scheme can identify the dictionary from a finite number of training signals, that is, the sources from the mixtures. Following the first dictionary identification results for the ℓ_1-minimisation principle, which was suggested in Zibulevsky and Pearlmutter [33]/Plumbley [22], by Gribonval and Schnass [13], Geng et al. [10], Jenatton et al. [16] and for the ER-SPuD algorithm for learning a basis in [28], 2013 has seen a number of interesting developments. First in Schnass [24] it was shown that the K-SVD minimisation principle suggested in Aharon et al. [3] can locally identify overcomplete tight dictionaries, then in Arora et al. [4], Agarwal et al. [2] algorithms with *global* identification guarantees for coherent dictionaries were presented and finally in Agarwal et al. [1] it was shown that an alternating minimisation method is locally convergent to the correct generating dictionary. One aspect that all these results have in common is that the sparsity level of the training signals required for successful identification is of order $O(\mu^{-1})$ or $O(\sqrt{d})$ for incoherent dictionaries. Considering that on average sparse recovery in a given dictionary is successful for sparsity levels $O(\mu^{-2})$, [29, 25], and that for dictionary learning we usually have a lot of training signals at our disposal, the same sparsity level should be sufficient for dictionary learning and indeed in this paper we provide the first indication that global dictionary identification could be possible for sparsity levels $O(\mu^{-2})$ by proving that it is locally possible. Further we show that in experiments a very simple iterative algorithm, based on thresholding and K signed means, is locally successful.

The paper is organised as follows. After introducing all necessary notation in Section 2 we present a new optimisation criterion, motivated by the analysis of the K-SVD principle,

[24], in Section 3. In Section 4 we give asymptotic identification results both for exact and stable recovery, which in Section 5 are extended to results to finite sample sizes. Section 6 provides an algorithm for actually finding a local optimum and some experiments confirming the theoretical results. Finally in the last section we compare the results for the new criterion to existing identification results, discuss the implications of these local results for global dictionary identification algorithms and point out directions for future research.

2. Notations and Conventions

Before we jump into the fray, we collect some definitions and lose a few words on notations; usually subscripted letters will denote vectors with the exception of c and ε where they are numbers, eg. $(x_1, \dots, x_K) = X \in \mathbb{R}^{d \times K}$ vs. $c = (c_1, \dots, c_K) \in \mathbb{R}^K$, however, it should always be clear from the context what we are dealing with.

We consider a **dictionary** Φ a collection of K unit norm vectors $\phi_i \in \mathbb{R}^d$, $\|\phi_i\|_2 = 1$. By abuse of notation we will also refer to the $d \times K$ matrix collecting the atoms as its columns as the dictionary, that is $\Phi = (\phi_i, \dots \phi_K)$. The maximal absolute inner product between two different atoms is called the **coherence** μ of the dictionary, $\mu = \max_{i \neq j} |\langle \phi_i, \phi_j \rangle|$. By Φ_I we denote the restriction of the dictionary to the atoms indexed by I, that is $\Phi_I = (\phi_{i_1} \dots \phi_{i_S})$, $i_j \in I$. We indicate the conjugate transpose of a matrix with a *, for example $\Phi^\star$ would be the transpose of Φ.

The set of all dictionaries of a given size $(d \times K)$ is denoted by $\mathcal{D}$. For two dictionaries $\Phi, \Psi \in \mathcal{D}$ we define the distance between each other as the maximal distance between two corresponding atoms,

$$d(\Phi, \Psi) := \max_i \|\phi_i - \psi_i\|_2.$$

We consider a **frame** F a collection of $K \geq d$ vectors $f_i \in \mathbb{R}^d$ for which there exist two positive frame constants A, B such that for all $v \in \mathbb{R}^d$ we have

$$A\|v\|_2^2 \leq \sum_{i=1}^{K} |\langle f_i, v \rangle|^2 \leq B\|v\|_2^2. \tag{1}$$

From (1) follows that F, interpreted as $d \times K$ matrix, has rank d and that its non-zero singular values are in the interval $[\sqrt{A}, \sqrt{B}]$. If B can be chosen equal to A, that is $B = A$, the frame is called **tight**. If all frame elements f_i have unit norm, we call F a unit norm frame. For more details on frames, see e.g. [6].

Finally we introduce the Landau symbols O, o to characterise the growth of a function. We write

$$\begin{aligned} f(t) = O(g(t)) \quad &\text{if} \quad \lim_{t \to 0/\infty} f(t)/g(t) = C < \infty \\ \text{and} \quad f(t) = o(g(t)) \quad &\text{if} \quad \lim_{t \to 0/\infty} f(\varepsilon)/g(\varepsilon) = 0. \end{aligned}$$

.

3. A Response Maximisation Criterion

One of the origins of dictionary learning can be found in the field of vector quantisation, where the aim is to find a codebook (dictionary) such that the codewords (atoms) closely represent the data, that is

$$\min_{\Phi,X} \|Y-\Phi X\|_F^2 \quad \text{s.t.} \quad x_n \in \{e_i\}_i.$$

Indeed the vector quantisation problem can be seen as an extreme case of dictionary learning, where we do not only want all our signals to be approximately 1-sparse but also want the single non-zero coefficient equal to one. On the other hand we allow the atoms (codewords) to have any length. The problem above is usually solved by a K-means algorithm, which alternatively separates the training data into K clusters, each assigned to one codeword, and then updates the codeword to be the mean of the associated train signals. For more detailed information about vector quantisation or the K-means algorithm see for instance [12] or the introduction of [3]. If we relax the condition that each coefficient has to be positive and allow also negative -1 coefficients, but in turn ask for the atoms to have unit norm, we are already getting closer to the concept of 1-sparse dictionary learning,

$$\min_{\Phi\in\mathcal{D},X} \|Y-\Phi X\|_F^2 \quad \text{s.t.} \quad x_n \in \{\pm e_i\}_i,$$

This minimisation problem can be rewritten as

$$\begin{aligned}\min_{\Phi\in\mathcal{D}}\sum_n \min_{i,\sigma_i=\pm1}\|y_n-\sigma_i\phi_i\|_2^2 &= \min_{\Phi\in\mathcal{D}}\sum_n \min_{i,\sigma_i=\pm1}\|y_n\|_2^2-2\sigma_i\langle y_n,\phi_i\rangle+\|\phi_i\|_2^2\\ &= \|Y\|_F^2+N-2\max_{\Phi\in\mathcal{D}}\sum_n\max_i|\langle y_n,\phi_i\rangle|,\end{aligned}$$

and is therefore equivalent to the maximisation problem

$$\max_{\Phi\in\mathcal{D}}\sum_n\max_i|\langle y_n,\phi_i\rangle|. \tag{2}$$

A local maximum of (2) can be found with a signed K-means algorithm, which assigns each training signal to the atom of the current dictionary giving the largest response in absolute value and then updating the atom as normalised signed mean of the associated training signals, see Section 6 for more details. The question now is how do we go from these 1-sparse dictionary learning formulations to S-sparse formulations with $S>1$. The most common generalisation, which provides the starting point for the MOD and the K-SVD algorithm, is to give up all constraints on the coefficients except for S-sparsity and to minimise,

$$(P_{P2}) \qquad \min_{\Phi\in\mathcal{D},X}\|Y-\Phi X\|_F^2 \quad \text{s.t.} \quad \|x_n\|_0\le S. \tag{3}$$

However, rewriting the problem we see that this formulation does not reduce to the same maximisation problem in case $S=1$. Then the best one term approximation in the given dictionary is simply the largest projection onto one atom and we have

$$\begin{aligned}\min_{\Phi\in\mathcal{D}}\sum_n \min_{i,x_i}\|y_n - x_i\phi_i\|_2^2 &= \min_{\Phi\in\mathcal{D}}\sum_n \min_i \|y_n - \langle\phi_i, y_n\rangle\phi_i\|_2^2\\ &= \|Y\|_F^2 - \max_{\Phi\in\mathcal{D}}\sum_n \max_i |\langle y_n,\phi_i\rangle|^2,\end{aligned}$$

leading instead to the maximisation problem

$$\max_{\Phi\in\mathcal{D}}\sum_n \max_i |\langle y_n,\phi_i\rangle|^2 \qquad \text{vs.} \qquad \max_{\Phi\in\mathcal{D}}\sum_n \max_i |\langle y_n,\phi_i\rangle|.$$

A local maximum can now be found using the same partitioning strategy as before but updating the atoms as largest singular vector rather than signed mean of the associated training signals, requiring K SVDs as opposed to K means. While the minimisation problem (3) is definitely the most effective generalisation for dictionary learning, when the goal is compression, it brings with it some complications when used as analysis tool. Indeed in [24] it has been shown that for $S = 1$ the K-SVD criterion (3) can only identify the underlying dictionary from sparse random mixtures to arbitrary precision (given enough training samples) if this dictionary is tight and it is conjectured that the same holds for $S \geq 1$. Roughly simplified the reason for this is that for random sparse signals $\Phi_I x_I$ and an ε-perturbation Ψ the average of the largest squared response behaves like

$$\frac{1}{2}\left(1-\frac{\varepsilon^2}{2}+c(\Psi)\right)^2+\frac{1}{2}\left(1-\frac{\varepsilon^2}{2}-c(\Psi)\right)^2 = 1-\varepsilon^2+\frac{\varepsilon^4}{4}+c(\Psi)^2.$$

If Φ is tight the term $c(\Psi)$ is constant over all dictionaries and therefore there is a local maximum at Φ. From the above we also see that the average of the largest *absolute* response should behave like

$$\frac{1}{2}\left|1-\frac{\varepsilon^2}{2}+c(\Psi)\right|+\frac{1}{2}\left|1-\frac{\varepsilon^2}{2}-c(\Psi)\right| = 1-\frac{\varepsilon^2}{2},$$

meaning that we should have a maximum at Φ also if it is non-tight. This suggests as alternative way to generalise the K-means optimisation principle for dictionary identification to simply maximise the absolute norm of the S-largest responses,

$$(P_{R1}) \qquad \max_{\Psi\in\mathcal{D}}\sum_n \max_{|I|=S}\|\Psi_I^\star y_n\|_1. \tag{4}$$

Other than for the K-SVD criterion it is not obvious that there should be a local optimum of (4) at Φ even if all signals y_n are perfectly S-sparse in Φ. Therefore it is quite intriguing that we will not only be able to prove local identifiability of any generating dictionary via (4) from randomly sparse signals, but that these identifiability properties are stable under coherence and noise. However, before we get to the main result in Theorem 5 on page 11, we first have to lay the foundation, by providing suitable random signal models and by studying the asymptotic identifiability properties of the new principle.

4. Asymptotic Results

We can get to the asymptotic version of the S-response maximisation principle in (4) simply by replacing the sum over the training signals with the expectation, leading to

$$\max_{\Psi\in\mathcal{D}} \mathbb{E}_y \left(\max_{|I|=S} \|\Psi_I^\star y\|_1\right). \tag{5}$$

Next we need a random sparse coefficient model to generate our signals y. We make the following definition, see also [24].

Definition 1 *A probability distribution (measure) ν on the unit sphere $S^{K-1} \subset \mathbb{R}^K$ is called symmetric if for all measurable sets $\mathcal{X} \subseteq S^{K-1}$, for all sign sequences $\sigma \in \{-1,1\}^K$ and all permutations p we have*

$$\begin{aligned} \nu(\sigma\mathcal{X}) = \nu(\mathcal{X}), &\quad \textit{where} \quad \sigma\mathcal{X} := \{(\sigma_1 x_1, \dots, \sigma_K x_K) : x \in \mathcal{X}\}, \quad \textit{and} \\ \nu(p(\mathcal{X})) = \nu(\mathcal{X}), &\quad \textit{where} \quad p(\mathcal{X}) := \{(x_{p(1)}, \dots, x_{p(K)}) : x \in \mathcal{X}\}. \end{aligned}$$

Setting $y = \Phi x$ where x is drawn from a symmetric probability measure ν on the unit sphere has the advantage that for dictionaries, which are an orthonormal basis, the resulting signals have unit norm and for general dictionaries the signals have unit square norm in expectation, that is $\mathbb{E}(\|y\|_2^2) = 1$. This reflects the situation in practical application, where we would normalise the signals in order to equally weight their importance.

One example of such a probability measure can be constructed from a non-negative, non-increasing sequence $c \in \mathbb{R}^K$ with $\|c\|_2 = 1$, which we permute uniformly at random and provide with random ± signs. To be precise for a permutation $p : \{1, ..., K\} \to \{1, ..., K\}$ and a sign sequence σ, $\sigma_i = \pm 1$, we define the sequence $c_{p,\sigma}$ component-wise as $c_{p,\sigma}(i) := \sigma_i c_{p(i)}$, and set $\nu(x) = (2^K K!)^{-1}$ if there exist p, σ such that $x = c_{p,\sigma}$ and $\nu(x) = 0$ otherwise. While being very simple this measure exhibits all the necessary structure and indeed in our proofs we will reduce the general case of a symmetric measure to this simple case.

So far we have not incorporated any sparse structure in our coefficient distribution. To motivate the sparsity requirements on our coefficients we will recycle the simple negative example of a sparse coefficient distribution for which the original generating dictionary is not at a local maximum of (5) with $S = 1$ from [24].

