Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms

Rémi Gribonval, Holger Rauhut, Karin Schnass, Pierre Vandergheynst

▶ **To cite this version:**

Rémi Gribonval, Holger Rauhut, Karin Schnass, Pierre Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. [Research Report] PI 1848, 2007, pp.34. inria-00146660

HAL Id: inria-00146660
https://inria.hal.science/inria-00146660

Submitted on 15 May 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

I R I S A
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION
INTERNE
N° 1848

ATOMS OF ALL CHANNELS, UNITE! AVERAGE CASE ANALYSIS OF MULTI-CHANNEL SPARSE RECOVERY USING GREEDY ALGORITHMS

RÉMI GRIBONVAL , HOLGER RAUHUT , KARIN SCHNASS , PIERRE VANDERGHEYNST

ISSN 1166-8687

IRISA
CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms

Rémi Gribonval [*] , Holger Rauhut[**] , Karin Schnass[***] , Pierre Vandergheynst[****]

Systèmes cognitifs
Projet Metiss

Publication interne n ° 1848 — Mai 2007 — 34 pages

Abstract: This paper provides new results on computing simultaneous sparse approximations of multichannel signals over redundant dictionaries using two greedy algorithms. The first one, p-thresholding, selects the S atoms that have the largest p-correlation while the second one, p-simultaneous matching pursuit (p-SOMP), is a generalisation of an algorithm studied by Tropp in [28]. We first provide exact recovery conditions as well as worst case analyses of all algorithms. The results, expressed using the standard cumulative coherence, are very reminiscent of the single channel case and, in particular, impose stringent restrictions on the dictionary. We unlock the situation by performing an *average* case analysis of both algorithms. First, we set up a general probabilistic signal model in which the coefficients of the atoms are drawn at random from the standard gaussian distribution. Second, we show that under this model, and with mild conditions on the coherence, the probability that p-thresholding and p-SOMP fail to recover the correct components is overwhelmingly small and gets smaller as the number of channels increases. Furthermore, we analyse the influence of selecting the set of correct atoms at random. We show that, if the dictionary satisfies a uniform uncertainty principle [5], the probability that simultaneous OMP fails to recover any sufficiently sparse set of atoms gets increasingly smaller as the number of channels increases. To conclude, we study the robustness of these algorithms to an imperfect knowledge of the dictionary, a situation met in sparsity-based blind source separation where the dictionary, which corresponds to a mixing matrix, is only approximately known. In this framework, we estimate the probability of failure of the considered algorithms as a function of the similarity between the reference dictionary and the approximate one, which we measure with the smallest correlation between corresponding pairs of atoms.

Key-words: sparse representation, simultaneous approximation, multichannel data, overcomplete dictionary, matching pursuit, thresholding, greedy algorithm, identifiability, robustness.

(Résumé : tsvp)

This paper has been submitted for possible publication to the Journal of Fourier Analysis and Applications. This work is supported in part by the European Union's Human Potential Programme, under contract HPRN-CT-2002-00285 (HASSIP). H. Rauhut is supported by an Individual Marie Curie Fellowship from the European Union under contract MEIF-CT 2006-022811. R. Gribonval, K. Schnass and P. Vandergheynst are members of the INRIA funded partner team SPARS between IRISA and EPFL.

[*] remi.gribonval@irisa.fr.
[**] Universität Wien, Faculty of Mathematics, Nordbergstrasse 15, 1090 Vienna, Austria, holger.rauhut@univie.ac.at
[***] Ecole Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Institute - ITS, 1015 Lausanne, Switzerland, karin.schnass@epfl.ch
[****] EPFL, Signal Processing Institute - ITS, 1015 Lausanne, Switzerland, pierre.vandergheynst@epfl.ch

Centre National de la Recherche Scientifique (UMR 6074) Université de Rennes 1 – Insa de Rennes
Institut National de Recherche en Informatique et en Automatique – unité de recherche de Rennes

Atomes de tous les canaux, unissez-vous! Une analyse au cas moyen des algorithmes gloutons de décomposition parcimonieuse multi-canal

Résumé : Cet article est consacré à l'analyse du comportement de deux algorithmes gloutons pour le calcul d'approximations parcimonieuses simultanées de signaux multicanaux avec des dictionnaires redondants. Le premier algorithme, *p-thresholding*, sélectionne les S atomes qui ont la plus grande p-corrélation avec le signal analysé, tandis que le second, *p-Simultaneous Orthonormal Matching Pursuit* (p-SOMP) est une généralisation d'un algorithme étudié par Tropp [28]. Pour chaque algorithme, nous fournissons des conditions d'identifiabilité de représentations parcimonieuses conjointes, ainsi qu'une analyse au pire cas de ces conditions. Les résultats, exprimés à l'aide de la cohérence cumulative classique, sont similaires au cas monocanal et imposent des restrictions très fortes sur le dictionnaire pour garantir le succès des algorithmes considérés.

Nous débloquons la situation en effectuant une analyse *au cas moyen* de ces deux algorithmes. Pour commencer, nous proposons un modèle probabiliste des signaux où les coefficients des atomes de la représentation sont générés aléatoirement selon une distribution Gaussienne. Ensuite, dans le cadre de ce modèle, nous prouvons que la probabilité que p-thresholding et p-SOMP ne retrouvent pas les bons atomes devient négligeable lorsque le nombre de canaux croît, sous des conditions assez faibles de cohérence du dictionnaire.

En outre, nous analysons un modèle génératif où l'ensemble des atomes de la représentation est lui-même tiré au hasard. Dans ces conditions, nous montrons que si le dictionnaire satisfait un principe d'incertitude uniforme [5] alors la probabilité que p-SOMP ne retrouve pas le bon jeu d'atomes devient négligeable lorsque le nombre de canaux croît.

Pour finir, nous considérons la question de la robustesse de ces algorithmes vis-à-vis d'une connaissance imparfaite du dictionnaire servant à modéliser les signaux, comme par exemple dans le cadre de la séparation aveugle de sources où le dictionnaire, qui correspond la matrice de mélange, est seulement approximativement connu. Dans ce cadre nous estimons la probabilité d'échec des algorithmes considérés en fonction de la ressemblance entre le dictionnaire de référence et le dictionnaire estimé, représentée par la plus faible corrélation entre paires d'atomes associés.

Mots clés : représentation parcimonieuse, approximation simultanée, traitement du signal multicanal, dictionnaire redondant, matching pursuit, seuillage, algorithme glouton, identifiabilité, robustesse.

Contents

1 Introduction **3**
1.1 Signal model 4
1.2 Recovery problem 5

2 Technical tools and notations **7**
2.1 Matrix norms 7
2.2 Babel functions and isometry constants 8

3 Main results **9**

4 Worst Case Analysis **15**

5 Average case analysis for thresholding **17**
5.1 Spirit of the proof 17
5.2 Concentration of measure 18
5.3 Main result for p-thresholding 18

6 Average case analysis of OMP **20**
6.1 Spirit of the proof 20
6.2 A general recovery result 21
6.3 Bounds on $c_0(\Lambda)$ and $d_0(\Lambda)$ 23
6.4 Proof of Theorem 3.4 24
6.5 Proof of Theorem 3.5 25
6.6 Proof of Theorem 3.6 26

7 Robustness with respect to the dictionary **26**

8 Conclusions and Outlook **29**

A Proof of Theorem 5.1 **29**

B Computation of $A_p(N)$ and $C_p(N)$ **31**

1 Introduction

Transform coding is one of the most successful paradigms in signal processing. Generally speaking, it asserts that many signals can be efficiently compressed because they have a sparse representation on some fixed basis. A simple transform coder would then decompose the signal over this optimal basis and threshold all projections to locate and keep only the K strongest ones. This simple algorithm is at the core of the success of modern image and video coders such as JPEG2000 where a wavelet basis is used [23, 11]. Recently though, new problems have come to challenge that paradigm. Restricting our models to decompositions over fixed bases drastically narrows the class of signals that

[0] *Math Subject Classifications:* 41A28, 41A46, 60D05.
[0] *Keywords and Phrases:* Greedy algorithms, OMP, Thresholding, multi-channel, average analysis.

can be efficiently processed. A lively strand of research advocates richer models based on redundant dictionaries, which can capture a much broader range of signals. A dictionary $\mathbf{\Phi}$ is a large collection of unit norm vectors $\|\varphi_n\|_2 = 1$, $n = 1, ..., K$ in $\mathbb{R}^d$, usually with $K \gg d$. Handling arbitrary dictionaries is no easy task, though. First, uniqueness of a signal representation is not guaranteed anymore. Second, even computing a decomposition becomes a complicated issue: several algorithms, most notably greedy algorithms and convex relaxation techniques can be used, but analysing their performances remained a daunting challenge. The situation unlocked with the realisation that sparse models solve these problems. To illustrate the role of sparsity, let us introduce the *coherence* of the dictionary, i.ethe strongest correlation between any two distinct vectors in $\mathbf{\Phi}$: $\mu = \max_{i \neq j} |\langle \varphi_i, \varphi_j \rangle|$. Schematically, if a signal is a superposition of less than μ^{-1} elements of $\mathbf{\Phi}$, this representation is unique and can be recovered by standard algorithms [24, 26, 10].

In parallel to developments in sparse signal models, various application scenarios motivated renewed interest in processing not just a single signal, but many signals or channels at the same time. A striking example is sensor networks, where signals are monitored by low complexity devices whose observations are transfered to a central collector [17]. This central node thus faces the task of analysing many, possibly high-dimensional, signals. Moreover, signals measured in sensor networks are typically not uncorrelated: there are global trends or components that appear in all signals, possibly in slightly altered forms. Modeling multichannel signals by means of redundant dictionaries, generalising existing mono-channel algorithms and understanding their properties are thus important challenges.

In this paper we analyse the theoretical performances of two classes of simultaneous greedy algorithms, p-thresholding and p-SOMP. In both cases, we provide worst case recovery conditions, but our main contribution with respect to prior art is a rigorous average case analysis of both classes of algorithms. The spirit of our results, described in Section 3, is that by allowing an overwhelmingly small probability of error, we get more favourable recovery conditions, far better than what had been previously reported in the worst case.

Our analysis is based on studying the average case instead of the worst case and the spirit of our results is the following: We show that given a dictionary of coherence μ, p-thresholding can recover superpositions of up to μ^{-2} atoms *with overwhelming probability*, provided that the *dynamic range* of the signal coefficients is somewhat limited. Our conditions on $\mathbf{\Phi}$ are thus much less restrictive than in the worst case. In particular, we provide quantitative versions of the results for distributed compressed sensing in [3], which even allow to work with deterministic measurement matrices.

1.1 Signal model

Suppose we are to design a network of N sensors monitoring a common phenomenon. Each of our sensors observes a d-dimensional signal $y_n \in \mathbb{R}^d$, $n = 1, ..., N$. As explained in the previous section, a sparsity hypothesis will be the central assumption of our model: we will assume that each signal y_n admits a sparse approximation over a single dictionary $\mathbf{\Phi}$,

$$y_n = \mathbf{\Phi} x_n + e_n,\ n = 1, ..., N.$$

Sparsity in this case is embodied in each of the coefficient vectors x_n, which are assumed to have few non zero entries as measured by their ℓ_0 "norm"[1]: $\|x\|_0 \leq S$. In order to model correlations between signals, we will refine this model by imposing that all signals share a common sparse support, i.e

$$y_n = \mathbf{\Phi}_\Lambda x_n + e_n,$$

where $\mathbf{\Phi}_\Lambda$ is the restriction of the synthesis matrix $\mathbf{\Phi}$ to the columns listed in the set Λ. In this case, sparsity is conveyed by the size of the support set, $|\Lambda| \leq S$, and there is thus no restriction on the coefficient vectors. This model is inspired by a recent series of papers on distributed sensing, see [2] and references therein. It describes a network of sensors monitoring a signal with a strong global component that appears at each node. Localised effects are modelled by letting synthesis coefficients x_n vary across nodes and through the innovations e_n. As an illustrative example, imagine sensors measuring the chemical composition of the atmosphere at some locations of a geographical area. There is a common component, say a mean regular chemical composition, modelled by the fixed support Λ. But it changes slightly from node to node because of differences in sensor location (latitude, altitude, ...); these are modelled by varying the amplitudes x_n of components from node to node. Localised effects, like pollution or forest fires, can drastically alter the signal and are captured by transient innovations e_n. The very nature of these innovation signals e_n will thus depend on the exact problem one wants to solve. However, and for simplicity, we will in this paper assume that they are orthogonal to the subspace spanned by Λ. Hence $\mathbf{\Phi}_\Lambda x_n$ is the best approximation of y_n by elements of Λ in mean squared sense. Note that we will sometimes refer to e_n as noise, in a clear but hopefully not misleading abuse of language.

Let us now turn towards describing a generative model for the synthesis coefficients x_n. In order to obtain a sufficiently general model, we will assume that the components $x_n(i)$, $i \in \Lambda$ of the random vector x_n are independent Gaussian variables of variance σ_i. This model is fairly general to accommodate various practical problems: the Gaussian assumption is one of the most widely used in signal processing, while incorporating different variances allows us to shape the synthesis coefficients, imposing statistical decay for example on the $x_n(i)$.

In order to simplify our analysis we will adopt a global matrix notation. We will collect all signals on the columns of the $d \times N$ matrix $Y = [y_1, \ldots, y_N]$. Let U be a $S \times N$ random matrix with independent standard gaussian entries and let Σ be a $S \times S$ diagonal matrix whose diagonal entries σ_i^2 are positive real numbers. Our model can then be written in compact form:

$$Y = \mathbf{\Phi}_\Lambda \cdot \Sigma^{\frac{1}{2}} \cdot U + E, \tag{1.1}$$

where E is a $d \times N$ matrix collecting innovation (noise) signals e_n on its columns.

1.2 Recovery problem

A typical problem consists in recovering either the support Λ (this is a recovery problem) or the coefficients X (this is an estimation problem) from the observation Y. For that, algorithms must be designed, and their success must be characterised depending on the

[1]Note that we adopted a common abuse of language, since $\|\cdot\|_0$ is not a norm, neither a quasi-norm.

noise level and other characteristics of the multichannel sparse signal model. Typical (single channel) sparse approximation algorithms rely on the computation of the inner products $\langle y, \varphi_k \rangle$ between the signal y and the atoms φ_k of the dictionary, which are the entries of the vector $\mathbf{\Phi}^\star y$. In the multichannel setting, we will consider algorithms that similarly rely on the entries $\langle y_n, \varphi_k \rangle$ of the matrix $\mathbf{\Phi}^\star Y$. Instead of inner products with the atoms φ_k involved in the signal model, it is also interesting to consider variants where other atoms ψ_k, which we will call *sensing atoms*, are used in the algorithms, cp. [22]. In other words, the algorithms will rely on the entries of $\mathbf{\Psi}^\star Y$. One of the reasons for introducing such sensing atoms is that, in some cases, the signal model is only approximately known so one cannot use the (unknown) dictionary $\mathbf{\Phi}$ in an algorithm. Another reason is that an added freedom in the choice of the sensing matrix may also improve the provable performance of the considered algorithms.

Thresholding algorithm. Of the two families of sparse approximation algorithms considered in this paper, the family of simultaneous thresholding algorithms is certainly the simplest one. In the single channel case, thresholding amounts to selecting the atoms of the dictionary which are most correlated with the signal y. In the multichannel setting, the main change is that one should combine the correlation of the atom with the different channels to get a single interchannel correlation criterion for the selection of the most correlated atoms. For any $1 \leq p \leq \infty$ one can consider the p-correlation

$$\|\psi_k^\star Y\|_p := \left(\sum_{n=1}^{N} |\langle \psi_k, y_n \rangle|^p \right)^{1/p} \tag{1.2}$$

with the standard modification for $p = \infty$. The p-thresholding algorithm simply amounts to selecting a set Λ_M of M atoms whose p-correlations with Y are among the M largest

$$\|\psi_k^\star Y\|_p \geq \|\psi_l^\star Y\|_p, \forall k \in \Lambda_M, \forall l \notin \Lambda_M. \tag{1.3}$$

In addition to an estimated support Λ_S, p-thresholding can also be used to provide an estimate of the coefficients X, which is most easily done by least squares optimisation, leading to $X_M := \mathbf{\Phi}_{\Lambda_M}^\dagger Y$ where $\mathbf{\Phi}_{\Lambda_M}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{\Phi}_{\Lambda_M}$.

Greedy algorithm. Simultaneous Orthogonal Matching Pursuit (SOMP) is a somewhat more elaborate iterative algorithm for sparse signal approximation. At each iteration, an atom index k_m is selected, and a residual is updated. At the first iteration the residual is simply $Y_0 := Y$. After M iterations, the set of selected atoms being $\Lambda_M := \{k_m\}_{k=1}^M$, the new residual is computed as $Y_M = Y - \mathbf{\Phi}_{\Lambda_M} X_M = (\mathbf{I} - \mathbf{P}_{\Lambda_M}) Y$ where $X_M := \mathbf{\Phi}_{\Lambda_M}^\dagger Y$ and $\mathbf{P}_{\Lambda_M} = \mathbf{\Phi}_{\Lambda_M} \mathbf{\Phi}_{\Lambda_M}^\dagger$ is the orthogonal projection onto the linear span of the selected atoms. In p-SOMP, the next selected atom k_{M+1} is the one which maximises the p-correlation with the residual Y_M

$$\|\psi_{k_{M+1}}^\star Y_M\|_p = \max_{1 \leq k \leq K} \|\psi_k^\star Y_M\|_p. \tag{1.4}$$

Recovering the right support. Given the model $Y = \mathbf{\Phi}_\Lambda X + E$, we will say by definition that p-thresholding (respectively p-SOMP) "recovers" Λ if when we set $M = |\Lambda|$, the selected set Λ_M exactly matches Λ. Occasionally we may also be interested in partial recovery, meaning that for some $M \leq |\Lambda|$ the algorithms only select "good" atoms, i.e $\Lambda_M \subset \Lambda$.

2 Technical tools and notations

This section provides the main tools and notations which will be used over and over in the remaining of this article to state and prove our results.