Example 1 *Let U be an orthonormal basis and let the signals be constructed as $y = \Phi x$. If x is randomly 2-sparse with 'flat' coefficients, that is drawn from the simple symmetric probability measure with base sequence c, where $c_1 = c_2 = 1/\sqrt{2}$, $c_i = 0$ for $i \geq 3$, then U is not a local maximum of (5) with $S = 1$.*

Indeed, since the signals are all 2-sparse, the maximal inner product with all atoms in U is the same as the maximal inner product with only $d - 1$ atoms. This degree of freedom we can use to construct an ascent direction. Choose $U_\varepsilon = (u_1, \dots, u_{d-1}, (u_d + \varepsilon u_1)/\sqrt{1+\varepsilon^2})$, *then we have*

$$\begin{aligned}\mathbb{E}_y\left(\|U_\varepsilon^\star y\|_\infty\right) &= \mathbb{E}_x\left(\left\|\left(x_1,\dots,x_{d-1},\frac{x_d+\varepsilon x_1}{\sqrt{1+\varepsilon^2}}\right)\right\|_\infty\right)\\ &= \mathbb{E}_x \max\left\{\frac{1}{\sqrt{2}},\left|\frac{x_d+\varepsilon x_1}{\sqrt{1+\varepsilon^2}}\right|\right\}\\ &= \frac{1}{\sqrt{2}}\left(1-\frac{1}{d(d-1)}+\frac{1}{d(d-1)}\frac{1+\varepsilon}{\sqrt{1+\varepsilon^2}}\right)\\ &\geq \frac{1}{\sqrt{2}}\left(1+\frac{1}{d(d-1)}\frac{\varepsilon-\varepsilon^2}{1+\varepsilon^2}\right) > \frac{1}{\sqrt{2}} = \mathbb{E}_y\left(\|U^\star y\|_\infty\right).\end{aligned}$$

From the above example we see that, in order to have a local maximum of (5) with $S=1$ at the original dictionary, we need our signals to be truly 1-sparse, that is, we need to have a decay between the first and the second largest coefficient. In the following sections we will study how large this decay should be to have a local maximum exactly at or near to the generating dictionary for more general dictionaries and sparsity levels.

4.1 Exact Recovery

To warm up we first provide an asymptotic exact dictionary identification result for (5) for incoherent dictionaries in the noiseless setting.

Theorem 2 *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ. Let the coefficients x be drawn from a symmetric probability distribution ν on the unit sphere $S^{K-1}\subset\mathbb{R}^K$ and assume that the signals are generated as $y=\Phi x$. If there exists $\beta>0$ such that for $c_1(x)\geq c_2(x)\geq\ldots\geq c_K(x)\geq 0$ the non-increasing rearrangement of the absolute values of the components of x we have $c_S(x)-c_{S+1}(x)-2\mu\|x\|_1\geq\beta$ almost surely, that is*

$$\nu\left(c_S(x)-c_{S+1}(x)-2\mu\|x\|_1\geq\beta\right)=1, \tag{6}$$

then there is a local maximum of (5) at Φ.
Moreover for $\Psi\neq\Phi$ we have $\mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right)<\mathbb{E}_y\left(\max_{|I|=S}\|\Phi_I^\star y\|_1\right)$ as soon as

$$\varepsilon < \frac{\beta}{1+3\sqrt{\log\left(\frac{25K^2S\sqrt{B}}{\beta(\bar{c}_1+\ldots+\bar{c}_S)}\right)}}, \tag{7}$$

where $\bar{c}_i := \mathbb{E}_x(c_i(x))$.

Proof idea We briefly sketch the main ideas of the proof, which are the same as for the corresponding theorem for the K-SVD principle in [24]. For self-containedness of the paper the full proof is included in Appendix A.1.
Assume that we have the case of a simple probability measure based on one sequence c, that is $x=c_{p,\sigma}$. For any fixed permutation p the condition in (6) ensures that for all sign sequences σ, and consequently all signals, the maximal S responses of the original dictionary Φ are attained at $I_p=p^{-1}\left(\{1\ldots S\}\right)$ and that there is a gap of size β to the remaining responses.

For an ε-perturbation of the generating dictionary we have $\psi_i \approx (1-\varepsilon_i^2/2)\phi_i + \varepsilon_i z_i$ for some unit vectors z_i with $\langle z_i, \phi_i \rangle = 0$ and $\varepsilon_i \leq \varepsilon$. Now for most sign sequences the contribution of $\varepsilon_i z_i$ to the response $\langle \psi_i, \Phi c_{p,\sigma}\rangle$ will be smaller than $\beta/2$ so the maximal S responses will still be attained at I_p. Comparing the loss of the perturbed dictionary over the typical sign sequences of all permutations, which scales as $\frac{(c_1+\ldots+c_S)}{2K}\sum \varepsilon_i^2$, to the maximal gain $S\varepsilon\sqrt{B}$ over the approximately $2\sum_i \exp\left(-\beta^2/\varepsilon_i^2\right)$ atypical sign sequences shows that there is a maximum at the original dictionary. The general result follows from an integration argument. ■

As already mentioned, while for the K-SVD criterion (3) there is always an optimum at the generating dictionary if all training signals are S-sparse, this it is not obvious for the response principle. Indeed, in the special case where all the training signals are exactly S-sparse, $c_{S+1}(x) = 0$ almost surely, we get as additional condition to ensure asymptotic recoverability,

$$c_S(x) - 2\mu \sum_{s=1}^{S} c_s(x) \geq \beta > 0, \qquad \text{almost surely.}$$

To get a better feeling for this constraint we bound the sum over the S largest reponses by S times the largest response, $\sum_{s=1}^{S} c_s(x) \leq S c_1(x)$ and arrive at the condition

$$\frac{c_S(x)}{c_1(x)} \gtrsim 2S\mu, \tag{8}$$

which is the classical condition under which simple thresholding will find the support of an exactly S-sparse signal, compare for instance Schnass and Vandergheynst [26].

4.2 Stability under Coherence and Noise

While giving a first insight into the identification properties of the response principle, Theorem 2 suffers from two main limitations.
First, the required condition on the coherence of the dictionary with respect to the decay of the coefficients, $c_S(x) - c_{S+1}(x) - 2\mu\|c(x)\|_1 > 0$, is unfortunately quite strict. In the most favourable case of exactly S sparse signals with equally sized coefficients, $c_1(x) = c_S(x) = 1/\sqrt{S}$, we see from (8) that we can only identify dictionaries from very sparse signals, where $S =\lesssim \mu^{-1}$. In case of very incoherent dictionaries with $\mu = O(1/\sqrt{d})$ this means that $S \lesssim \sqrt{d}$. However, typically, that is for most sign sequences σ we have

$$|\langle \phi_i, \Phi c_{p,\sigma}\rangle| = \Big|\sigma_i c_{p(i)} + \sum_{j\neq i} \sigma_j c_{p(j)} \langle \phi_i, \phi_j\rangle\Big| \approx c_{p(i)} \pm \left(\sum_{j\neq i} c_{p(j)}^2 |\langle \phi_i, \phi_j\rangle|^2\right)^{1/2} \approx c_{p(i)} \pm \mu,$$

which indicates that a condition of the form $\mu \lesssim c_S - c_{S+1}$ may be strong enough to guarantee (approximate) recoverability of the dictionary. Assuming again the most favourable case of equally sized coefficients, we could therefore identify dictionaries from signals with sparsity levels of the order $S \lesssim \mu^{-2}$, which in case of incoherent dictionaries means of the order of the ambient dimension, $S \lesssim d$.

The second limitation of Theorem 2 is that, even if it allows for not exact S-sparseness of the signals, it does not take into account noise. Our next goal is therefore to extend the exact identification result in Theorem 2 to a stable identification result for less sparse (larger S) and noisy signals. For this task we first need to amend our signal model to incorporate noise. We would like to consider unbounded white noise, but also keep the property that in expectation the signals have unit square norm. Further for the next section, where we want to transform our asymptotic identification results to results for finite sample sizes, it will be convenient if our signals are bounded. These considerations lead to the following model,

$$y = \frac{\Phi x + r}{\sqrt{1 + \|r\|_2^2}}, \tag{9}$$

where $r = (r(1) \dots r(d))$ is a centred random subgaussian vector with parameter ρ, that is the entries $r(i)$ are independent and satisfy $\mathbb{E}(e^{t \cdot r(i)}) \leq e^{t^2 \rho^2 / 2}$.
Employing this noisy signal model and formalising the ideas about the typical gap size between responses of the generating dictionary within and without the true support, leads to the following theorem.

Theorem 3 *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ. Let the coefficients x be drawn from a symmetric probability distribution ν on the unit sphere $S^{K-1} \subset \mathbb{R}^K$. Further let $r = (r(1) \dots r(d))$ be a centred random subgaussian noise-vector with parameter ρ and assume that the signals are generated according to the noisy signal model in (9). If there exists $\beta > 0$ such that for $c_1(x) \geq c_2(x) \geq \ldots \geq c_K(x) \geq 0$ the non-increasing rearrangement of the absolute values of the components of x we have $c_S(x) - c_{S+1}(x) \geq \beta$ almost surely and*

$$\max\{\mu, \rho\} \leq \frac{\beta}{\sqrt{72(\log a + \log\log a)}} \quad \text{for} \quad a = \frac{112 K^2 S(\sqrt{B} + 1)}{C_r \beta (\bar{c}_1 + \ldots + \bar{c}_S)}, \tag{10}$$

where $C_r = \mathbb{E}_r\left((1 + \|r\|_2^2)^{-1/2}\right)$ and $\bar{c}_i := \mathbb{E}_x(c_i(x))$, then there is a local maximum of (5) at $\tilde{\Psi}$ satisfying,

$$d(\tilde{\Psi}, \Phi) \leq \frac{12 S K^2 \sqrt{B}}{C_r (\bar{c}_1 + \ldots + \bar{c}_S)} \exp\left(\frac{-\beta^2}{72 \max\{\mu^2, \rho^2\}}\right). \tag{11}$$

Proof idea As outlined at the beginning of the section the main ingredient we have to add to the proof idea of Theorem 2 is a probabilistic argument to substitute the condition guaranteeing that the S largest responses of the generating dictionary are I_p. Due to concentration of measure we get that for most sign sequences, and therefore most signals, the maximum is still attained at I_p. Moreover the gap to the remaining responses is actually large enough to accommodate relatively high levels of noise and/or perturbations.
The detailed proof can be found in Appendix A.2. ■

Let us make some observations about the last result.
First, we want to point out that for sub-gaussian noise with parameter ρ, the quantity $C_r = \mathbb{E}_r\left((1 + \|r\|_2^2)^{-1/2}\right)$ in the statement above is well behaved. If for example the $r(i)$

are iid Bernoulli-variables, that is $P(r(i) = \pm\rho) = \frac{1}{2}$, we have $C_r = (1+d\rho^2)^{-1/2}$. In general we have the following estimate due for instance to Theorem 1 in [15]. Since we have

$$\mathbb{P}\left(\|x\|_2^2 \geq \rho^2(d + 2\sqrt{dt} + 2t\right) \leq e^{-t},$$

setting $t = d$, we get $\mathbb{P}\left(\|x\|_2^2 \geq 5d\rho^2\right) \leq e^{-d}$, which leads to

$$\mathbb{E}_r\left(\frac{1}{\sqrt{1+\|r\|_2^2}}\right) \geq \frac{(1-e^{-d})}{\sqrt{1+5d\rho^2}}.$$

Also to illustrate the result we again specialise it to the most favourable case of exactly S-sparse signals with balanced coefficients, that is $c_S(x) = S^{-1/2}$. Assuming white Gaussian noise with variance ρ_G^2 we see that identification is possible even for expected signal to noise ratios of the order $O(\frac{S}{d})$, that is

$$\frac{\mathbb{E}(\|\Phi x\|_2^2)}{\mathbb{E}(\|r\|_2^2)} \gtrsim \frac{S}{d}.$$

Similarly by specialising Theorem 3 to the case of exactly S-sparse and noiseless signals we get - to the best of our knowledge - the first result establishing that locally it is possible to stably identify dictionaries from signals with sparsity levels beyond the spark of the generating dictionary. Indeed, even if some of the S-sparse signals could have representations in Φ that require less than S atoms, there will still be a local maximum of the asymptotic criterion close to the original dictionary as long as the smallest coefficient of each signal is of the order $O(\mu)$, which in the most favourable case means that we can have $S \lesssim \mu^{-2}$ or $S \lesssim d$. The quality of this result is on par with the best results for finding sparse approximations in a *given* dictionary, which say that on average Basis Pursuit or thresholding can find the correct sparse support even for signals with sparsity levels of the order of the ambient dimension [29, 25].

Next note that with the available tools it would be possible to consider also a signal model where a small fraction of the coefficients violates the decay condition $c_S(x) - c_{S+1}(x) \geq \beta$ and still have stability. However, we leave explorations into that direction to the interested reader and instead turn to the study of the criterion for a finite number of training samples.

5. Finite Sample Size Results

In this section we will transform the two asymptotic results from the last section into results for finite sample sizes, that is, we will study when Φ is close to a local maximum of

$$\max_{\Psi\in\mathcal{D}} \frac{1}{N}\sum_{n=1}^{N} \max_{|I|=S} \|\Psi_I^\star y_n\|_1, \tag{12}$$

assuming that the y_n are following either the noise-free or the noisy signal model. For convenience we will do the analysis for the normalised version (12) of the S-response criterion (4).