2.1 Matrix norms

In order to be able to neatly analyse the algorithms in the next sections it will be convenient to define the following matrix norms. Let A be a $n \times m$-matrix with rows $(A_i)_{1\dots n}$ then we define

$$\|A\|_{p,\infty} := \max_{i=1\dots n} \|A_i\|_p = \max_{i=1\dots n} \big(\sum_{j=1}^{m} |A_{ij}|^p \big)^{\frac{1}{p}}. \tag{2.1}$$

Note that this matrix norm should not be confused with the operator norm $\|\!|A|\!\|_{p\to\infty}$, which for general $1 \le p, q \le \infty$ is defined as:

$$\|\!|A|\!\|_{p\to q} = \max_{\|x\|_p=1} \|Ax\|_q. \tag{2.2}$$

However, there exists a connection between the two norm types which we will exploit later to prove some easy inequalities. Namely if $\frac{1}{p} + \frac{1}{p'} = 1$ we have

$$\|A\|_{p,\infty} = \|\!|A|\!\|_{p'\to\infty}. \tag{2.3}$$

Among the p, q-operator norms the $2, 2$-operator norm will play an important role as it is connected to the spectrum of the matrix, i.e,

$$\|\!|A|\!\|_{2\to 2} = \lambda_{\max}(A) = \text{largest singular value of } A. \tag{2.4}$$

Also we will write for shortness $\|\!|\cdot|\!\| := \|\!|\cdot|\!\|_{2\to 2}$. The following lemma collects two useful properties of operator norms.

Lemma 2.1. *1. For two matrices A, B we have*

$$\|\!|AB|\!\|_{p\to q} \le \|\!|B|\!\|_{p\to s} \|\!|A|\!\|_{s\to q}. \tag{2.5}$$

2. If $A^\dagger$ denotes the Moore-Penrose pseudo-inverse of A we have

$$\|\!|A^\dagger|\!\|_{2\to 2} = \frac{1}{\lambda_{\min}(A)}, \tag{2.6}$$

where $\lambda_{\min}(A)$ denotes the smallest non-zero singular value of A.

The following trivial Corollary will be essential for some recovery results in this paper.

Corollary 2.2. *For two matrices A, B we have*

$$\frac{\|AB\|_{p,\infty}}{\|B\|_{p,\infty}} \le \|\!|A|\!\|_{\infty\to\infty} = \|A\|_{1,\infty} = \max_{i=1..n} \sum_{j=1}^{m} |A_{ij}|. \tag{2.7}$$

2.2 Babel functions and isometry constants

A few essential tools have emerged from the literature to characterise which sparse representations from a redundant dictionary can be recovered with typical algorithms such as ℓ^1-minimization and greedy algorithms. Here we recall the definitions of the Babel function, also known as cumulative coherence, and the restricted/global isometry constants of a dictionary. Where necessary, we adapt these tools to handle pairs $(\mathbf{\Phi}, \mathbf{\Psi})$ made of a dictionary $\mathbf{\Phi}$, from which the sparse signals $Y \approx \mathbf{\Phi} X$ are built, and a sensing dictionary $\mathbf{\Psi}$, used to compute correlations with the signal Y.

p-Babel functions.
The p-Babel function for a subset Λ is the most tangible characteristics of a given pair of dictionaries $(\mathbf{\Phi}, \mathbf{\Psi})$ of equal size. It is defined in the computationally explicit form as

$$\mu_p(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda) := \sup_{\ell \notin \Lambda} \big(\sum_{j \in \Lambda} |\langle \varphi_j, \psi_\ell \rangle|^p \big)^{\frac{1}{p}} \tag{2.8}$$

and measures the amount of correlation between sensing atoms ψ_ℓ *outside* the support Λ and modeling atoms φ_j *inside* the support Λ. A complement to the p-Babel function measures the amount of correlation between atoms *inside* the support Λ

$$\mu_p^{in}(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda) := \sup_{i \in \Lambda} \mu_p(\mathbf{\Phi}_\Lambda, \mathbf{\Psi}_\Lambda, \Lambda \backslash \{i\}). \tag{2.9}$$

Taking the supremum over all possible subsets of size at most S, we get the definition of the p-Babel function for an integer S as

$$\mu_p(\mathbf{\Phi}, \mathbf{\Psi}, S) := \sup_{|\Lambda| \leq S} \mu_p(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda). \tag{2.10}$$

A similar definition is used for $\mu_p^{in}(\mathbf{\Phi}, \mathbf{\Psi}, S)$, which trivially yields the relation

$$\mu_p^{in}(\mathbf{\Phi}, \mathbf{\Psi}, S) \leq \mu_p(\mathbf{\Phi}, \mathbf{\Psi}, S-1). \tag{2.11}$$

Most interesting for us will be the cases $p = 1$ and $p = 2$. When the sensing dictionary $\mathbf{\Psi}$ equals the modeling one $\mathbf{\Phi}$, the reader can easily check that the p-Babel function for $p = 1$ matches the standard definition of the Babel function which can be found, e.g., in Tropp's enjoyable paper [24].

Shorthands.
In several sections of this article, we will omit the reference to the dictionary pair $(\mathbf{\Phi}, \mathbf{\Psi})$ if it is clear which one we are considering and will write simply $\mu_p(\Lambda)$, $\mu_p^{in}(\Lambda)$, $\mu_p(S)$ and $\mu_p^{in}(S)$. Similar shorthands will be used for the notations introduced hereafter.

Similarity between sensing and modeling dictionaries.
While p-Babel functions measure the similarity between non-corresponding atoms in the original and the sensing dictionary, which we will want to be small to obtain recovery

results, we will also need a measure for the similarity between matching atom pairs φ_k, ψ_k, which we will then want to be large. For that we consider

$$\begin{aligned} \beta_k &:= \langle \varphi_k, \psi_k \rangle > 0, & (2.12) \\ \beta(\Lambda) &:= \min_{i \in \Lambda} \beta_i & (2.13) \end{aligned}$$

The assumption that $\beta_k > 0$ is merely a convention which can always be guaranteed by slightly changing the definition of the sensing dictionary $\mathbf{\Psi}$, replacing ψ_k by $-\psi_k$ if necessary.

Isometry constants.
In order to bound the spectrum of a subdictionary $\mathbf{\Phi}_\Lambda$ we define the isometry constant $\delta_\Lambda = \delta_\Lambda(\mathbf{\Phi})$ as the smallest quantity such that

$$(1 - \delta_\Lambda) \cdot \|x\|_2^2 \leq \|\mathbf{\Phi}_\Lambda x\|_2^2 \leq (1 + \delta_\Lambda) \cdot \|x\|_2^2 \qquad \forall x \neq 0. \tag{2.14}$$

Note that the definition above provides the following bound on the extremal singular values of $\mathbf{\Phi}_\Lambda$

$$\lambda_{\min}(\mathbf{\Phi}_\Lambda) \geq \sqrt{1 - \delta_\Lambda} \qquad \text{and} \qquad \lambda_{\max}(\mathbf{\Phi}_\Lambda) \leq \sqrt{1 + \delta_\Lambda}. \tag{2.15}$$

Since we also want a uniform estimate over all possible subdictionaries of a given size, we define for an integer S the global (restricted) isometry constant

$$\delta_S := \sup_{|\Lambda| = S} \delta_\Lambda \tag{2.16}$$

and easily check that δ_S is a non-decreasing function of S. Restricted isometry constants were introduced by Candès, Romberg and Tao in [4, 5] in order to study recovery by Basis Pursuit (ℓ_1) in the context of compressed sensing. Indeed if $\delta_{3S} + 3\delta_{4S} < 2$ then Basis Pursuit recovers all S-sparse (mono-channel) signals [4]. Good estimates of these numbers were obtained for random Gaussian and Bernoulli $d \times K$ matrices $\mathbf{\Phi}$: If

$$S \leq C_\delta \frac{d}{\log\left(\frac{K}{S\epsilon}\right)} \tag{2.17}$$

then with probability at least $1 - \epsilon$ the restricted isometry constant of $\mathbf{\Phi}$ satisfies $\delta_S \leq \delta$, see e.g. [5, 1, 19]. A similar result holds for random partial Fourier matrices under the condition $S \leq C_\delta d \log^{-4}(K) \log^{-1}(\epsilon^{-1})$, see [5, 20, 18].

3 Main results

The analysis of both p-thresholding and p-SOMP follows a similar pattern. First, we provide subtle sufficient conditions which guarantee that the considered algorithm (partially) recovers the desired support. In addition to the noise level, these recovery conditions depend on subtle joint properties of the analysis and synthesis dictionaries, of the ideal support Λ, of the signal coefficients X, etc. Next we proceed with a worst case analysis which provides coarser worst case recovery conditions that depend more globally on

the sparsity of X, on its "dynamic range", etc. Such a worst case analysis gives results expressed in terms of cumulative coherence of the dictionary which are essentially of the same strength and flavour as similar results for recovery in the monochannel setting. Last, we show how to switch from a worst case analysis to an average case analysis: assuming a specific probabilistic model on the coefficients X, we provide conditions on the sparsity of X that guarantee that the subtle recovery conditions are satisfied with high probability. This drastically changes the strength of the required conditions, since by allowing a small amount of failure of the algorithms for non typical coefficients, this significantly increases the size of the supports that can be recovered.

In order to give a more quantitative feeling of our results, we will highlight them with the example of a dictionary composed of the union of the Dirac and DCT bases (hereby simply referred to as the Dirac-DCT dictionary). More precisely, $\mathbf{\Phi}_{\mathrm{DDCT}}$ is the $d \times 2d$ matrix obtained by concatenating the $d \times d$ identity matrix and the $d \times d$ DCT matrix whose k-th column is:

$$\varphi_k(n) = \sqrt{\frac{2}{d}}\Omega_k \cos\left(\frac{\pi}{2d}(2n-1)(k-1)\right),\ n = 1,\ ...,\ d,$$

w ith $\Omega_k = 1/\sqrt{2}$ for $k = 1$ and $\Omega_k = 1$ for $2 \leq k \leq d$. This dictionary has coherence $\mu = \sqrt{2/d}$ and it is also easy to see that $\mu_p(S) = S^{1/p} \cdot \mu$.

Recovery conditions for p-thresholding. The success of p-thresholding at recovering the good support Λ is guaranteed for a given signal model $Y = \mathbf{\Phi}_\Lambda X + E$ as soon as the minimum p-correlation with good atoms $\min_{i\in\Lambda} \|\psi_i^\star Y\|_p$ exceeds the maximum p-correlation with "bad" atoms $\|\mathbf{\Psi}_{\overline{\Lambda}}^\star Y\|_{p,\infty}$ where $\overline{\Lambda} := \{1 \leq k \leq K, k \notin \Lambda\}$. By the triangle inequalities

$$\|\mathbf{\Psi}_{\overline{\Lambda}}^\star Y\|_{p,\infty} \leq \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda X\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty}$$

and

$$\min_{i\in\Lambda} \|\psi_i^\star Y\|_p \geq \min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda X\|_p - \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty},$$

we get the recovery condition

$$\|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} < \min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda X\|_p - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda X\|_{p,\infty}. \tag{3.1}$$

Recovery conditions for p-SOMP. As far as p-SOMP is concerned, it partially recovers the good support Λ after M steps if the set Λ_M only contains "good" atoms, that is to say if $\Lambda_M \subset \Lambda$. Since $\Lambda_{M+1} = \Lambda_M \cup \{k_{M+1}\}$, partial recovery after $M+1$ steps is equivalent to partial recovery after M steps with an additional good choice of the $M+1$-th atom, which is only guaranteed if $\|\mathbf{\Psi}_\Lambda^\star Y_M\|_{p,\infty} > \|\mathbf{\Psi}_{\overline{\Lambda}}^\star Y_M\|_{p,\infty}$. Denoting $\mathbf{Q}_{\Lambda_M} := \mathbf{I} - \mathbf{P}_{\Lambda_M}$ the orthogonal projection onto the complement of the span of the selected atoms (by convention $\mathbf{Q}_\emptyset = \mathbf{I}$), by the triangle inequalities

$$\|\mathbf{\Psi}_\Lambda^\star Y_M\|_{p,\infty} \geq \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty} - \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_{\Lambda_M} E\|_{p,\infty}$$

and

$$\|\mathbf{\Psi}_{\overline{\Lambda}}^\star Y_M\|_{p,\infty} \leq \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_{\Lambda_M} E\|_{p,\infty}$$

we get the recovery condition

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_{\Lambda_M} E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_{\Lambda_M} E\|_{p,\infty} < \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty}. \quad (3.2)$$

Under the simplifying assumption that $\mathbf{\Phi}_\Lambda^\star E = 0$, which we discuss below, if the first M steps of p-SOMP have been successful (that is to say if $\Lambda_M \subset \Lambda$) then $\mathbf{Q}_{\Lambda_M} E = E$, and we obtain that the $M+1$-th atom is guaranteed to be correct provided that

$$\|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} < \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty}. \quad (3.3)$$

Remark 3.1. The assumption that $\mathbf{\Phi}_\Lambda^\star E = 0$ might seem a bit artificial if one considers E as additive noise in the model, in which case it would seem more natural to assume it is a realization of, *e.g.*, a random Gaussian process. In contrast, from an approximation theory perspective, E would typically represent the error of best approximation of Y with the atoms in Λ, that is to say $E = Y - \mathbf{\Phi}_\Lambda X$ with $X = \arg\min_Z \|Y - \mathbf{\Phi}_\Lambda Z\|$ for some norm $\|\cdot\|$. When this norm is given by $\|Y - \mathbf{\Phi}_\Lambda X\| = (\sum_{n=1}^N \|y_n - \mathbf{\Phi}_\Lambda x_n\|_2^q)^{1/q}$ for some q, (*e.g.*, $q = 2$ for the Froebenius norm), this implies that E satisfies $\mathbf{\Phi}_\Lambda^\star e_n = 0$ for each n.

Both condition (3.1) and (3.3) mean that the noise level, as measured by $\|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty}$, should be small enough compared to some upper limit which jointly depends on the analysis and synthesis dictionaries $\mathbf{\Phi}$, $\mathbf{\Psi}$, the supports Λ and $\Lambda_M \subset \Lambda$, the coefficients X, etc. Next, we express simpler conditions that somehow untangle the role of the different objects that we are manipulating.

To state the worst case analysis of thresholding, we introduce a specific notation

$$\mathcal{X}_p^i = \left(\sum_{n=1}^N |X_{in}|^p \right)^{1/p}, \quad i \in \Lambda \quad (3.4)$$

for the p-norms of the rows of X, i.e $\mathcal{X}_p^i$ is the p-norm of the vector of coefficients associated to the i-th atom φ_i. A detailed analysis is carried out in the next section, yielding Theorem 4.1. We state below a somewhat simpler form of this result, assuming $\mathbf{\Psi} = \mathbf{\Phi}$.

Theorem 3.1 (Worst case analysis for thresholding)**.** *Assume that* $Y = \mathbf{\Phi}_\Lambda X + E$ *with*

$$\|\mathbf{\Phi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{p,\infty} < \min_{i\in\Lambda} \mathcal{X}_p^i - \max_{i\in\Lambda} \mathcal{X}_p^i \cdot (\mu_1(S) + \mu_1(S-1)), \quad (3.5)$$

where $S := |\Lambda|$*. Then, p-thresholding with* $\mathbf{\Psi} = \mathbf{\Phi}$ *exactly recovers the support* Λ*.*

Observe that we want to maximise the right hand side of (3.5) and, in particular, we want:

$$\frac{\min_{i\in\Lambda} \mathcal{X}_p^i}{\max_{i\in\Lambda} \mathcal{X}_p^i} > \mu_1(S) + \mu_1(S-1).$$

Since the ratio on the l.h.s of this equation is at most one, the most favourable case arises when the dynamic range of the coefficients is *small*, i.e when the components of Λ have the same strength. In the same expression, incoherence rears its ugly head, for even in the best case we have to assume

$$\mu_1(S) + \mu_1(S-1) < 1. \quad (3.6)$$

Since $\mu_1(S) \leq S\mu$, the sparsity of recoverable signals is thus roughly confined to the realm

$$S < \frac{1}{2}(\mu^{-1} + 1),$$

making it nearly useless for dictionaries one would use in practice. On the other hand experiments show that the range of useful sparsity is *much* bigger and confirm the intuition that typical results are much more favourable [28]. Understanding the average performance of simultaneous thresholding under the probabilistic signal model introduced in Section 1.1 is precisely our next contribution, detailed in Section 5, and summarised by the following result:

Theorem 3.2 (Average case analysis for 1-thresholding)**.** *Let $p = 1$ and $S = |\Lambda|$. Assume that $Y = \mathbf{\Phi}_\Lambda \, \Sigma^{\frac{1}{2}} \, U + E$ with U a $S \times N$ matrix of standard Gaussian random variables and $\Sigma = \mathrm{diag}(\sigma_i^2)_{i\in\Lambda}$, and suppose that*

$$\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} + \|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} < \sqrt{\frac{2}{\pi}} N \cdot \left(\min_{i\in\Lambda} \sigma_i - \max_{i\in\Lambda} \sigma_i \cdot \mu_2(S) \right). \tag{3.7}$$

Then the probability that p-thresholding with $\mathbf{\Psi} = \mathbf{\Phi}$ fails to exactly recover the support Λ does not exceed $K \exp(-N\gamma^2/\pi)$ with K the number of atoms in $\mathbf{\Phi}$ and

$$\gamma := \frac{\min_{i\in\Lambda} \sigma_i - \max_{i\in\Lambda} \sigma_i \cdot \mu_2(S) - \sqrt{\frac{\pi}{2}} N^{-1} \cdot \left(\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} + \|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} \right)}{\min_{i\in\Lambda} \sigma_i + \max_{i\in\Lambda} \sigma_i \cdot \mu_2(S)}. \tag{3.8}$$

Similar results hold for $1 < p \leq \infty$ where $\sqrt{\frac{2}{\pi}} N$ is replaced with a constant $C_p(N)$. Clearly, there is a common flavour with worst case results: we want to maximise the r.h.s of (3.7) and, for any fixed number of channels N, this implies

$$\frac{\min_{i\in\Lambda} \sigma_i}{\max_{i\in\Lambda} \sigma_i} > \mu_2(S).$$

The most favourable situation is once again reached when all components of Λ have the same strength, i.e when the ratio on the l.h.s gets close to one. This time however, observe that the range of allowed sparsity is constrained by the 2-Babel function $\mu_2(S) < 1$. Since $\mu_2(S)$ grows *much slower* than $\mu_1(S)$, we can now recover much more atoms, up to roughly $S = \mu^{-2}$, with high probability. When the number of channels N grows, condition (3.7) demands that the average noise per channel $N^{-1}(\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} + \|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty})$ be small enough, but once this is satisfied the probability of failure decreases exponentially fast with the number of channels N.