Theorem 4 *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ. Let the coefficients x_n be drawn from a symmetric probability distribution ν on the unit sphere*

$S^{K-1} \subset \mathbb{R}^K$ and assume that the signals are generated as $y_n = \Phi x_n$. If there exists $\beta > 0$ such that for $c_1(x_n) \geq c_2(x_n) \geq \ldots \geq c_K(x_n) \geq 0$ the non-increasing rearrangement of the absolute values of the components of x_n we have $c_S(x_n) - c_{S+1}(x_n) - 2\mu\|x_n\|_1 \geq \beta$ almost surely and the target precision $\tilde{\varepsilon}$ satisfies

$$\tilde{\varepsilon} \leq \frac{\beta}{1 + 3\sqrt{\log\left(\frac{50K^2S\sqrt{B}}{\beta(\bar{c}_1+\ldots+\bar{c}_S)}\right)}},$$

where $\bar{c}_i := \mathbb{E}_{x_n}(c_i(x_n))$, then except with probability,

$$2\exp\left(-\frac{N\tilde{\varepsilon}^2(\bar{c}_1 + \ldots + \bar{c}_S)^2}{129S^2K^2B} + Kd\log\left(\frac{25SK\sqrt{B}}{\tilde{\varepsilon}(\bar{c}_1 + \ldots + \bar{c}_S)}\right)\right),$$

there is a local maximum of (12) *respectively* (4) *at $\tilde{\Psi}$ satisfying,*

$$d(\tilde{\Psi}, \Phi) \leq \tilde{\varepsilon} + \frac{\tilde{\varepsilon}^2}{4K}.$$

Theorem 5 *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ. Let the coefficients x_n be drawn from a symmetric probability distribution ν on the unit sphere $S^{K-1} \subset \mathbb{R}^K$. Further let $r_n = (r_n(1) \ldots r_n(d))$ be i.i.d. centred random subgaussian noise-vectors with parameter ρ and assume that the signals are generated according to the noisy signal model in* (9)*. If there exists $\beta > 0$ such that for $c_1(x_n) \geq c_2(x_n) \geq \ldots \geq c_K(x_n) \geq 0$ the non-increasing rearrangement of the absolute values of the components of x we have $c_S(x_n) - c_{S+1}(x_n) \geq \beta$ almost surely and if the target precision $\tilde{\varepsilon}$, the noiseparameter ρ and the coherence μ satisfy*

$$\tilde{\varepsilon} \leq \frac{\beta}{\frac{9}{4} + 9\sqrt{\log a}} \qquad \textit{and} \tag{13}$$
$$\max\{\mu, \rho\} \leq \frac{\beta}{\sqrt{72(\log a + \log\log a)}} \qquad \textit{for} \quad a = \frac{150K^2S(\sqrt{B}+1)}{C_r\beta(\bar{c}_1 + \ldots + \bar{c}_S)},$$

where $C_r = \mathbb{E}_{r_n}\left((1 + \|r_n\|_2^2)^{-1/2}\right)$ and $\bar{c}_i := \mathbb{E}_{x_n}(c_i(x_n))$, then except with probability

$$2\exp\left(-\frac{N\tilde{\varepsilon}_{\mu,\rho}^2(\bar{c}_1 + \ldots + \bar{c}_S)^2}{513C_r^2S^2K^2\left(\sqrt{B}+1\right)^2} + Kd\log\left(\frac{49SK\left(\sqrt{B}+1\right)}{\varepsilon_{\mu,\rho}(\bar{c}_1 + \ldots + \bar{c}_S)}\right)\right),$$
$$\textit{where} \qquad \tilde{\varepsilon}_{\mu,\rho} = \max\left\{\tilde{\varepsilon}, \frac{16SK^2\left(\sqrt{B}+1\right)}{C_r(\bar{c}_1 + \ldots + \bar{c}_S)}\exp\left(-\frac{\beta^2}{72\max\{\mu^2,\rho^2\}}\right)\right\},$$

there is a local maximum of (12) *respectively* (4) *at $\tilde{\Psi}$, satisfying*

$$d(\tilde{\Psi}, \Phi) \leq \tilde{\varepsilon}_{\mu,\rho} + \frac{\tilde{\varepsilon}_{\mu,\rho}^2}{16K}.$$

Proof idea The proofs, which can be found in Appendix A.3, are based on three ingredients, a Lipschitz property for the mapping $\Psi \to \frac{1}{N}\sum_{n=1}^{N} \max_{|I|=S} \|\Psi_I^\star y_n\|_1$ for the respective signal model, the concentration of the sum around its expectation for a δ-net covering the space of all admissible dictionaries close to Φ and a triangle inequality argument to show that the finite sample response differences are close to the expected response differences and therefore larger than 0 for all $\varepsilon \gtrsim \varepsilon_{\mu,\rho}$. ■

To see better how the sample complexity behaves we simplify the two theorems to the special case of noiseless exactly S-sparse signals with balanced coefficients for various orders of magnitude of S.
If we have $S = O(1)$, Theorem 4 implies that in order to have a maximum within radius $\tilde{\varepsilon}$ to the original dictionary Φ with probability e^{-Kd} we need $N = O\big(K^3 d\tilde{\varepsilon}^{-2}\big)$ samples. Conversely given N training signals we can expect the distance between generating dictionary and closest local maximum to be of the order $O\big(K^2 N^{-1/2}\big)$.
If we assume a very incoherent dictionary where $\mu = O(d^{-1/2})$ and thus let the sparsity level be of the order $O(\sqrt{d})$ the sample complexity rises to $N = O\big(K^3 d^{3/2}\tilde{\varepsilon}^{-2}\big)$. Taking into account that by (13) the target precision $\tilde{\varepsilon}$ needs to be of order $O\left(S^{-1/2}\right) = O\left(d^{-1/4}\right)$ this means that we need at least $N = O\big(K^3 d^2\big)$ training signals and once this initial level is reached, the error goes to zero at rate $N^{-1/2}$.
For an even lower sparsity level, $S = O(d)$, again assuming a very incoherent dictionary, the sample complexity for target precision $\tilde{\varepsilon}$ implied now by Theorem 5 rises to $N = O\big(K^3 d^2\tilde{\varepsilon}^{-2}\big)$. In this regime, however, we cannot reach arbitrarily small errors by choosing N large enough but only approach the asymptotic precision $\tilde{\varepsilon}_\mu = 16K^2\sqrt{SB}\exp\left(-d/72S\right)$.
Following these promising theoretical results, in the next section we will finally see how theory translates into practice.

6. Experiments

After showing that the optimisation criterion in (4) is locally suitable for dictionary identification, in this section we present an iterative thresholding& K-means type algorithm (ITKM) to actually find the local maxima of (4) and conduct some experiments to illustrate the theoretical results. We recall that given the input signals $Y = (y_1 \dots y_N)$ and a fixed sparsity parameter S we want to solve,

$$\max_{\Psi\in\mathcal{D}} \sum_n \max_{|I|=S} \|\Psi_I^\star y_n\|_1.$$

Using LaGrange multipliers,

$$\frac{\partial}{\partial\psi_k}\left(\sum_n \max_{|I|=S} \|\Psi_I^\star y_n\|_1\right) = \sum_{n:k\in I(\Psi,y_n)} \operatorname{sign}(\langle\psi_k, y_n\rangle) y_n^\star,$$
$$\frac{\partial}{\partial\psi_k}\left(\|\psi_k\|_2^2\right) = 2\psi_k^\star,$$

where $I(\Psi, y_n) := \mathrm{argmax}_{|I|=S} \|\Phi_I^\star y_n\|_1$, we arrive at the following update rule,

$$\psi_k^{new} = \lambda_k \cdot \sum_{n:k\in I(\Psi^{old},y_n)} \mathrm{sign}(\langle \psi_k^{old}, y_n\rangle) y_n, \tag{14}$$

where λ_k is a scaling parameter ensuring that $\|\psi_k^{new}\|_2 = 1$.

In practice, when we do not have an oracle giving us the generating dictionary as initialisation, we also need to safeguard against bad initialisations resulting in a zero-update $\psi_k^{new} = 0$. For example we can choose the zero-updated atom uniformly at random from the unit sphere or from the input signals.

Note that finding the sets $I(\Psi^{old}, y_n)$ corresponds to N thresholding operations while updating according to (14) corresponds to K signed means. Altogether this means that each iteration of ITKM has computational complexity determined by the matrix multiplication $\Psi^\star Y$, meaning $O(dKN)$. This is light in comparison to K-SVD, which even when using tresholding instead of OMP as sparse approximation procedure still requires the calculation of the maximal singular vector of K on average $d \times \frac{N}{K}$ matrices. It is also more computationally efficient than the similarly on averaging based algorithm for local dictionary refinement, proposed in [4]. Furthermore it is straightforward to derive online or parallelised versions of ITKM. In an online version for each newly arriving signal y_n we calculate $I(\Psi^{old}, y_n)$ using thresholding and update $\psi_k^{new} = \psi_k^{new} + \mathrm{sign}(\langle \psi_k^{old}, y_n\rangle) y_n$ for $k \in I(\Psi^{old}, y_n)$. After N signals have been processed we renormalise the atoms ψ_k^{new} to have unit norm and set $\Psi^{old} = \Psi^{new}$. Similarly to parallelise we divide the training samples into m sets of size $\frac{N}{m}$ and on each node m learn a dictionary Ψ_m^{new} according to (14) but with $\lambda_k = 1$. We then calculate the sum of these dictionaries $\Psi_0^{new} = \sum_m \Psi_m^{new}$ and renormalise the atoms in Ψ_0^{new} to have unit norm.

Armed with this very simple algorithm we will now conduct four experiments to illustrate our theoretical findings[1].

6.1 ITKM vs. K-SVD

In our first experiment we compare the local recovery error of ITKM and K-SVD for 3-dimensional bases with increasing condition numbers.

The bases are perturbations of the canonical basis $\Phi = (e_1, e_2, e_3)$ with the vector $v = (1, 1, 1)$, that is $\Phi^t = (e_1^t, e_2^t, e_3^t)$, where $e_i^t = (e_i + tv)/\|(e_i + tv)\|_2$ and t varying from 0 to 0.5 in steps of 0.1, which corresponds to condition numbers $\kappa(\Phi^t)$ varying from 1 to 2.5. We generate $N = 4096$ approximately 1-sparse noiseless signals from the signal model described in Table 1 with $S = 1$, $T = 2$, $\rho = 0$ and $b = 0.1/0.2$ and run both ITKM and K-SVD with 1000 iterations, sparsity parameter $S = 1$ and the true dictionary (basis) as initialisation. Figure 1(a) shows the recovery error $d(\Phi^t, \tilde{\Psi})$ between the original dictionary and the output of the respective algorithm averaged over 10 runs.

As predicted by the theoretical results on the corresponding underlying minimisation principles, the recovery errors of ITKM and K-SVD are roughly the same for Φ^0, which is an orthogonal basis and therefore tight. However, while for ITKM the recovery error stays

[1] A Matlab penknife (mini-toolbox) for playing around with ITKM and reproducing the experiments can be found at `http://homepage.uibk.ac.at/~c7021041/ITKM.zip`

Signal Model

Given the generating dictionary Φ our signal model further depends on four coefficient parameters,

S - the effective sparsity or number of comparatively large coefficients,
b - deciding the decay factor of these sparse coefficients,
T - the total number of non-zero coefficients ($T \geq S$) and
ρ - the noiselevel.

Given these parameters we choose a decay factor c_b uniformly at random in the interval $[1-b, 1]$. We set $c_i = c_b^i/\sqrt{S}$ for $1 \leq i \leq S$ and $c_i = 0$ for $T < i \leq K$. If $T = S$ we renormalise the sequence to have unit norm, while if $T > S$ we choose the vector $(c_{S+1}, \dots, c_T)$ uniformly at random on the sphere of radius R, where R is chosen such that the resulting sequence c has unit norm. We then choose a permutation p and a sign sequence σ uniformly at random and set $y = \Phi c_{p,\sigma}$, respectively $y = (\Phi c_{p,\sigma} + r)/\sqrt{1 + \|r\|_2}$ where r is a Gaussian noise-vector with variance ρ^2 if $\rho > 0$.

Table 1: Signal Model

constantly low over all condition numbers, for K-SVD it increases with increasing condition number or non-tightness.

6.2 Recovery Error and Sample Size

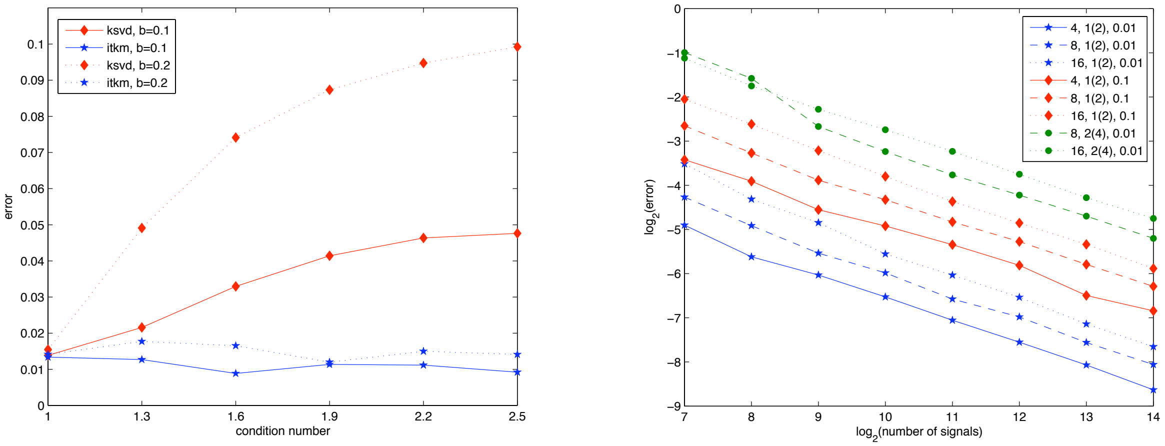
The next experiment is designed to show how fast the maximiser $\tilde{\Psi}$ near the original dictionary Φ converges to Φ with increasing sample size N.