Even though the conditions for recovering typical signals with p-thresholding are milder than their worst case counterpart, the constraint that each component of the support be equally important remains quite a limitation of the algorithm. This motivates turning our attention to p-SOMP in hope that this more complex technique will allow us to relax those restrictions. We start by stating the worst case results for OMP which are proved in Section 4. For $p = 1$ they match the results by Tropp et al. [28], and for all p they generalise the results of Chen and Huo [7] to the noisy setting.

Theorem 3.3 (Worst case analysis for p-SOMP)**.** *Assume that $Y = \mathbf{\Phi}_\Lambda X + E$ where the atoms in Λ are linearly independent and*

$$\|\mathbf{\Phi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{p,\infty} < \min_{i\in\Lambda} \mathcal{X}_p^i \cdot \left(1 - \mu_1(\Lambda) - \mu_1^{in}(\Lambda)\right). \tag{3.9}$$

Then $S := |\Lambda|$ steps of p-SOMP with $\mathbf{\Psi} = \mathbf{\Phi}$ recover the support Λ.

This result is expressed in slightly different and finer terms than Theorem 3.1: here we give a characterisation of recoverable index sets by explicitly controlling the correlations among atoms on the support through the quantity $\mu_1^{in}(\Lambda)$ and correlations of the support with the rest of the dictionary through $\mu_1(\Lambda)$. Comparing (3.9) and (3.5) clearly shows the main advantage of OMP over thresholding: both conditions require the noise level to be small enough compared to some measure of dictionary coherence, but the restriction on the dynamic range of the signal has disappeared in (3.9). However, there is no quantitative gain on the size of S. If we give up our fine characterisation of Λ and estimate the r.h.s of (3.9) in terms of S, the right most term becomes $1 - \mu_1(S) - \mu_1(S-1)$ and we are back to (3.6). Once again, the obvious way to transcend this barrier is to understand the behaviour of the algorithm for typical signals and not in the worst case. A detailed analysis is performed in Section 6, but a simplified version of our result reads as follows.

Theorem 3.4. *Let $p = 1$, $S := |\Lambda|$ and $Y = \mathbf{\Phi}_\Lambda \, \Sigma^{\frac{1}{2}} \, U + E$ with U a $S \times N$ matrix of standard Gaussian random variables, $\Sigma = \operatorname{diag}(\sigma_i^2)_{i\in\Lambda}$, and E an error term orthogonal to the atoms in Λ. Suppose*

$$\kappa := 1 - \frac{\mu_2^{in}(\Lambda) + \mu_2(\Lambda)}{1 - \delta_\Lambda} > 0$$

and in addition

$$\|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty} < \sqrt{\frac{2}{\pi}} N \kappa \min_{i\in\Lambda} \sigma_i. \tag{3.10}$$

Then the probability that S steps of 1-SOMP with $\mathbf{\Psi} = \mathbf{\Phi}$ fail to exactly recover the support Λ does not exceed $K \cdot 2^S \cdot \exp(-N\gamma^2/\pi)$ with K the number of atoms in $\mathbf{\Phi}$ with

$$\gamma := \frac{\kappa - \left(\sqrt{\frac{2}{\pi}} N \cdot \min_{i\in\Lambda} \sigma_i\right)^{-1} \cdot \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty}}{\kappa}. \tag{3.11}$$

This theorem gives a characterization of those index sets Λ that can be recovered with high probability. As expected, there are similarities with the worst case: we see that the main requirement embodied by (3.10) is that the approximation error be sufficiently small compared to a measure of correlations of atoms on the support and correlations of the support with the rest of the dictionary. However, observe that these correlations are now measured using the 2-Babel function and that we are basically asking that:

$$\mu_2^{in}(\Lambda) + \mu_2(\Lambda) < 1 - \delta_\Lambda.$$

If that is the case, and the average approximation error per channel $N^{-1} \cdot \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty}$ is small enough, then the probability that 1-SOMP fails to recover Λ becomes increasingly smaller as the number of channels grows. It might be more convenient to state a condition

on the dictionary as a whole, and not on a given support. If the dictionary satisfies a uniform uncertainty principle [5], that is to say if the S-restricted isometry constants δ_S are small, the following result shows that the probability that 1-SOMP fails to recover any support of size S decays exponentially fast with the number of channels.

Theorem 3.5 (Average case analysis of 1-SOMP)**.** *Let $p = 1$ and $S = |\Lambda|$. Assume that the dictionary $\mathbf{\Phi}$ obeys a uniform uncertainty principle with S-restricted isometry constants $\delta_{S+1} < 1/3$ and*

$$\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} < \sqrt{\frac{2}{\pi}} N \cdot \min_{i\in\Lambda} \sigma_i \cdot (1 - 3\delta_{S+1})\,. \tag{3.12}$$

Then the probability that S steps of 1-SOMP with $\mathbf{\Psi} = \mathbf{\Phi}$ fail to exactly recover the support Λ does not exceed $K \cdot 2^S \cdot \exp(-N\gamma^2/\pi)$ with K the number of atoms in $\mathbf{\Phi}$ and

$$\gamma := 1 - 3\delta_{S+1} - \big(\sqrt{\frac{2}{\pi}} N \cdot \min_{i\in\Lambda} \sigma_i\big)^{-1} \cdot \|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty}. \tag{3.13}$$

The previous result provides a quantitative average case analysis of multi-channel OMP based on the restricted isometry constants δ_S alone. Together with the condition (2.17) for random Gaussian or Bernoulli matrices to have small δ_S it therefore gives a theoretical explanation to numerical results in the context of distributed compressed sensing conducted in [3].

Note that because of the term 2^S in the probability bound above, which also appears in Theorem 3.4, the required number of channels must be quite high, typically $N \approx S$. Getting rid of this factor would therefore be highly desirable, but the technique we used to prove the theorems does not seem to be easily adaptable to do so, and it remains an open question whether this can be done at all.

In practice, computing the S-restricted isometry constant of $\mathbf{\Phi}$ is a daunting task. Fortunately, when $\mathbf{\Phi}$ is a tight frame and for any support of size at most S selected at random, our last result shows that the behaviour of 1-SOMP is essentially controlled by the 2-Babel function.

Theorem 3.6. *Assume $\mathbf{\Phi}$ to be a tight frame. Let $Y = \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U$ with U a $S \times N$ matrix of standard Gaussian random variables and Λ drawn at random among all supports of size at most S. Assume that $\mu_2(S) < 1/3$ and*

$$\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} < \sqrt{\frac{2}{\pi}} N \cdot \min_{i\in\Lambda} \sigma_i \cdot (1 - 3\mu_2(S)) \qquad \text{and} \qquad S < d/37. \tag{3.14}$$

Then the probability that S steps of 1-OMP with $\mathbf{\Psi} = \mathbf{\Phi}$ fail to exactly recover the support Λ does not exceed $K \cdot 2^S \cdot \exp(-N\gamma^2/\pi) + 2\exp(-\tilde{\gamma}^2)$ with

$$\gamma = 0.9 \left(1 - 3\mu_2(S) - \big(\sqrt{\frac{2}{\pi}} N \cdot \min_{i\in\Lambda} \sigma_i\big)^{-1} \cdot \|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty}\right).$$

and $\tilde{\gamma} = (\frac{1}{37} - \frac{S}{d})/(\mu\sqrt{S})$.

Before proceeding to the technical core of this paper, let us synthesise our findings using the Dirac-DCT dictionary introduce above. Since in that case we have $\mu_q(S) = S^{1/q}\sqrt{2/d}$, for $q = 1, 2$, worst case analysis tell us that both p-thresholding and p-SOMP can recover supports of size $S \approx \sqrt{d}$. For 1-thresholding however, average case analysis when all Gaussian coefficient have equal variances asserts that the probability of recovering supports of size $S \approx d$ gets overwhelmingly large as the number of channels grows. We reach the same conclusion for 1-SOMP by inspecting equation (3.14). Average case analysis confirms a large body of experiments that illustrate the effectiveness of simultaneous approximations with greedy algorithms. In particular, strong hypotheses on either the size of Λ or the incoherence of the dictionary are relaxed. Note, though, that for both p-thresholding or p-SOMP our bounds require a large number of channels to be effective. It is not absolutely clear, as of this writing, whether that is an inherent limit of the algorithms or an artefact of our proofs and more experimental results are needed to draw a decisive conclusion.

4 Worst Case Analysis

In this section we develop conditions that ensure recovery of all signals with a certain support set Λ. Our main contribution is an extension of existing results to the case where noise is present on the signal. In contrast to the expository Section 3 we now work with a sensing matrix $\mathbf{\Psi}$ (possibly different from $\mathbf{\Phi}$) and a general $p \in [1, \infty]$ to measure multichannel correlations. We will need some assumptions on $\{\mathcal{X}_p^{(m)}\}_{m=1}^{|\Lambda|}$, a non-increasing rearrangement of the row p-norms $\mathcal{X}_p^k$, $k \in \Lambda$ of the signal coefficients X. The shorthands $\mu_p(\Lambda)$ and $\mu_p^{in}(\Lambda)$ will respectively denote $\mu_p(\mathbf{\Psi}, \mathbf{\Phi}, \Lambda)$ and $\mu_p^{in}(\mathbf{\Psi}, \mathbf{\Phi}, \Lambda)$.

Theorem 4.1 (**Worst case recovery with p-thresholding.**). *If*

$$\|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} < \min_{i\in\Lambda} \left\{\mathcal{X}_p^i \cdot |\langle \psi_i, \varphi_i\rangle|\right\} - \max_{k\in\Lambda} \left\{\mathcal{X}_p^k \cdot \left(\mu_1(\Lambda) + \mu_1^{in}(\Lambda)\right)\right\} \quad (4.1)$$

then p-thresholding recovers the support set Λ from $Y = \mathbf{\Phi}_\Lambda X + E$. Moreover, the reconstructed coefficients $\tilde{X}$ satisfy

$$\|X - \tilde{X}\|_{\infty,2} \leq \|\mathbf{\Phi}_\Lambda^\dagger E\|_{\infty,2} \leq \left(1 + \mu_1^{in}(\mathbf{\Phi}, \mathbf{\Phi}, \Lambda)\right) \cdot \|E\|_{\infty,2}.$$

Note: *the latter inequality involves $\mu_1^{in}(\mathbf{\Phi}, \mathbf{\Phi}, \Lambda)$ and* not *$\mu_1^{in}(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda)$.*

Proof 1. *Denoting $\mathbf{B} := diag\left(\langle \psi_k, \varphi_k\rangle\right)_{k\in\Lambda}$, observe that $\|\psi_i^\star \mathbf{\Phi}_\Lambda X\|_p$ is the p-norm of the i-th row of $\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda X = \mathbf{B}X + (\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda - \mathbf{B})X$. Since the p-norm of the i-th row of $\mathbf{B}X$ is $|\langle \psi_i, \varphi_i\rangle| \cdot \mathcal{X}_p^i$ we get*

$$\|\psi_i^\star \mathbf{\Phi}_\Lambda X\|_p \geq |\langle \psi_i, \varphi_i\rangle| \cdot \mathcal{X}_p^i - \|(\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda - \mathbf{B})X\|_{p,\infty}.$$

Therefore, the recovery condition (3.1) *is satisfied whenever*

$$\|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} < \min_{i\in\Lambda} \left\{\langle \psi_i, \varphi_i\rangle| \cdot \mathcal{X}_p^i\right\} - \|(\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda - \mathbf{B})X\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda X\|_{p,\infty}. \quad (4.2)$$

To conclude, we use Corollary 2.2 to estimate

$$\begin{aligned} \|(\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda - \mathbf{B})X\|_{p,\infty} \;+\; & \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda X\|_{p,\infty} \leq \left(\|\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda - \mathbf{B}\|_{1,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda\|_{1,\infty}\right) \cdot \|X\|_{p,\infty} \\ \leq\; & \left(\sup_{k\in\Lambda} \sum_{j\in\Lambda\setminus\{k\}} |\langle \psi_k, \varphi_j\rangle| + \sup_{k\notin\Lambda} \sum_{j\in\Lambda} |\langle \psi_k, \varphi_j\rangle|\right) \cdot \|X\|_{p,\infty} \end{aligned}$$

and identify with the definitions of $\mu_1^{in}(\Lambda)$ *and* $\mu_1(\Lambda)$*. For the claim on the error of the reconstructed coefficients we note that* $\tilde{X} = \mathbf{\Phi}_\Lambda^\dagger(\mathbf{\Phi}_\Lambda X + E) = X + \mathbf{\Phi}_\Lambda^\dagger E$*. Moreover,* $\|\mathbf{\Phi}_\Lambda^\dagger\| \leq 1 + \mu_1^{in}(\mathbf{\Phi}, \mathbf{\Phi}, \Lambda)$*, see for instance [25, Proposition 4.3] or [9]. This completes the proof.*

The success of p-thresholding is thus governed by the condition that the noise level should be smaller than a threshold determined both by the dynamic range of the coefficients $\mathcal{X}_p^i$ and by the sum of correlations among atoms on the support as well as between the support and the remaining of $\mathbf{\Phi}$. The conditions on the correlations between the sensing and synthesis dictionaries are expressed in terms of the cumulative coherence and are very reminiscent of Tropp's recovery condition [24]. These conditions are based on worst case analysis and are fairly restrictive. The cumulative coherence in particular is an ℓ_1 norm and can be very big even for reasonably small Λ. In the next sections, we develop an average case analysis of p-thresholding and show that the *typical* recovery conditions are much less restrictive.

Theorem 4.2. Worst case recovery with p-SOMP *Assume that, for the support set* Λ*, the sensing matrix and the dictionary matrix are such that* $\mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_\Lambda^\star$ *is invertible and*

$$\sup_{k \notin \Lambda} \|(\mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_\Lambda)^{-1} \mathbf{\Phi}_\Lambda^\star \psi_k\|_1 < 1. \tag{4.3}$$

Consider a multichannel signal $Y = \mathbf{\Phi}_\Lambda X + E$ *and suppose that* $M \leq |\Lambda|$ *satisfies*

$$\|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} + \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} < \mathcal{X}_p^{(M)} \cdot \left(1 - \sup_{k \notin \Lambda} \|(\mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_\Lambda)^{-1} \mathbf{\Phi}_\Lambda^\star \psi_k\|_1\right) \cdot \|(\mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_\Lambda)^{-1}\|_{1\to 1}^{-1}. \tag{4.4}$$

Then the first M *steps of p-OMP recover distinct elements of the support* Λ*. If* (4.4) *is valid for* $M = |\Lambda|$ *then in addition the reconstructed coefficients* $\tilde{X}$ *satisfy* $\|X - \tilde{X}\|_{\infty,2} \leq (1 + \mu_1^{in}(\mathbf{\Phi}, \mathbf{\Phi}, \Lambda)) \cdot \|E\|_{\infty,2}$*.*

Proof 2. *We will proceed by induction. Suppose we have performed* M *iterations successfully, i.e.,* $\Lambda_M \subset \Lambda$ *(this assumption is clearly true for* $M = 0$ *since* $\Lambda_0 = \emptyset$ *when no iteration of SOMP has been performed yet) and, with only a slight abuse of notations, let* $\mathbf{\Phi} X_M = \mathbf{\Phi}_\Lambda X_M$ *be an approximant of* Y *generated by SOMP after* M *iterations, i.e.,* $X_M = \mathbf{\Phi}_{\Lambda_M}^\dagger Y$ *on its support* Λ_M *and zero outside. Further, let* $Y_M = \mathbf{Q}_{\Lambda_M} Y = Y - \mathbf{\Phi} X_M$ *be the associated residual. If* $M = |\Lambda|$ *there is nothing to prove, so we consider the case* $M < |\Lambda|$*. The next selected atom is in* Λ *as soon as* $\|\mathbf{\Psi}_\Lambda^\star Y_M\|_{p,\infty} > \|\mathbf{\Psi}_{\overline{\Lambda}}^\star Y_M\|_{p,\infty}$*. Decomposing the residual, we just need*

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda (X - X_M) + \mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} > \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda (X - X_M) + \mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty}.$$

Using triangle inequalities and rearranging we get the stronger condition

$$\|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} + \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} < \|\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda (X - X_M)\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda (X - X_M)\|_{p,\infty}. \tag{4.5}$$

From Corollary 2.2 we have $\|X - X_M\|_{p,\infty} \leq \|(\mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_\Lambda)^{-1}\|_{1\to 1} \cdot \|\mathbf{\Psi}_\Lambda^\star \mathbf{\Phi}_\Lambda (X - X_M)\|_{p,\infty}$*, and using the fact*[2] *that* X_M *has at most* M *nonzero entries, we also get* $\|X - X_M\|_{p,\infty} \geq$

[2] see also [10, Lemma 4.4]

$\mathcal{X}^{(M+1)}$. *Combining these facts with an estimate due to Tropp [24, 28] and Chen and Huo [6] (which is also recovered using Corollary 2.2)*

$$\frac{\|\mathbf{\Psi}_{\overline{\Lambda}}^{\star}\mathbf{\Phi}_{\Lambda}(X-X_M)\|_{p,\infty}}{\|\mathbf{\Psi}_{\Lambda}^{\star}\mathbf{\Phi}_{\Lambda}(X-X_M)\|_{p,\infty}} \leq \sup_{Z} \frac{\|\mathbf{\Psi}_{\overline{\Lambda}}^{\star}\mathbf{\Phi}_{\Lambda}(\mathbf{\Psi}_{\Lambda}^{\star}\mathbf{\Phi}_{\Lambda})^{-1}Z\|_{p,\infty}}{\|Z\|_{p,\infty}} = \sup_{k\notin\Lambda} \|(\mathbf{\Phi}_{\Lambda}^{\star}\mathbf{\Psi}_{\Lambda})^{-1}\mathbf{\Phi}_{\Lambda}^{\star}\psi_k\|_1 \quad (4.6)$$

shows that the r.h.s in (4.5) *is lower bounded by*

$$\Big(1-\sup_{k\notin\Lambda}\|(\mathbf{\Phi}_{\Lambda}^{\star}\mathbf{\Psi}_{\Lambda})^{-1}\mathbf{\Phi}_{\Lambda}^{\star}\psi_k\|_1\Big)\cdot\|\!|(\mathbf{\Phi}_{\Lambda}^{\star}\mathbf{\Psi}_{\Lambda})^{-1}|\!\|^{-1}\cdot\mathcal{X}^{(M+1)}$$

which yields the sufficient condition (4.4). *The statement on the approximation error of the reconstructed coefficients* $\tilde{X}$ *is shown in the same way as in the proof of Theorem 4.1.*

Standard techniques based on von Neumann series, see e.g. [24, 12] can be used to prove that

$$\Big(1-\sup_{k\notin\Lambda}\|(\mathbf{\Phi}_{\Lambda}^{\star}\mathbf{\Psi}_{\Lambda})^{-1}\mathbf{\Phi}_{\Lambda}^{\star}\psi_k\|_1\Big)\cdot\|\!|(\mathbf{\Phi}_{\Lambda}^{\star}\mathbf{\Psi}_{\Lambda})^{-1}|\!\|^{-1} \leq 1-\mu_1(\Lambda)+\mu_1^{in}(\Lambda).$$

This enables us to obtain Theorem 3.3 as a corollary of Theorem 4.2, since the main assumption (3.9) of Theorem 3.3 will imply both that (4.4) is satisfied for $M=|\Lambda|$ and that (4.3) holds true.