The generating dictionaries consist of the canonical basis in $\mathbb{R}^d$ for $d = 4, 8, 16$ and the first $d/2$ elements of the Hadamard basis and as such are not tight. For every set of parameters $d, S(T), b$ we generate N noiseless signals with N varying from $2^7 = 128$ to $2^{14} = 16384$ and run ITKM with 1000 iterations, sparsity parameter S equal to the coefficient parameter S and the true dictionary as initialisation. Figure 1(b) shows the recovery error $d(\Phi, \tilde{\Psi})$ between the original dictionary Φ and the output of ITKM $\tilde{\Psi}$ averaged over 10 runs.

As predicted by Theorem 4 the recovery error decays as $N^{-1/2}$. However, the separation of the curves for $d = 4, 8, 16$ and almost exactly sparse signals ($b = 0.01$) by a factor around $\sqrt{2}$ instead of 4, as suggested by the estimate $\tilde{\varepsilon} \approx K^2 N^{-1/2}$, indicates that the cubic dependence of the sampling complexity on the number of atoms K may be too pessimistic and could be lowered.

6.3 Stability of Recovery Error under Coherence and Noise

With the last two experiments we illustrate the stability of the maximisation criterion under coherence and noise. As generating dictionaries we use again the canonical basis plus half Hadamard dictionaries described in the last experiment, which have coherence $\mu = d^{-1/2}$.To test the stability under coherence we use a large enough number of noiseless training signals $N = 16384$, such that the distance between the local maximum of the criterion near the generating dictionary, that is the output of ITKM with oracle initialisation, and the generating dictionary is mainly determined by the ratio between the gapsize β and the coherence.



(a)

(b)

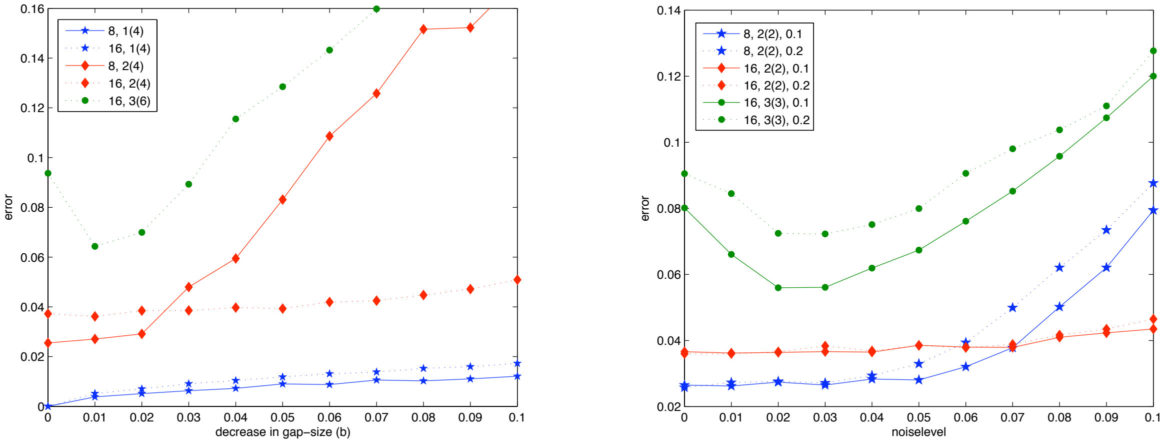
Figure 1: (a) Local recovery error of K-SVD and ITKM for two different types of decaying coefficients and bases with varying condition numbers in $\mathbb{R}^3$, (b) Decay of recovery error of ITKM with increasing number of training signals

For each set of parameters $d, S(T)$ we create N training signals with decreasing gap-sizes β by increasing b from 0 to 0.1 in steps of 0.01 and run ITKM with oracle initialisation, parameter S and 1000 iterations. Figure 2(a) shows the recovery error $d(\Phi, \tilde{\Psi})$ between the original dictionary Φ and the output of ITKM $\tilde{\Psi}$ again averaged over 10 trials.

Again the experiments reflect our theoretical results. For $d = 8, 16$ with $S = 1$ or $d = 16$ with $S = 2$ the gap size is large enough that over the whole range of parameters the recovery error stays constantly low at the level defined by the number of samples. Note that this is quite good, since for $b = 0.1$ we are already far beyond the gap-size coherence ratios where the stable theoretical results hold. On the other hand for $d = 8$ with $S = 2$ or $d = 16$ with $S = 3$ early on the gap decreases enough to become the error determining factor and so we see an increase in recovery error as b grows.

Conversely to test the stability under noise we use a large enough number of exactly sparse training signals, such that the recovery error will be mainly determined by the noiselevel. For each set of parameters $d, S(S), b$ we create $N = 16384$ training signals with Gaussian noise of variance (noiselevel) ρ^2 going from 0 to 0.1 in steps of 0.01 and run ITKM with oracle initialisation, parameter S and 1000 iterations. Figure 2(b) shows the recovery error $d(\Phi, \tilde{\Psi})$ between the original dictionary Φ and the output of ITKM $\tilde{\Psi}$ this time averaged over 20 trials.

The curves again correspond to the prediction of the theoretical results, that is the recovery error stays at roughly the same level defined by the number of samples until the noise becomes large enough and then increases. What is maybe interesting to observe in both experiments is the dithering effect for $d = 16$ with $S = 3$, which is due to the special structure of the dictionary. Indeed using almost equally sized, almost exactly sparse coefficients, it is possible to build signals using only the canonical basis, that have almost the same response in only half the Hadamard basis and the other way round. This indicates that slight perturbations of one with the other lead to even better responses and therefore a larger recovery error. After showing that the theoretical results translate into algorithmic practice, we finally turn to a discussion of our results in the context of existing work and point out directions of future research.



(a) (b)

Figure 2: Increase of recovery error with (a) decreasing ratio between coefficient gap and coherence and (b) increasing noiselevel

7. Discussion

We have introduced a new response maximisation principle for dictionary learning and shown that this is locally suitable to identify a generating μ-coherent dictionary from approximately S-sparse training signals to arbitrary precision as long as the sparsity level is of the order $O(\mu^{-1})$. We have also presented the - to the best of our knowledge - first results showing that stable dictionary identification is locally possible not only for signal to noise ratios of the order $O(\sqrt{d})$ but also for sparsity levels of the order $O(\mu^{-2})$.

The derived sample complexity (omitting log factors) of $O(K^3 d\tilde{\varepsilon}^{-2})$, for signals with sparsity levels $S = O(1)$ is roughly the same as for the K-SVD criterion, [24], or the ℓ_1-minimisation criterion, [16], but somewhat large compared to recently developed dictionary algorithms that have a sample complexity of $O(K^2)$, [4, 1], or $O(K\varepsilon^{-2})$, [2]. However as the sparsity approaches and goes beyond $\mu^{-1} \sim \sqrt{d}$ the derived sample complexity of $O(K^3 d^2 \tilde{\varepsilon}^{-2})$ compares quite favourably to the sample complexity of $O(K^{1/(4\eta)})$ for a sparsity level $d^{1/2-\eta}$ as projected in [4]. Given that also our experimental results suggest that $O(K^3 d\tilde{\varepsilon}^{-2})$ is quite pessimistic, one future direction of research aims to lower the sample complexity. In particular on-going work suggests that for the ITKM *algorithm* a sample size of order K^2 is enough to guarantee local recovery with high probability.

Another strong point of the results is that the corresponding maximisation algorithm ITKM

(Iterative Thresholding& K signed Means) is locally successful, as demonstrated in several experiments, and computationally very efficient. The most complex step in each iteration is the matrix multiplication $\Phi^\star Y$ of order $O(dKN)$, which is even lighter than the iterative averaging algorithm described in [4].
However, the serious drawback is that ITKM is only a local algorithm and that all our results are only local. Also while for the K-SVD criterion and the ℓ_1-minimisation criterion there is reason to believe that all local minima might be equivalent, the response maximisation principle has a lot of smaller local maxima, which is confirmed by preliminary experiments with random initialisations. There ITKM fails but with grace, that is, it outputs local maximisers that have not all, but only most atoms in common with what seems to be the global maximiser near the generating dictionary. This behaviour is in strong contrast to the algorithms presented in [4, 2], that have global success guarantees at a computational cost of the order $O(dN^2)$, and leads to several very important research directions.
First we want to confirm that ITKM has a convergence radius of the order $O(1/\sqrt{S})$. This is suggested by the derived radius of the area on which the generating dictionary is the optimal maximiser as well as preliminary experiments. Alternatively, we could investigate how the results for the local iterative algorithms in [4, 1] could be extended to larger sparsity levels and convergence radii using our techniques. The associated important question is how to extend the results for the algorithms presented in [4, 2] to sparsity levels $O(\mu^{-2})$, if possible at lower cost than $O(dN^2)$. Given the conjectured size of the convergence radius for ITKM it would even be sufficient for the output of the algorithm to arrive at a dictionary within distance $O(1/\sqrt{S})$ to the generating dictionary, since the output could then be used as initialisation for ITKM.
A parallel approach for getting global identification results for sparsity levels $O(\mu^{-2})$, that we are currently pursuing, is to analyse a version of ITKM using residual instead of pure signal means, which in preliminary experiments exhibits global convergence properties.
The last research directions we want to point out are concerned with the realism of the signal model. The fact that for an input sparsity S a gap of order $O(\mu^{-2})$ between the S and $S+1$ largest coefficient is sufficient can be interpreted as a relaxed dependence of the algorithm on the sparsity parameter, since a gap of order μ^{-2} can occur quite frequently. To further decrease this sensitivity to the sparsity parameter in the criterion and the algorithm we would therefore like to extend our results to the case where we can only guarantee a gap of order $O(\mu^{-2})$ between the S largest and the $S+T$ largest coefficient for some $T>1$. Last but not least we would like to exactly reflect the practical situation, where we would normalise our training signals to equally weight their importance, and analyse the unit norm signal model where $y = \Phi x + r/\|\Phi x + r\|_2$.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) under Grant no. J3335 and improved thanks to the reviewers' comments and suggestions.
Thanks also go to the Computer Vision Laboratory of the University of Sassari, Italy, which provided the surroundings, where almost all of the presented work was done.
Finally, big thanks go to Sebastian Kaiser and Felix Krahmer for pointing out the error in

the original proof of Proposition 7, which lead to this correction.

Appendix A. Proofs

A.1 Proof of Theorem 2

We first reformulate and prove the theorem for the simple case of a symmetric coefficient distribution based on one sequence and then use an integration argument to extend it to the general case.

Proposition 6 *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ. Let $x \in \mathbb{R}^K$ be a random permutation of a sequence c, where $c_1 \geq c_2 \geq c_3 \ldots \geq c_K \geq 0$ and $\|c\|_2 = 1$, provided with random $\pm$ signs, that is $x = c_{p,\sigma}$ with probability $\mathbb{P}(p,\sigma) = (2^K K!)^{-1}$. Assume that the signals are generated as $y = \Phi x$. If c satisfies $c_S > c_{S+1} + 2\mu\|c\|_1$ then there is a local maximum of (5) at Φ.*
Moreover for $\Psi \neq \Phi$ we have $\mathbb{E}_y\left(\max_{|I|=S} \|\Psi_I^\star y\|_1\right) < \mathbb{E}_y\left(\max_{|I|=S} \|\Phi_I^\star y\|_1\right)$ as soon as

$$d(\Phi,\Psi) \leq \frac{c_S - c_{S+1} - 2\mu\|c\|_1}{1 + 3\sqrt{\log\left(\frac{25K^2S\sqrt{B}}{(c_S - c_{S+1} - 2\mu\|c\|_1)(c_1+\ldots+c_S)}\right)}}. \tag{15}$$

Proof We start by evaluating the objective function at the original dictionary Φ.

$$\mathbb{E}_y\left(\max_{|I|=S} \|\Phi_I^\star y\|_1\right) = \mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S} \|\Phi_I^\star \Phi c_{p,\sigma}\|_1\right) = \mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S} \sum_{i\in I} |\langle \phi_i, \Phi c_{p,\sigma}\rangle|\right).$$

To estimate the sum of the (in absolute value) largest S inner products, we first assume that p is fixed. Setting $I_p = p^{-1}(\{1,\ldots S\})$ we have,

$$|\langle \phi_i, \Phi c_{p,\sigma}\rangle| = \left|\sigma_i c_{p(i)} + \sum_{j\neq i} \sigma_j c_{p(j)}\langle \phi_i, \phi_j\rangle\right| \begin{array}{ll} \geq c_S - \mu\|c\|_1 & \forall i \in I_p \\ \leq c_{S+1} + \mu\|c\|_1 & \forall i \notin I_p \end{array}.$$

Together with the condition that $c_S > c_{S+1} + 2\mu\|c\|_1$ these estimates ensure that the S maximal inner products in absolute value are attained at I_p and so we get for the expectation,

$$\begin{aligned}\mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S} \|\Phi_I^\star \Phi c_{p,\sigma}\|_1\right) &= \mathbb{E}_p\mathbb{E}_\sigma\left(\|\Phi_{I_p}^\star \Phi c_{p,\sigma}\|_1\right) \\ &= \mathbb{E}_p\mathbb{E}_\sigma\left(\sum_{i\in I_p}\left|c_{p(i)} + \sigma_i \sum_{j\neq i} \sigma_j c_{p(j)}\langle \phi_i, \phi_j\rangle\right|\right) = c_1 + \ldots + c_S.\end{aligned}$$