5 Average case analysis for thresholding

In this section we will study the average performances of simultaneous p-thresholding. Our goal, as announced in Section 4, is to show that under the multichannel Gaussian signal model $X=\Sigma^{\frac{1}{2}}U$, the *typical* behaviour of the algorithm is much better than in the worst case. More precisely, we will prove that the probability that p-thresholding fails to identify a sparse superposition of atoms decays exponentially with the number of channels. Interestingly, the hypotheses under which our result holds are reminiscent of the worst case conditions (4.1) but involve switching from the usual cumulative coherence μ_1 to the milder 2-cumulative coherence μ_2.

5.1 Spirit of the proof

Let us first streamline our reasoning so the busy or lazy readers can get enough insight and intuition to go directly to Theorem 5.2, which can be simplified to get Featured Theorem 3.2, and skip its proof. If we want thresholding to succeed we need to show that

$$\min_{i\in\Lambda}\|\psi_i^{\star}\mathbf{\Phi}_{\Lambda}\Sigma^{\frac{1}{2}}U\|_p - \max_{\ell\in\overline{\Lambda}}\|\psi_{\ell}^{\star}\mathbf{\Phi}_{\Lambda}\Sigma^{\frac{1}{2}}U\|_p > \|\mathbf{\Psi}_{\Lambda}^{\star}E\|_{p,\infty}+\|\mathbf{\Psi}_{\overline{\Lambda}}^{\star}E\|_{p,\infty}.$$

The main idea of the proof is based on concentration of measure appearing when the number of channels N is sufficiently large. Then for each p-correlation of the noiseless multichannel signal with a sensing atom we have with very large probability

$$\|\psi_j^{\star}\mathbf{\Phi}_{\Lambda}\Sigma^{\frac{1}{2}}U\|_p \approx C_p(N)\cdot\|\psi_j^{\star}\mathbf{\Phi}_{\Lambda}\Sigma^{\frac{1}{2}}\|_2,$$

	$p=1$	$p=2$	$p=\infty$
$C_p(N)$	$\sqrt{\frac{2}{\pi}N}$	$\sqrt{2}\frac{\Gamma(N/2)}{\Gamma((N-1)/2)} \sim \sqrt{N}$	$\asymp \sqrt{\log(N)}$
$A_p(N)$	$\frac{N}{\pi}$	$\frac{\Gamma^2(N/2)}{\Gamma^2((N-1)/2)} \sim N/2$	$\asymp \log(N)$

Table 1: Constants $A_p(N)$ and $C_p(N)$, the computations can be found Appendix B

where $C_p(N)$ grows with N. Therefore the recovery condition will be satisfied with high probability as long as

$$\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2 - \max_{\ell\notin\Lambda} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2 \gtrsim \frac{\|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} + \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty}}{C_p(N)},$$

and all we need to check is under which conditions on the dictionary and the coefficient ranges the left hand side in the above is large enough.
The next section will supply us with tools to estimate the typicality and precision of the approximation $\|\psi_j^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_p \approx C_p(N)\cdot \|\psi_j^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2$ in order to give a fully detailed proof.

5.2 Concentration of measure

As mentioned above the corner stone on which both the average case analyses of thresholding and OMP rely are the following concentration of measure inequalities. Their actual proofs in all gory mathematical detail are awaiting the interested reader in Appendix A.

Theorem 5.1. *Let U be an $N\times S$ matrix with independent standard Gaussian entries, and $\{v_k\}_{k\in\Omega}\subset\mathbb{R}^S$ a finite family of nonzero vectors. Then for $\varepsilon_1>0$ and $0<\varepsilon_2<1$,*

$$\mathbb{P}\Big(\|v_k^\star U\|_p \geq (1+\varepsilon_1)C_p(N)\|v_k\|_2\Big) \leq \exp(-\varepsilon_1^2 A_p(N)) \tag{5.1}$$

$$\mathbb{P}\Big(\|v_k^\star U\|_p \leq (1-\varepsilon_2)C_p(N)\|v_k\|_2\Big) \leq \exp(-\varepsilon_2^2 A_p(N)) \tag{5.2}$$

for each vector v_k, and

$$\mathbb{P}\Big(\max_{k\in\Omega}\|v_k^\star U\|_p \geq (1+\varepsilon_1)C_p(N)\max_{k\in\Omega}\|v_k\|_2\Big) \leq |\Omega|\cdot\exp(-\varepsilon_1^2 A_p(N)) \tag{5.3}$$

$$\mathbb{P}\Big(\max_{k\in\Omega}\|v_k^\star U\|_p \leq (1-\varepsilon_2)C_p(N)\max_{k\in\Omega}\|v_k\|_2\Big) \leq \exp(-\varepsilon_2^2 A_p(N)) \tag{5.4}$$

$$\mathbb{P}\Big(\min_{k\in\Omega}\|v_k^\star U\|_p \geq (1+\varepsilon_1)C_p(N)\min_{k\in\Omega}\|v_k\|_2\Big) \leq \exp(-\varepsilon_1^2 A_p(N))$$

$$\mathbb{P}\Big(\min_{k\in\Omega}\|v_k^\star U\|_p \leq (1-\varepsilon_2)C_p(N)\min_{k\in\Omega}\|v_k\|_2\Big) \leq |\Omega|\cdot\exp(-\varepsilon_2^2 A_p(N)). \tag{5.5}$$

5.3 Main result for p-thresholding

To keep the notational mess in the proof to a minimum we use the following abbreviations. We capture all the noise related terms in

$$\eta := \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} + \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty}, \tag{5.6}$$

and to deal with the coefficients more efficiently we use for the minimal and maximal entry in $\Sigma = \mathrm{diag}(\sigma_i^2)_{i\in\Lambda}$

$$\sigma_{\min} := \min_{i\in\Lambda} \sigma_i \quad \text{and} \quad \sigma_{\max} := \max_{i\in\Lambda} \sigma_i.$$

Theorem 5.2. *Assume that the noise level η is sufficiently small, i.e*

$$\eta < C_p(N) \cdot \big(\beta \cdot \sigma_{\min} - \mu_2(\Lambda) \cdot \sigma_{\max}\big). \tag{5.7}$$

Then, under the multichannel Gaussian signal model $X = \Sigma^{\frac{1}{2}} U$, the probability that p-thresholding fails to recover the indices of the atoms in Λ does not exceed

$$\mathbb{P}(p - \textit{thresholding fails}) \leq K \cdot \exp\big(-A_p(N) \cdot \gamma^2\big)$$

with

$$\gamma := \frac{\beta \cdot \sigma_{\min} - \mu_2(\Lambda) \cdot \sigma_{\max} - \eta/C_p(N)}{\beta \cdot \sigma_{\min} + \mu_2(\Lambda) \cdot \sigma_{\max}} \tag{5.8}$$

Proof 3. *We can bound the probability that thresholding fails with the following trick,*

$$\begin{aligned}
&\mathbb{P}\big(\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_p - \max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_p \leq \eta\big) \\
&\leq \mathbb{P}\big(\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_p \leq C\big) + \mathbb{P}\big(\max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_p \geq C - \eta\big).
\end{aligned}$$

Motivated by the concentration of measure results we set

$$C = (1-\varepsilon_1) \cdot C_p(N) \cdot \min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2,$$

where we choose ε_1 later. Using (5.5) *we can bound the first probability in the above as:*

$$\mathbb{P}\big(\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_p \leq (1-\varepsilon_1) \cdot C_p(N) \cdot \min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2\big) \leq |\Lambda| \cdot \exp\big(-A_p(N) \cdot \varepsilon_1^2\big).$$

To bound the second probability we have to work a little bit more before applying (5.3).

$$\begin{aligned}
&\mathbb{P}\big(\max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_p \geq C - \eta\big) \\
&= \mathbb{P}\big(\max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_p \geq \underbrace{\frac{C-\eta}{C_p(N) \cdot \max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2}}_{=:1+\varepsilon_2} \cdot C_p(N) \cdot \max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2\big) \\
&\leq |\overline{\Lambda}| \cdot \exp\big(-A_p(N) \cdot \varepsilon_2^2\big).
\end{aligned}$$