To compute the expectation for a perturbation of the original dictionary we use the following parametrisation of all ε-perturbations Ψ of the original dictionary Φ. If $d(\Psi,\Phi) = \varepsilon$ then $\|\psi_i - \phi_i\|_2 = \varepsilon_i$ with $\max_i \varepsilon_i = \varepsilon$ and we have z_i with $\langle \phi_i, z_i\rangle = 0$, $\|z_i\|_2 = 1$ and $\alpha_i := 1 - \varepsilon_i^2/2$ and $\omega_i := (\varepsilon_i^2 - \varepsilon_i^4/4)^{\frac{1}{2}}$, such that

$$\psi_i = \alpha_i \phi_i + \omega_i z_i.$$

Expanding the expectation as before we get,

$$\mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right)=\mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S}\|\Psi_I^\star\Phi c_{p,\sigma}\|_1\right)=\mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S}\sum_{i\in I}|\langle\psi_i,\Phi c_{p,\sigma}\rangle|\right). \tag{16}$$

The tried and tested strategy applied now is showing that for small perturbations and most sign patterns σ the maximal inner products are still attained by $i\in I_p$. We have

$$\begin{aligned}\forall i\in I_p:\quad & |\langle\psi_i,\Phi c_{p,\sigma}\rangle|\geq\alpha_i(c_S-\mu\|c\|_1)-\omega_i|\langle z_i,\Phi c_{p,\sigma}\rangle|\\ \forall i\notin I_p:\quad & |\langle\psi_i,\Phi c_{p,\sigma}\rangle|\leq\alpha_i(c_{S+1}+\mu\|c\|_1)+\omega_i|\langle z_i,\Phi c_{p,\sigma}\rangle|.\end{aligned}$$

Using Hoeffding's inequality we can estimate the typical sizes of the terms $|\langle z_i,\Phi c_{p,\sigma}\rangle|$,

$$\begin{aligned}\mathbb{P}(|\langle z_i,\Phi c_{p,\sigma}\rangle|\geq t)&=\mathbb{P}(|\sum_{j\neq i}\sigma_j c_{p(j)}\langle z_i,\phi_j\rangle|>t)\\ &\leq 2\exp\left(-\frac{t^2}{2\sum_{j\neq i}c_{p(j)}^2\langle z_i,\phi_j\rangle^2}\right)\leq 2\exp\left(-\frac{t^2}{2}\right).\end{aligned}$$

In case $\omega_i\neq 0$ or equivalently $\varepsilon_i\neq 0$, we set $t=s/\omega_i$ to arrive at

$$\mathbb{P}(\omega_i|\langle z_i,\Phi c_{p,\sigma}\rangle|\geq s)\leq 2\exp\left(-\frac{s^2}{2\omega_i^2}\right)\leq 2\exp\left(-\frac{s^2}{2\varepsilon_i^2}\right),$$

where we have used that $\omega_i^2=\varepsilon_i^2-\varepsilon_i^4/4\leq\varepsilon_i^2$, while in case $\varepsilon_i=0$ we trivially have that $\mathbb{P}(\omega_i|\langle z_i,\Phi c_{p,\sigma}\rangle|\geq s)=0$. Summarising these findings we see that except with probability

$$\eta:=2\sum_{i:\varepsilon_i\neq 0}\exp\left(-\frac{s^2}{2\varepsilon_i^2}\right)$$

we have,

$$\begin{aligned}\forall i\in I_p:\quad & |\langle\psi_i,\Phi c_{p,\sigma}\rangle|\geq\alpha_i(c_S-\mu\|c\|_1)-s\\ \forall i\notin I_p:\quad & |\langle\psi_i,\Phi c_{p,\sigma}\rangle|\leq\alpha_i(c_{S+1}+\mu\|c\|_1)+s.\end{aligned}$$

This means that as long as $\min_{i\in I_p}\alpha_i(c_S-\mu\|c\|_1)-s\geq\max_{i\notin I_p}\alpha_i(c_{S+1}+\mu\|c\|_1)+s$, which is for instance implied by setting $s:=\frac{1}{2}(c_S-c_{S+1}-2\mu\|c\|_1-\frac{\varepsilon^2}{2})$, we have

$$\max_{|I|=S}\|\Psi_I^\star\Phi c_{p,\sigma}\|_1=\|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1.$$

We now use this result for the calculation of the expectation over σ in (16). For any permutation p we define the set,

$$\Sigma_p:=\bigcup_i\{\sigma\text{ s.t. }\omega_i|\langle z_i,\Phi c_{p,\sigma}\rangle|\geq s\}.$$

We then have

$$\begin{aligned}\mathbb{E}_\sigma\left(\max_{|I|=S}\|\Psi_I^\star\Phi c_{p,\sigma}\|_1\right) &= \sum_{\sigma\in\Sigma_p}\mathbb{P}(\sigma)\cdot\max_{|I|=S}\|\Psi_I^\star\Phi c_{p,\sigma}\|_1 + \sum_{\sigma\notin\Sigma_p}\mathbb{P}(\sigma)\cdot\|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1\\ &= \sum_{\sigma\in\Sigma_p}\mathbb{P}(\sigma)\cdot\left(\max_{|I|=S}\|\Psi_I^\star\Phi c_{p,\sigma}\|_1 - \|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1\right) + \mathbb{E}_\sigma\left(\|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1\right).\end{aligned}\tag{17}$$

To estimate the sum over Σ_p, note that we have the following bounds,

$$|\langle\psi_i,\Phi c_{p,\sigma}\rangle| = |\alpha_i\langle\phi_i,\Phi c_{p,\sigma}\rangle + \omega_i\langle z_i,\Phi c_{p,\sigma}\rangle| \left\{\begin{array}{l}\le (1-\frac{\varepsilon^2}{2})|\langle\phi_i,\Phi c_{p,\sigma}\rangle| + \varepsilon\sqrt{B}\\ \ge (1-\frac{\varepsilon^2}{2})|\langle\phi_i,\Phi c_{p,\sigma}\rangle| - \varepsilon\sqrt{B}\end{array}\right. ,$$

leading to,

$$\begin{aligned}&\max_{|I|=S}\|\Psi_I^\star\Phi c_{p,\sigma}\|_1 \le (1-\frac{\varepsilon^2}{2})\max_{|I|=S}\|\Phi_I^\star\Phi c_{p,\sigma}\|_1 + S\cdot\varepsilon\sqrt{B} = (1-\frac{\varepsilon^2}{2})\|\Phi_{I_p}^\star\Phi c_{p,\sigma}\|_1 + S\cdot\varepsilon\sqrt{B}\\ &\qquad\|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1 \ge (1-\frac{\varepsilon^2}{2})\|\Phi_{I_p}^\star\Phi c_{p,\sigma}\|_1 - S\cdot\varepsilon\sqrt{B}.\end{aligned}$$

Substituting these estimates into (17) we get

$$\begin{aligned}\mathbb{E}_\sigma\left(\max_{|I|=S}\|\Psi_I^\star\Phi c_{p,\sigma}\|_1\right) &\le \sum_{\sigma\in\Sigma_p}\mathbb{P}(\sigma)\cdot 2\varepsilon S\sqrt{B} + \mathbb{E}_\sigma\left(\|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1\right)\\ &\le \eta\cdot 2\varepsilon S\sqrt{B} + \mathbb{E}_\sigma\left(\|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1\right).\end{aligned}$$

Next we calculate $\mathbb{E}_\sigma\left(\|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1\right)$,

$$\begin{aligned}\mathbb{E}_\sigma\left(\|\Psi_{I_p}^\star\Phi c_{p,\sigma}\|_1\right) &= \mathbb{E}_\sigma\left(\sum_{i\in I_p}|\langle\psi_i,\Phi c_{p,\sigma}\rangle|\right)\\ &= \mathbb{E}_\sigma\left(\sum_{i\in I_p}\left|\alpha_i c_{p(i)} + \sigma_i\langle\alpha_i\phi_i+\omega_i z_i,\sum_{j\neq i}\sigma_j c_{p(j)}\phi_j\rangle\right|\right) = \sum_{i\in I_p}\alpha_i c_{p(i)},\end{aligned}\tag{18}$$

where have used that $\varepsilon \le ((1-\frac{\varepsilon^2}{2})c_S - \mu\|c\|_1)/\sqrt{B}$ guarantees the expression within absolute values in (18) to always be positive. Collecting all these results we arrive at the following estimate for the value of the objective function at Ψ.

$$\begin{aligned}\mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right) &= \mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S}\|\Psi_I^\star\Phi c_{p,\sigma}\|_1\right)\\ &\le \mathbb{E}_p\left(4\varepsilon S\sqrt{B}\sum_{i:\varepsilon_i\neq 0}\exp\left(-\frac{(c_S-c_{S+1}-2\mu\|c\|_1-\frac{\varepsilon^2}{2})^2}{8\varepsilon_i^2}\right) + \sum_{i\in I_p}\alpha_i c_{p(i)}\right)\\ &\le 4\varepsilon S\sqrt{B}\sum_{i:\varepsilon_i\neq 0}\exp\left(-\frac{(c_S-c_{S+1}-2\mu\|c\|_1-\frac{\varepsilon^2}{2})^2}{8\varepsilon_i^2}\right) + \frac{c_1+\ldots+c_S}{K}\sum_i\alpha_i.\end{aligned}$$

Finally we are able to compare the expectation at the original dictionary to that at an ε-perturbation. Remembering that $\alpha_i = 1 - \frac{\varepsilon_i^2}{2}$, we get

$$\begin{aligned}
&\mathbb{E}_y\left(\max_{|I|=S}\|\Phi_I^\star y\|_1\right) - \mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right)\\
&\qquad\geq \frac{c_1+\ldots+c_S}{K}\sum_i \frac{\varepsilon_i^2}{2} - 4\varepsilon S\sqrt{B}\sum_{i:\varepsilon_i\neq 0}\exp\left(-\frac{(c_S-c_{S+1}-2\mu\|c\|_1-\frac{\varepsilon^2}{2})^2}{8\varepsilon_i^2}\right)\\
&\qquad\geq \varepsilon^2\frac{c_1+\ldots+c_S}{2K} - 4\varepsilon SK\sqrt{B}\exp\left(-\frac{(c_S-c_{S+1}-2\mu\|c\|_1-\frac{\varepsilon^2}{2})^2}{8\varepsilon^2}\right).
\end{aligned}$$

Thus to have a local maximum at the original dictionary we need that

$$\varepsilon > \frac{8SK^2\sqrt{B}}{c_1+\ldots+c_S}\exp\left(-\frac{(c_S-c_{S+1}-2\mu\|c\|_1-\frac{\varepsilon^2}{2})^2}{8\varepsilon^2}\right),$$

and all that remains to be shown is that this is implied by (15). Since $K \geq 2$, (15) implies that $\frac{\varepsilon^2}{2} < \frac{c_S-c_{S+1}-2\mu\|c\|_1}{2(1+3\sqrt{\log 96})^2} \leq \frac{c_S-c_{S+1}-2\mu\|c\|_1}{100}$ and it suffices to show that (15) further implies

$$\varepsilon > \frac{8SK^2\sqrt{B}}{c_1+\ldots+c_S}\exp\left(-\frac{(c_S-c_{S+1}-2\mu\|c\|_1)^2\cdot 99^2}{8\varepsilon^2\cdot 100^2}\right). \tag{19}$$

Applying Lemma A.3 from [24], which says that for $a, b, \varepsilon > 0$,

$$\varepsilon \leq \frac{4b}{1+\sqrt{1+16\log(\frac{a}{b})}} \quad \text{implies that} \quad a\exp\left(\frac{-b^2}{\varepsilon^2}\right) < \varepsilon,$$

to the situation at hand, where $a = \frac{8SK^2\sqrt{B}}{c_1+\ldots+c_S}$ and $b = \frac{(c_S-c_{S+1}-2\mu\|c\|_1)\cdot 99}{\sqrt{8}\cdot 100}$, we get that (19) is ensured by

$$\varepsilon < \frac{c_S-c_{S+1}-2\mu\|c\|_1}{\sqrt{8}\cdot\frac{25}{99}\left(1+\sqrt{16\log\left(\frac{8\sqrt{8}\cdot\frac{100}{99}e^{1/16}SK^2\sqrt{B}}{(c_S-c_{S+1}-2\mu\|c\|_1)(c_1+\ldots+c_S)}\right)}\right)},$$

which simplifies to (15). ■

Proof [of Theorem 2]
Using the symmetry of ν, our strategy is to reduce the general to the simple coefficient model. Let c denote the mapping that assigns to each $x \in S^{K-1}$ the non increasing rearrangement of the absolute values of its components, that is $c_i(x) = |x_{p(i)}|$ for a permutation p such that $c_1(x) \geq c_2(x) \geq \ldots \geq c_K(x) \geq 0$. Then the mapping c together with the probability measure ν on S^{K-1} induces a pull-back probability measure ν_c on $c(S^{K-1})$, by $\nu_c(\Omega) := \nu(c^{-1}(\Omega))$ for any measurable set $\Omega \subseteq c(S^{K-1})$. With the help of this new measure

we can rewrite the expectations we need to calculate as,

$$\begin{aligned}\mathbb{E}_y\left(\max_{|I|=S}\|\Phi_I^\star y\|_1\right) &= \mathbb{E}_x\left(\max_{|I|=S}\|\Phi_I^\star\Phi x\|_1\right)\\ &= \int_x \max_{|I|=S}\|\Phi_I^\star\Phi x\|_1 d\nu = \int_{c(x)}\mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S}\|\Phi^\star\Phi c_{p,\sigma}(x)\|_1\right)d\nu_c.\end{aligned}$$

The expectation inside the integral should seem familiar. Indeed we have calculated it already in the proof of Proposition 6 for $c(x)$ a fixed decaying sequence satisfying $c_S(x) > c_{S+1}(x) + 2\mu\|x\|_1$. Since this property is satisfied almost surely we have,

$$\begin{aligned}\mathbb{E}_y\left(\max_{|I|=S}\|\Phi_I^\star y\|_1\right) &= \int_{c(x)}\mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S}\|\Phi^\star\Phi c_{p,\sigma}(x)\|_1\right)d\nu_c\\ &= \int_{c(x)} c_1(x)+\ldots+c_S(x)d\nu_c := \bar{c}_1+\ldots+\bar{c}_S.\end{aligned}$$

For the expectation of a perturbed dictionary Ψ we get in analogy

$$\begin{aligned}\mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right) &= \int_{c(x)}\mathbb{E}_p\mathbb{E}_\sigma\left(\max_{|I|=S}\|\Psi^\star\Phi c_{p,\sigma}(x)\|_1\right)d\nu_c\\ &\leq \int_{c(x)}\eta(x)+(c_1(x)+\ldots+c_S(x))\frac{1}{K}\sum_i\alpha_i\, d\nu_c,\end{aligned}$$

where

$$\eta(x) := 4\varepsilon S\sqrt{B}\sum_{i:\varepsilon_i\neq 0}\exp\left(-\frac{(c_S(x)-c_{S+1}(x)-2\mu\|x\|_1-\frac{\varepsilon^2}{2})^2}{8\varepsilon_i^2}\right).$$

Since $c_S(x)-c_{S+1}(x)-2\mu\|x\|_1 \geq \beta$ almost surely we have

$$\eta(x)\leq 4\varepsilon S\sqrt{B}\sum_{i:\varepsilon_i\neq 0}\exp\left(-\frac{(\beta-\frac{\varepsilon^2}{2})^2}{8\varepsilon_i^2}\right) := \eta_\beta,$$

almost surely and therefore

$$\mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right)\leq \eta_\beta+(\bar{c}_1+\ldots+\bar{c}_S)\frac{1}{K}\sum_i\alpha_i.$$

Following the same argument as in the proof of Proposition 6 we see that $\mathbb{E}_y\left(\max_{|I|=S}\|\Phi_I^\star y\|_1\right) > \mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right)$ as soon as

$$\varepsilon < \frac{\beta}{1+3\sqrt{\log\left(\frac{25K^2S\sqrt{B}}{\beta(\bar{c}_1+\ldots+\bar{c}_S)}\right)}}.$$

■

A.2 Proof of Theorem 3

Again we first reformulate and prove the theorem for the case of a symmetric coefficient distribution based on one sequence and then extend it with an integration argument.

Proposition 7 *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ. Let $x \in \mathbb{R}^K$ be a random permutation of a sequence c, where $c_1 \geq c_2 \geq c_3 \ldots \geq c_K \geq 0$ and $\|c\|_2 = 1$, provided with random $\pm$ signs, that is $x = c_{p,\sigma}$ with probability $\mathbb{P}(p,\sigma) = (2^K K!)^{-1}$. Further let $r = (r(1) \ldots r(d))$ be a centred random subgaussian noise-vector with parameter ρ and assume that the signals are generated according to the noisy signal model in* (9). *If we have*

$$\max\{\mu, \rho\} \leq \frac{c_S - c_{S+1}}{\sqrt{72(\log a + \log\log a)}} \quad \textit{for} \quad a = \frac{112K^2S(\sqrt{B}+1)}{C_r(c_S - c_{S+1})(c_1 + \ldots + c_S)}, \tag{20}$$

where $C_r = \mathbb{E}_r\left((1+\|r\|_2^2)^{-1/2}\right)$, then there is a local maximum of (5) *at $\tilde{\Psi}$ satisfying,*

$$d(\tilde{\Psi}, \Phi) \leq \frac{12SK^2\sqrt{B}}{C_r(c_1 + \ldots + c_S)} \exp\left(\frac{-(c_S - c_{S+1})^2}{72\max\{\mu^2, \rho^2\}}\right).$$

Proof To prove the proposition we digress from the conventional scheme of first calculating the expectation of our objective function for both the original and a perturbed dictionary and then comparing and instead bound the difference of the expectations directly.

$$\begin{aligned}
&\mathbb{E}_y\left(\max_{|I|=S}\|\Phi_I^\star y\|_1\right) - \mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right) \\
&\quad = \mathbb{E}_{p,\sigma,r}\left(\max_{|I|=S}\left\|\frac{\Phi_I^\star(\Phi c_{p,\sigma} + r)}{\sqrt{1+\|r\|_2^2}}\right\|_1 - \max_{|I|=S}\left\|\frac{\Psi_I^\star(\Phi c_{p,\sigma} + r)}{\sqrt{1+\|r\|_2^2}}\right\|_1\right) \\
&\quad = \mathbb{E}_{p,\sigma,r}\left(\frac{\max_{|I|=S}\|\Phi_I^\star(\Phi c_{p,\sigma}+r)\|_1 - \max_{|I|=S}\|\Psi_I^\star(\Phi c_{p,\sigma}+r)\|_1}{\sqrt{1+\|r\|_2^2}}\right) := \mathbb{E}_{p,\sigma,r}(\Delta_{p,\sigma,r})
\end{aligned}$$

Again our strategy is to show that for a fixed p for most σ and r the maximal response of both the original dictionary and the perturbation is attained at I_p. The expressions we therefore need to lower (upper) bound for $i \in I_p$ ($i \notin I_p$) are

$$\begin{aligned}
|\langle \phi_i, \Phi c_{p,\sigma} + r\rangle| &= \big|\sigma_i c_{p(i)} + \sum_{j\neq i}\sigma_j c_{p(j)}\langle\phi_i,\phi_j\rangle + \langle\phi_i, r\rangle\big|, \\
&= \big|c_{p(i)} + \sigma_i\sum_{j\neq i}\sigma_j c_{p(j)}\langle\phi_i,\phi_j\rangle + \sigma_i\langle\phi_i, r\rangle\big|, \\
|\langle \psi_i, \Phi c_{p,\sigma} + r\rangle| &= \big|\alpha_i\sigma_i c_{p(i)} + \alpha_i\sum_{j\neq i}\sigma_j c_{p(j)}\langle\phi_i,\phi_j\rangle + \omega_i\langle z_i, \Phi c_{p,\sigma}\rangle + \langle\psi_i, r\rangle\big| \\
&= \big|\alpha_i c_{p(i)} + \sigma_i\alpha_i\sum_{j\neq i}\sigma_j c_{p(j)}\langle\phi_i,\phi_j\rangle + \sigma_i\omega_i\langle z_i, \Phi c_{p,\sigma}\rangle + \sigma_i\langle\psi_i, r\rangle\big|.
\end{aligned}$$

However, instead of using a worst case estimate for the gap between the responses of the original dictionary within and without I_p, we now make use of the fact that for most sign sequences we have a gap size of order $c_S - c_{S+1}$. This means that as soon as $|\sum_{j\neq i}\sigma_j c_{p(j)}\langle \phi_i,\phi_j\rangle|$, $\omega_i|\langle z_i, \Phi c_{p,\sigma}\rangle|$ and the noise related terms $|\langle \phi_i, r\rangle|$ and $|\langle \psi_i, r\rangle|$ are of order $(c_S - c_{S+1})$ the maximal response of both the original dictionary and the perturbation is attained at I_p. In particular, setting $\delta_p(i) = -1$ for $i \in I_p$ and $\delta_p(i) = 1$ for $i \notin I_p$ defining the sets,

$$\Sigma_p := \bigcup_i \left\{ \sigma \text{ s.t. } \sigma_i \delta_p(i) \sum_{j\neq i} \sigma_j c_{p(j)} \langle \phi_i, \phi_j\rangle \geq \frac{c_S - c_{S+1}}{6} \right.$$
$$\left. \text{or } \sigma_i\delta_p(i)\omega_i\langle z_i, \Phi c_{p,\sigma}\rangle \geq \frac{c_S - c_{S+1} - \frac{3\varepsilon^2}{2}}{6} \right\},$$

for a fixed permutation p and

$$R := \bigcup_i \left\{ r \text{ s.t. } |\langle \phi_i, r\rangle| \geq \frac{c_S - c_{S+1}}{3} \text{ or } |\langle \psi_i, r\rangle| \geq \frac{c_S - c_{S+1}}{6} \right\},$$

we see that both maxima are attained at I_p as long as $\sigma \notin \Sigma_p$ and $r \notin R$. Using Hoeffding's inequality we get that

$$\mathbb{P}\left(\sigma_i\delta_p(i)\sum_{j\neq i}\sigma_j c_{p(j)}\langle\phi_i,\phi_j\rangle > t\right) \leq \exp\left(\frac{-t^2}{2\sum_{j\neq i} c^2_{p(j)}|\langle \phi_i,\phi_j\rangle|^2}\right) \leq \exp\left(\frac{-t^2}{2\mu^2}\right),$$

and similarly (compare also Proposition 6) we get that for $\varepsilon_i \neq 0$

$$\mathbb{P}(\sigma_i\delta_p(i)\omega_i\langle z_i, \Phi c_{p,\sigma}\rangle \geq s) \leq \exp\left(\frac{-s^2}{2\varepsilon_i^2}\right).$$

Setting $t = (c_S - c_{S+1})/6$, $s = (c_S - c_{S+1} - \frac{3\varepsilon^2}{2})/6$ and using a union bound then leads to

$$\mathbb{P}(\Sigma_p) \leq K\exp\left(-\frac{\left(c_S - c_{S+1} - \frac{3\varepsilon^2}{2}\right)^2}{72\varepsilon^2}\right) + K\exp\left(\frac{-(c_S - c_{S+1})^2}{72\mu^2}\right). \tag{21}$$

Since the $r(i)$ are subgaussian with parameter ρ we have for any $v = (v_1 \dots v_d)$ and $t \geq 0$, $\mathbb{P}(|\langle v, r\rangle| \geq t) \leq \exp\left(-\frac{t^2}{2\rho^2\|v\|_2^2}\right)$, see e.g. [31]. Taking a union bound over all ϕ_i, ψ_i with the corresponding choice for t then leads to the estimate

$$\mathbb{P}(R) \leq 2K\exp\left(-\frac{\left(c_S - c_{S+1}\right)^2}{72\rho^2}\right) + 2K\exp\left(-\frac{\left(c_S - c_{S+1}\right)^2}{18\rho^2}\right). \tag{22}$$

We now split the expectations over the sign and noise patterns for a fixed p to get

$$\begin{aligned}\mathbb{E}_{\sigma,r}(\Delta_{p,\sigma,r}) &= \int_{r\notin R}\mathbb{E}_\sigma\left(\Delta_{p,\sigma,r}\right)d\nu_r + \int_{r\in R}\mathbb{E}_\sigma\left(\Delta_{p,\sigma,r}\right)d\nu_r \\ &= \int_{r\notin R}\Big(\sum_{\sigma\notin\Sigma_p}\mathbb{P}(\sigma)\Delta_{p,\sigma,r}\Big)d\nu_r + \int_{r\notin R}\Big(\sum_{\sigma\in\Sigma_p}\mathbb{P}(\sigma)\Delta_{p,\sigma,r}\Big)d\nu_r \\ &\qquad + \mathbb{E}_\sigma\left(\int_{r\in R}\Delta_{p,\sigma,r}d\nu_r\right).\end{aligned} \tag{23}$$

We start by bounding $\sum_{\sigma\notin\Sigma_p}\Delta_{p,\sigma,r}$ for a fixed $r\notin R$. We have

$$\begin{aligned}
\sum_{\sigma\notin\Sigma_p}\Delta_{p,\sigma,r} &= \sum_{\sigma\notin\Sigma_p}\mathbb{P}(\sigma)\left(\frac{\left\|\Phi^\star_{I_p}(\Phi c_{p,\sigma}+r)\right\|_1-\left\|\Psi^\star_{I_p}(\Phi c_{p,\sigma}+r)\right\|_1}{\sqrt{1+\|r\|_2^2}}\right)\\
&=(1+\|r\|_2^2)^{-\frac{1}{2}}\sum_{\sigma\notin\Sigma_p}\sum_{i\in I_p}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-|\langle\psi_i,\Phi c_{p,\sigma}+r\rangle|\\
&=(1+\|r\|_2^2)^{-1/2}\sum_{i\in I_p}\sum_{\sigma\notin\Sigma_p}\sigma_i\langle\phi_i,\Phi c_{p,\sigma}+r\rangle-\sigma_i\langle\psi_i,\Phi c_{p,\sigma}+r\rangle\\
&=(1+\|r\|_2^2)^{-\frac{1}{2}}\sum_{i\in I_p}\sum_{\sigma\notin\Sigma_p}(1-\alpha_i)|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\sigma_i\omega_i\langle z_i,\Phi c_{p,\sigma}+r\rangle.
\end{aligned}$$

Define σ^i via setting $\sigma^i_i=-\sigma_i$ and $\sigma^i_j=\sigma_j$ for $j\neq i$. For every index $i\in I_p$ there are two types of sequences $\sigma\notin\Sigma_p$, those were $\sigma^i\notin\Sigma_p$ and those were $\sigma^i\in\Sigma_p$. Since $\omega_i\langle z_i,\Phi c_{p,\sigma}+r\rangle$ is independent of σ_i, in the sum over the first type of sequences the term is scaled once with σ_i and once with $\sigma^i_i=-\sigma_i$ and therefore cancels out. This leads to

$$\begin{aligned}
\sum_{\sigma\notin\Sigma_p}\Delta_{p,\sigma,r} &= (1+\|r\|_2^2)^{-\frac{1}{2}}\sum_{i\in I_p}\left(\sum_{\sigma\notin\Sigma_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\sum_{\sigma\notin\Sigma_p:\sigma^i\in\Sigma_p}\sigma_i\omega_i\langle z_i,\Phi c_{p,\sigma}+r\rangle\right)\\
&\geq(1+\|r\|_2^2)^{-\frac{1}{2}}\sum_{i\in I_p}\left(\sum_{\sigma\notin\Sigma_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\sum_{\sigma\notin\Sigma_p:\sigma^i\in\Sigma_p}\varepsilon_i\|\Phi c_{p,\sigma}+r\|_2\right)\\
&\geq(1+\|r\|_2^2)^{-\frac{1}{2}}\sum_{i\in I_p}\left(\sum_{\sigma\notin\Sigma_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\sum_{\sigma\in\Sigma_p}\varepsilon(\sqrt{B}+\|r\|_2)\right)\\
&\geq(1+\|r\|_2^2)^{-\frac{1}{2}}\left(\sum_{\sigma\notin\Sigma_p}\sum_{i\in I_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\sum_{\sigma\in\Sigma_p}\varepsilon S(\sqrt{B}+\|r\|_2)\right). \qquad (24)
\end{aligned}$$

Before substituting this bound into (23) we develop a general bound on $\Delta_{p,\sigma,r}$. We have

$$\begin{aligned}
\max_{|I|=S}\|\Psi^\star_I(\Phi c_{p,\sigma}+r)\|_1 &= \max_{|I|=S}\sum_{i\in I}|\langle\alpha_i\phi_i+\omega_i z_i,\Phi c_{p,\sigma}+r)\rangle|\\
&\leq\max_{|I|=S}\sum_{i\in I}\big(1-\frac{\varepsilon_i^2}{2}\big)|\langle\phi_i,\Phi c_{p,\sigma}+r)\rangle|+\varepsilon_i\|\Phi c_{p,\sigma}+r\|_2\\
&\leq\max_{|I|=S}\sum_{i\in I}\big(1-\frac{\varepsilon^2}{2}\big)|\langle\phi_i,\Phi c_{p,\sigma}+r)\rangle|+\varepsilon\big(\sqrt{B}+\|r\|_2\big)\\
&=\big(1-\frac{\varepsilon^2}{2}\big)\max_{|I|=S}\|\Phi^\star_I(\Phi c_{p,\sigma}+r)\|_1+\varepsilon S\big(\sqrt{B}+\|r\|_2\big),
\end{aligned}$$

which immediately leads to the lower bound

$$\begin{aligned}\Delta_{p,\sigma,r} &\geq (1+\|r\|_2^2)^{-\frac{1}{2}}\left(\max_{|I|=S}\|\Phi_I^\star(\Phi c_{p,\sigma}+r)\|_1\frac{\varepsilon^2}{2}-\varepsilon S\big(\sqrt{B}+\|r\|_2\big)\right)\\ &\geq (1+\|r\|_2^2)^{-\frac{1}{2}}\left(\|\Phi_{I_p}^\star(\Phi c_{p,\sigma}+r)\|_1\frac{\varepsilon^2}{2}-\varepsilon S\big(\sqrt{B}+\|r\|_2\big)\right)\\ &\geq (1+\|r\|_2^2)^{-\frac{1}{2}}\left(\sum_{i\in I_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\varepsilon S\big(\sqrt{B}+\|r\|_2\big)\right).\end{aligned}$$

Substituting the estimate above together with (24) into (23) we get

$$\begin{aligned}\mathbb{E}_{\sigma,r}(\Delta_{p,\sigma,r}) \geq &\int_{r\notin R}\frac{\mathbb{P}(\sigma)}{\sqrt{1+\|r\|_2^2}}\left(\sum_{\sigma\notin\Sigma_p}\sum_{i\in I_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\sum_{\sigma\in\Sigma_p}\varepsilon S(\sqrt{B}+\|r\|_2)\right)d\nu_r\\ &+\int_{r\notin R}\frac{\mathbb{P}(\sigma)}{\sqrt{1+\|r\|_2^2}}\sum_{\sigma\in\Sigma_p}\left(\sum_{i\in I_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\varepsilon S\big(\sqrt{B}+\|r\|_2\big)\right)d\nu_r\\ &+\mathbb{E}_\sigma\left(\int_{r\in R}\frac{1}{\sqrt{1+\|r\|_2^2}}\left(\sum_{i\in I_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|-\varepsilon S\big(\sqrt{B}+\|r\|_2\big)\right)d\nu_r\right),\end{aligned}$$

which after collecting all the terms, we can further bound as

$$\begin{aligned}\mathbb{E}_{\sigma,r}(\Delta_{p,\sigma,r}) \geq \mathbb{E}_{\sigma,r}&\left(\frac{\sum_{i\in I_p}\frac{\varepsilon_i^2}{2}|\langle\phi_i,\Phi c_{p,\sigma}+r\rangle|}{\sqrt{1+\|r\|_2^2}}\right)\\ &\qquad -2\mathbb{P}(\Sigma_p)\int_{r\notin R}\frac{\varepsilon S\big(\sqrt{B}+\|r\|_2\big)}{\sqrt{1+\|r\|_2^2}}d\nu_r-\int_{r\in R}\frac{\varepsilon S\big(\sqrt{B}+\|r\|_2\big)}{\sqrt{1+\|r\|_2^2}}d\nu_r\\ \geq \mathbb{E}_r&\left(\frac{\sum_{i\in I_p}c_{p(i)}\frac{\varepsilon_i^2}{2}}{\sqrt{1+\|r\|_2^2}}\right)-\varepsilon S\big(\sqrt{B}+1\big)\cdot\big(2\mathbb{P}(\Sigma_p)+\mathbb{P}(R)\big). \qquad (25)\end{aligned}$$

Taking the expectation over the permutations then yields

$$\begin{aligned}\mathbb{E}_{p,\sigma,r}(\Delta_{p,\sigma,r}) &\geq \mathbb{E}_r\mathbb{E}_p\left(\frac{\sum_{i\in I_p}c_{p(i)}\frac{\varepsilon_i^2}{2}}{\sqrt{1+\|r\|_2^2}}\right)-\varepsilon S\big(\sqrt{B}+1\big)\cdot\big(2\mathbb{E}_p\mathbb{P}(\Sigma_p)+\mathbb{P}(R)\big)\\ &\geq \mathbb{E}_r\left(\frac{1}{\sqrt{1+\|r\|_2^2}}\right)\frac{c_1+\ldots+c_S}{2K}\sum_i\varepsilon_i^2-\varepsilon S\big(\sqrt{B}+1\big)\cdot\big(2\mathbb{E}_p\mathbb{P}(\Sigma_p)+\mathbb{P}(R)\big).\end{aligned}$$

Using the probability estimates from (21)/(22) we see that $\mathbb{E}_{p,\sigma,r}(\Delta_{p,\sigma,r})>0$ is implied by

$$\varepsilon \geq \frac{4SK^2\big(\sqrt{B}+1\big)}{C_r\gamma}\left(\exp\left(\frac{-\big(\beta-\frac{3\varepsilon^2}{2}\big)^2}{72\varepsilon^2}\right)+\exp\left(\frac{-\beta^2}{72\mu^2}\right)+\exp\left(\frac{-\beta^2}{72\rho^2}\right)+\exp\left(\frac{-\beta^2}{18\rho^2}\right)\right),$$

where we have used the abbreviations $\gamma = c_1 + \ldots + c_S$, $\beta = c_S - c_{S+1}$ and $C_r = \mathbb{E}_r\left((1+\|r\|_2^2)^{-1/2}\right)$. We now proceed by splitting the above condition. We define $\varepsilon_{\min}$ by asking that

$$\frac{\varepsilon}{3} \geq \frac{4SK^2\big(\sqrt{B}+1\big)}{C_r\gamma}\exp\left(-\frac{\beta^2}{72\max\{\mu^2,\rho^2\}}\right) := \frac{\varepsilon_{\min}}{3}$$

and $\varepsilon_{\max}$ implicitely by asking that

$$\frac{\varepsilon}{3} - \frac{\varepsilon^4}{81} \geq \frac{4SK^2\big(\sqrt{B}+1\big)}{C_r\gamma}\exp\left(-\frac{\big(\beta-\frac{3\varepsilon^2}{2}\big)^2}{72\varepsilon^2}\right).$$

Following the line of argument in the proof of Proposition 6 we see that the above condition is guaranteed as soon as

$$\varepsilon \leq \frac{\beta}{\frac{5}{2}+9\sqrt{\log\left(\frac{112K^2S(\sqrt{B}+1)}{C_r\beta\gamma}\right)}} := \varepsilon_{\max}.$$

The statement follows from making sure that $\varepsilon_{\min} < \varepsilon_{\max}$. ■

Proof [of Theorem 3] Using the pull-back probability measure ν_c we can write

$$\mathbb{E}_y\left(\max_{|I|=S}\|\Phi_I^\star y\|_1\right) - \mathbb{E}_y\left(\max_{|I|=S}\|\Psi_I^\star y\|_1\right) = \int_{c(x)}\mathbb{E}_{p,\sigma,r}\left(\Delta_{p,\sigma,r,c(x)}\right)d\nu_c,$$

where $\Delta_{p,\sigma,r,c(x)}$ is defined analogue to $\Delta_{p,\sigma,r}$ in the last proof, that is replacing c by $c(x)$. The statement follows from employing the lower estimate for $\mathbb{E}_{p,\sigma,r}\left(\Delta_{p,\sigma,r,c(x)}\right)$ from (25) and replacing $c_1+\ldots+c_S$ by $\bar{c}_1+\ldots+\bar{c}_S$ resp. $c_S - c_{S+1}$ by its lower bound β in the proof of Proposition 7. ■

A.3 Proof of Theorems 4 and 5

Since the proofs of Theorems 4 and 5 are conceptually equivalent we will combine them into one and just split the argument for the inevitable juggling of constants.

Proof As outlined in the proof idea we need a Lipschitz property for the mapping $\Psi \to \frac{1}{N}\sum_{n=1}^N \max_{|I|=S}\|\Psi_I^\star y_n\|_1$ for both signal models, the concentration of the sum around its expectation for a δ net covering the space of all admissible dictionaries close to Φ and a triangle inequality argument to get to the final statement.

To show the Lipschitz property we use a reverse triangle inequality,

$$\begin{aligned}\left|\max_{|I|=S}\|\Psi_I^\star y_n\|_1 - \max_{|I|=S}\|\bar{\Psi}_I^\star y_n\|_1\right| &= \left|\max_{|I|=S}\|\bar{\Psi}_I^\star y_n - (\bar{\Psi}_I^\star - \Psi_I^\star)y_n\|_1 - \max_{|I|=S}\|\bar{\Psi}_I^\star y_n\|_1\right|\\ &\leq \max_{|I|=S}\|(\bar{\Psi}_I^\star - \Psi_I^\star)y_n\|_1\\ &\leq S\max_k\|\psi_k - \bar{\psi}_k\|_2\|y_n\|_2\\ &\leq d(\Psi,\bar{\Psi})S\big(\sqrt{B}+1\big).\end{aligned}$$

Note that for the noise-free signal model we can replace $\big(\sqrt{B}+1\big)$ by $\sqrt{B}$ in the last expression. By averaging over n we get that the mapping in question is Lipschitz with constant $S\big(\sqrt{B}+1\big)$ in the noisy and $S\sqrt{B}$ in the noise-free case, that is

$$\left|\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\Psi_I^\star y_n\|_1-\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\bar{\Psi}_I^\star y_n\|_1\right|\leq d(\Psi,\bar{\Psi})S\big(\sqrt{B}+1\big).$$

To show that the averaged sums concentrate around their expectations we use our favourite tool Hoeffding's inequality. Set $X_n=\max_{|I|=S}\|\Phi_I^\star y_n\|_1-\max_{|I|=S}\|\Psi_I^\star y_n\|_1$, then we have $|X_n|\leq\varepsilon S\big(\sqrt{B}+1\big)$, resp. $|X_n|\leq\varepsilon S\sqrt{B}$ in the noise-free case, and get the estimate,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N}\left(\max_{|I|=S}\|\Phi_I^\star y_n\|_1-\max_{|I|=S}\|\Psi_I^\star y_n\|_1\right)-\mathbb{E}\left(\max_{|I|=S}\|\Phi_I^\star y_1\|_1-\max_{|I|=S}\|\Psi_I^\star y_1\|_1\right)\right|\geq 2t\right)$$
$$\leq 2\exp\left(\frac{-2Nt^2}{\varepsilon^2S^2\big(\sqrt{B}+1\big)^2}\right).$$

Next we need to choose a δ-net for all perturbations Ψ with $d(\Phi,\Psi)\leq\varepsilon_{\max}$, that is a finite set of perturbations $\mathcal{N}$ such that for every Ψ we can find $\bar{\Psi}\in\mathcal{N}$ with $d(\Psi,\bar{\Psi})\leq\delta$. Recalling the parametrisation of all ε-perturbations from the proof of Proposition 6, we see that the space we need to cover is included in the product of K balls with radius $\varepsilon_{\max}$ in dimension d. From e.g. Lemma 2 in [31] we know that for the d dimensional ball of radius $\varepsilon_{\max}$ we can find a δ-net $\mathcal{N}_d$ satisfying $\sharp\mathcal{N}_d\leq\big(\varepsilon_{\max}+\frac{2\varepsilon_{\max}}{\delta}\big)^d$, so for our space of ε-perturbations we can find a δ-net $\mathcal{N}$ satisfying,

$$\sharp\mathcal{N}\leq\left(\varepsilon_{\max}+\frac{2\varepsilon_{\max}}{\delta}\right)^{Kd}\leq\left(\frac{3\varepsilon_{\max}}{\delta}\right)^{Kd}.$$

Taking a union bound we can now estimate the probability that we have concentration for all perturbations in the net as,

$$\mathbb{P}\left(\exists\Psi\in\mathcal{N}:\left|\frac{1}{N}\sum_{n=1}^{N}\left(\max_{|I|=S}\|\Phi_I^\star y_n\|_1-\max_{|I|=S}\|\Psi_I^\star y_n\|_1\right)\right.\right.$$
$$\left.\left.-\,\mathbb{E}\left(\max_{|I|=S}\|\Phi_I^\star y_1\|_1-\max_{|I|=S}\|\Psi_I^\star y_1\|_1\right)\right|\geq 2t\right)$$
$$\leq\left(\frac{3\varepsilon_{\max}}{\delta}\right)^{Kd}2\exp\left(\frac{-2Nt^2}{\varepsilon_{\max}^2S^2\big(\sqrt{B}+1\big)^2}\right).$$

Finally we are ready for the triangle inequality argument. For any Ψ with $d(\Psi,\Phi)=\varepsilon\leq\varepsilon_{\max}$ we can find $\bar\Psi\in\mathcal{N}$ with $d(\bar\Psi,\Psi)\leq\delta$ and $d(\Phi,\bar\Psi)=\bar\varepsilon$ and therefore get,

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\Phi_I^\star y_n\|_1-\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\Psi_I^\star y_n\|_1\\
&=\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\Phi_I^\star y_n\|_1-\mathbb{E}\left(\max_{|I|=S}\|\Phi_I^\star y_1\|_1\right)+\mathbb{E}\left(\max_{|I|=S}\|\Phi_I^\star y_1\|_1\right)-\mathbb{E}\left(\max_{|I|=S}\|\bar\Psi_I^\star y_1\|_1\right)\\
&+\mathbb{E}\left(\max_{|I|=S}\|\bar\Psi_I^\star y_1\|_1\right)-\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\bar\Psi_I^\star y_n\|_1+\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\bar\Psi_I^\star y_n\|_1-\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\Psi_I^\star y_n\|_1\\
&\qquad\geq\mathbb{E}\left(\max_{|I|=S}\|\Phi_I^\star y_1\|_1\right)-\mathbb{E}\left(\max_{|I|=S}\|\bar\Psi_I^\star y_1\|_1\right)-2t-\delta S\big(\sqrt{B}+1\big).
\end{aligned}$$

Depending on the signal model we now have to substitute the values for the asymptotic differences $\mathbb{E}\left(\max_{|I|=S}\|\Phi_I^\star y_1\|_1\right)-\mathbb{E}\left(\max_{|I|=S}\|\bar\Psi_I^\star y_1\|_1\right)$ calculated in the previous proofs. Under the conditions given in Theorem 4 we have,

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\Phi_I^\star y_n\|_1-\frac{1}{N}\sum_{n=1}^{N}\max_{|I|=S}\|\Psi_I^\star y_n\|_1\\
&\qquad\geq\bar\varepsilon^2\frac{\bar c_1+\ldots+\bar c_S}{2K}-4\bar\varepsilon SK\sqrt{B}\exp\left(-\frac{(\beta-\frac{\bar\varepsilon^2}{2})^2}{8\bar\varepsilon^2}\right)-2t-\delta S\sqrt{B}.
\end{aligned}\tag{26}$$

To make sure that the above expression is larger than zero, we split it into two conditions. The first condition,

$$\bar\varepsilon^2\frac{\bar c_1+\ldots+\bar c_S}{4K}>4\bar\varepsilon SK\sqrt{B}\exp\left(-\frac{(\beta-\frac{\bar\varepsilon^2}{2})^2}{8\bar\varepsilon^2}\right),$$

is satisfied as soon as

$$\bar\varepsilon\leq\frac{\beta}{\frac{25\sqrt{8}}{99}\left(1+4\sqrt{\log\left(\frac{50K^2S\sqrt{B}}{\beta(\bar c_1+\ldots+\bar c_S)}\right)}\right)}.$$

To concretise the second condition,

$$\bar\varepsilon^2\frac{\bar c_1+\ldots+\bar c_S}{4K}\geq 2t+\delta S\sqrt{B},$$

we choose $t=\tilde\varepsilon^2\frac{\bar c_1+\ldots+\bar c_S}{16K}$ and $\delta=\tilde\varepsilon^2\frac{\bar c_1+\ldots+\bar c_S}{8KS\sqrt{B}}$ to arrive at $\bar\varepsilon\geq 2\tilde\varepsilon$. Given that $\bar\varepsilon$ differs at most by δ from ε we see that (26) is larger than zero except with probability

$$2\exp\left(-\frac{N\tilde\varepsilon^4(\bar c_1+\ldots+\bar c_S)^2}{128\varepsilon_{\max}^2S^2K^2B}+Kd\log\left(\frac{24\varepsilon_{\max}KS\sqrt{B}}{\tilde\varepsilon^2(\bar c_1+\ldots+\bar c_S)}\right)\right),$$

as long as

$$\varepsilon_{\min} := \tilde{\varepsilon} + \frac{\tilde{\varepsilon}^2}{8K} \leq \varepsilon \leq \varepsilon_{\max} \leq \frac{\beta}{\frac{25\sqrt{8}}{99}\left(1 + 4\sqrt{\log\left(\frac{50K^2S\sqrt{B}}{\beta(\bar{c}_1+\ldots+\bar{c}_S)}\right)}\right)} - \frac{\tilde{\varepsilon}^2}{8K}. \tag{27}$$

While for the asymptotic results we tried to make $\varepsilon_{\max}$ as large as possible to indicate how large the basin of attraction could be, for the finite sample size results we want it as small as possible in order to keep the sampling complexity small and therefore choose $\varepsilon_{\max} = \varepsilon_{\min}$. The statement then follows from making sure that the right most inequality in (27) is satisfied and simplifications.

In case of the noisy signal model, that is under the conditions given in Theorem 5, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N} \max_{|I|=S} \|\Phi_I^\star y_n\|_1 - \frac{1}{N}\sum_{n=1}^{N} \max_{|I|=S} \|\Psi_I^\star y_n\|_1 \\
&\quad \geq \bar{\varepsilon}^2 \frac{\bar{c}_1+\ldots+\bar{c}_S}{2C_rK} - 2t - \delta S\big(\sqrt{B}+1\big) - 2\bar{\varepsilon}SK\big(\sqrt{B}+1\big)\cdot \\
&\qquad \cdot \left(\exp\left(\frac{-\big(\beta - \frac{3\bar{\varepsilon}^2}{2}\big)^2}{72\bar{\varepsilon}^2}\right) + \exp\left(\frac{-\beta^2}{72\mu^2}\right) + \exp\left(\frac{-\beta^2}{72\rho^2}\right) + \exp\left(-\frac{\beta^2}{18\rho^2}\right)\right).
\end{aligned} \tag{28}$$

Splitting equally gives us four conditions,

$$\begin{aligned}
\bar{\varepsilon} &\geq \frac{16C_rSK^2\big(\sqrt{B}+1\big)}{\bar{c}_1+\ldots+\bar{c}_S} \exp\left(-\frac{\beta^2}{72\mu^2}\right) := \varepsilon_\mu, \\
\bar{\varepsilon} &\geq \frac{16C_rSK^2\big(\sqrt{B}+1\big)}{\bar{c}_1+\ldots+\bar{c}_S} \exp\left(-\frac{\beta^2}{72\rho^2}\right) := \varepsilon_\rho, \\
\bar{\varepsilon}^2 &\geq \frac{8C_rK}{\bar{c}_1+\ldots+\bar{c}_S}\left(2t + \delta S\big(\sqrt{B}+1\big)\right), \\
\bar{\varepsilon}(1 - \frac{\bar{\varepsilon}^3}{64}) &> \frac{16C_rSK^2\big(\sqrt{B}+1\big)}{\bar{c}_1+\ldots+\bar{c}_S} \exp\left(-\frac{\big(\beta - \frac{3\bar{\varepsilon}^2}{2}\big)^2}{72\bar{\varepsilon}^2}\right).
\end{aligned} \tag{29}$$

Choosing $t = \tilde{\varepsilon}^2_{\mu,\rho}\frac{\bar{c}_1+\ldots+\bar{c}_S}{32C_rK}$ and $\delta = \tilde{\varepsilon}^2_{\mu,\rho}\frac{\bar{c}_1+\ldots+\bar{c}_S}{16KS(\sqrt{B}+1)}$ we can merge the first three conditions to $\bar{\varepsilon} \geq \tilde{\varepsilon}^2_{\mu,\rho}$, while following the usual argument, Condition (29) is satisfied once

$$\bar{\varepsilon} \leq \frac{\beta}{\frac{9}{4} + 9\sqrt{\log\left(\frac{150K^2S\left(\sqrt{B}+1\right)}{\beta C_r(\bar{c}_1+\ldots+\bar{c}_S)}\right)}}.$$

Given that $\bar{\varepsilon}$ differs at most by δ from ε we see that (28) is larger than zero except with probability

$$2\exp\left(-\frac{N\tilde{\varepsilon}^4_{\mu,\rho}(\bar{c}_1+\ldots+\bar{c}_S)^2}{512\varepsilon^2_{\max}C_r^2S^2K^2\big(\sqrt{B}+1\big)^2} + Kd\log\left(\frac{48\varepsilon_{\max}KS\big(\sqrt{B}+1\big)}{\tilde{\varepsilon}^2_{\mu,\rho}(\bar{c}_1+\ldots+\bar{c}_S)}\right)\right),$$

as long as

$$\varepsilon_{\min} := \tilde{\varepsilon}_{\mu,\rho} + \frac{\tilde{\varepsilon}^2_{\mu,\rho}}{16K} \leq \varepsilon \leq \varepsilon_{\max} \leq \frac{\beta}{\frac{9}{4} + 9\sqrt{\log\left(\frac{150K^2S\left(\sqrt{B}+1\right)}{\beta C_r(\bar{c}_1+\ldots+\bar{c}_S)}\right)}} - \frac{N^{-2q}}{K}. \tag{30}$$

Again the statement follows from choosing $\varepsilon_{\max} = \varepsilon_{\min}$, making sure that the right most inequality in (30) is satisfied and simplifications. ■

References

[1] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. *COLT 2014 (arXiv:1310.7991)*, 2014.

[2] A. Agarwal, A. Anandkumar, and P. Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *COLT 2014 (arXiv:1309.1952)*, 2014.

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

[4] S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. *COLT 2014 (arXiv:1308.6273)*, 2014.

[5] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[6] O. Christensen. *An Introduction to Frames and Riesz Bases*. Birkhäuser, 2003.

[7] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[8] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, January 2006.

[9] 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.

[10] Q. Geng, H. Wang, and J. Wright. On the local correctness of ℓ^1-minimization for dictionary learning. *arXiv:1101.5672*, 2011.

[11] 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.

[12] A. Gersho and R.M. Gray. *Vector quantization and signal compression.* Springer, 1992.

[13] R. Gribonval and K. Schnass. Dictionary identifiability - sparse matrix-factorisation via l_1-minimisation. *IEEE Transactions on Information Theory*, 56(7):3523–3539, July 2010.

[14] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert. Sample complexity of dictionary learning and other matrix factorizations. *arXiv:1312.3790*, 2013.

[15] D. Hsu, S.M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability (arXiv:1110.2842)*, 17 (14), 2012.

[16] R. Jenatton, F. Bach, and R. Gribonval. Sparse and spurious: dictionary learning with noise and outliers. *arXiv:1407.5155*, 2014.

[17] K. Kreutz-Delgado and B.D. Rao. FOCUSS-based dictionary learning algorithms. In *SPIE 4119*, 2000.

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

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[20] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

[21] N.A. Mehta and A.G. Gray. On the sample complexity of predictive sparse coding. *arXiv:1202.4050*, 2012.

[22] 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.

[23] R. Rubinstein, A. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[24] K. Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD. *Applied Computational Harmonic Analysis*, 37(3):464–491, 2014.

[25] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14(11):828–831, 2007.

[26] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Transactions on Signal Processing*, 56(5):1994–2002, 2008.

[27] K. Skretting and K. Engan. Recursive least squares dictionary learning algorithm. *IEEE Transactions on Signal Processing*, 58(4):2121–2130, April 2010.

[28] D. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *COLT 2012 (arXiv:1206.5882)*, 2012.

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

[30] D. Vainsencher, S. Mannor, and A.M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12(3259-3281), 2011.

[31] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, chapter 5. Cambridge University Press, 2012.

[32] M. Yaghoobi, T. Blumensath, and M.E. Davies. Dictionary learning for sparse approximations with the majorization method. *IEEE Transactions on Signal Processing*, 57 (6):2178–2191, June 2009.

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