For the last equality to hold we need to make sure that $\varepsilon_2 > 0$. We will do this by adjusting the choice of ε_1 so that $\varepsilon_2 = \varepsilon_1$,

$$\varepsilon_2 = \frac{(1-\varepsilon_1) \cdot C_p(N) \cdot \min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2 - \eta}{C_p(N) \cdot \max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2} - 1 = \varepsilon_1.$$

Solving the equation above for ε_1 we get

$$\varepsilon_1 := \frac{\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2 - \max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2 - \eta/C_p(N)}{\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2 + \max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2}. \tag{5.9}$$

To see that $\varepsilon_1 > 0$ observe that

$$\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2^2 = \min_{i\in\Lambda} \sum_{k\in\Lambda} |\sigma_k|^2 |\langle \varphi_k, \psi_i\rangle|^2 \geq \sigma_{\min}^2 \cdot \min_{i\in\Lambda} (|\langle \psi_i, \varphi_i\rangle|^2 + \|\mathbf{\Phi}_{\Lambda/i}^\star \psi_i\|_2^2) \geq \sigma_{\min}^2 \cdot \beta^2$$

$$\max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_2^2 = \max_{\ell\in\overline{\Lambda}} \sum_{k\in\Lambda} |\sigma_k|^2 |\langle \varphi_k, \psi_\ell\rangle|^2 \leq \sigma_{\max}^2 \cdot \max_{\ell\in\overline{\Lambda}} \sum_{k\in\Lambda} |\sigma_k|^2 |\langle \varphi_k, \psi_\ell\rangle|^2 \leq \sigma_{\max}^2 \cdot \mu_2^2(\Lambda).$$

Thus we can estimate ε_1 from below as,

$$\varepsilon_1 > \frac{\beta \cdot \sigma_{\min} - \mu_2(\Lambda) \cdot \sigma_{\max} - \eta/C_p(N)}{\beta \cdot \sigma_{\min} + \mu_2(\Lambda) \cdot \sigma_{\max}} =: \gamma. \tag{5.10}$$

This is larger than zero by condition (5.7) *and we get as final bound for the probability that thresholding fails,*

$$\mathbb{P}(p - \textit{thresholding fails}) \leq K \cdot \exp\big(-A_p(N) \cdot \varepsilon_1^2\big) \leq K \cdot \exp\big(-A_p(N) \cdot \gamma^2\big).$$

To get from the above theorem to Featured Theorem 3.2 we need to insert the expression for η and the concrete values for $C_p(N), A_p(N)$ for $p = 1$ and observe that because $\mu_2(\Lambda) \leq \mu_2(S)$ we can use it instead in the above formulas.

6 Average case analysis of OMP

In the previous section we have seen that even in the average case thresholding requires balanced coefficients in order to ensure viable recovery results. This is quite a strong limitation. Motivated by the fact that in the worst case OMP enabled us to overcome this restriction we will now analyse the average performance of OMP.

6.1 Spirit of the proof

A sufficient condition for OMP to succeed is that it will always pick another component in the support, whatever residual $R_J = \mathbf{Q}_J Y = (I - \mathbf{P}_J)(\mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U + E)$ we have. So for all $J \subset \Lambda$ we want to ensure

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} > \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J E\|_{p,\infty}. \tag{6.1}$$

Concentration of measure tells us that for any matrix A we have with very high probability

$$\|AU\|_{p,\infty} \approx C_p(N) \cdot \|A\|_{2,\infty}.$$

Therefore, condition (6.1) should be satisfied with high probability as long as

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} > \frac{\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J E\|_{p,\infty}}{C_p(N)}. \tag{6.2}$$

To ensure the condition above we need to find a lower bound for the left hand side that does not depend on J itself but only on its size.
The first term on the left hand side in (6.2) can be estimated from below as

$$\begin{aligned}\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty}^2 &= \sup_{i\in\Lambda} \sum_{k\in\Lambda} \sigma_k^2 \cdot |\langle \mathbf{Q}_J\varphi_k, \psi_i\rangle|^2 \\ &\geq \sup_{i\in\Lambda\setminus J} \sigma_i^2 \cdot |\langle \mathbf{Q}_J\varphi_i, \psi_i\rangle|^2 \geq \sup_{i\in\Lambda\setminus J} \sigma_i^2 \cdot \inf_{i\in\Lambda\setminus J} |\langle \mathbf{Q}_J\varphi_i, \psi_i\rangle|^2.\end{aligned}$$

Using $\mathbf{Q}_J\varphi_i = 0$ whenever $i \in J$, the second term can be estimated from above as

$$\begin{aligned}\|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty}^2 &= \sup_{\ell\notin\Lambda} \sum_{i\in\Lambda} \sigma_i^2 \cdot |\langle \mathbf{Q}_J\varphi_i, \psi_\ell\rangle|^2 \\ &= \sup_{\ell\notin\Lambda} \sum_{i\in\Lambda\setminus J} \sigma_i^2 \cdot |\langle \mathbf{Q}_J\varphi_i, \psi_\ell\rangle|^2 \leq \sup_{i\in\Lambda\setminus J} \sigma_i^2 \cdot \sup_{\ell\notin\Lambda} \sum_{i\in\Lambda\setminus J} |\langle \mathbf{Q}_J\varphi_i, \psi_\ell\rangle|^2 \\ &\leq \sup_{i\in\Lambda\setminus J} \sigma_i^2 \cdot \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_{\Lambda\setminus J}\|_{2,\infty}^2.\end{aligned}$$

The combination of these two bounds leads to

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} > \sup_{i\in\Lambda\setminus J} \sigma_i^2 \cdot \big(\inf_{i\in\Lambda\setminus J} |\langle \mathbf{Q}_J\varphi_i, \psi_i\rangle|^2 - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_{\Lambda\setminus J}\|_{2,\infty}^2 \big).$$

Now observe that if we denote with $\{\sigma^{(i)}\}_{i=1}^{|\Lambda|}$ the decreasing rearrangement of σ_i we have $\sup_{i\in\Lambda\setminus J} \sigma_i \geq \sigma^{(M)}$ for J of size at most $M-1$. Therefore defining the two constants

$$c_0(\Lambda) = \inf_{J\subsetneq\Lambda} \inf_{i\in\Lambda\setminus J} |\langle \mathbf{Q}_J\varphi_i, \psi_i\rangle|, \quad \text{and} \quad d_0(\Lambda) = \sup_{J\subsetneq\Lambda} \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_{\Lambda\setminus J}\|_{2,\infty} \tag{6.3}$$

we can finally lower bound the left hand side in (6.2) as

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} > \sigma^{(M)} \cdot \big(c_0(\Lambda) - d_0(\Lambda)\big).$$

Based on the bounds $c_0(\Lambda), d_0(\Lambda)$ we can now formulate a general recovery result.

6.2 A general recovery result

Theorem 6.1. *Assume that the noise is orthogonal to all the atoms in the support,* $\mathbf{\Phi}_\Lambda^\star E = 0$*, and that the noise level* η *is sufficiently small, i.e*

$$\eta < \big(c_0(\Lambda) - d_0(\Lambda)\big) \cdot C_p(N) \cdot \sigma^{(M)}. \tag{6.4}$$

Then, under the multichannel Gaussian signal model $X = \Sigma^{\frac{1}{2}} U$*, the probability that one of the first* M *atoms selected by p-OMP is incorrect (not in* Λ*) does not exceed*

$$\mathbb{P}(\textit{p-OMP fails after at most } M \textit{ steps}) \leq (1 + |\overline{\Lambda}|) \cdot \mathcal{C}_M \cdot \exp\big(-A_p(N) \cdot \gamma_M^2\big) \tag{6.5}$$

with $\mathcal{C}_M := \sum_{m=0}^{M-1} \binom{|\Lambda|}{m}$ *and*

$$\gamma_M := \frac{c_0(\Lambda) - d_0(\Lambda) - \eta \cdot \big(C_p(N) \cdot \sigma^{(M)}\big)^{-1}}{c_0(\Lambda) + d_0(\Lambda)}$$

Proof 4. *We have to show that for any subset J of size at most $M-1$ equation* (6.1) *holds. However since we assume that the noise is orthogonal to the span of the support we have $\mathbf{Q}_J E = E - \mathbf{P}_J E = E$ and so it suffices to show that*

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} > \|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} = \eta.$$

We can bound the probability that the above condition is violated using the same tricks as before for thresholding. Again we collect all the noise terms on the right hand side in η.

$$\begin{aligned}&\mathbb{P}\big(\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} < \eta\big) = \\ &= \mathbb{P}\big(\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} < C\big) + \mathbb{P}\big(\|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} > C - \eta\big).\end{aligned}$$

We choose $C = (1-\varepsilon_1)\cdot C_p(N)\cdot |\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty}$ and use concentration inequality (5.4) *to bound the first probability as*

$$\mathbb{P}\big(\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} < (1-\varepsilon_1)\cdot C_p(N)\cdot \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty}\big) \le \exp\big(-A_p(N)\cdot \varepsilon_1^2\big).$$

To bound the second probability we proceed as for thresholding and use inequality (5.3)*,*

$$\begin{aligned}&\mathbb{P}\big(\|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} > C - \eta\big) = \\ &= \mathbb{P}\Big(\|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} > \underbrace{\frac{C-\eta}{C_p(N)\cdot \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty}}}_{=:1+\varepsilon_2}\cdot C_p(N)\cdot \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty}\Big) \\ &\le |\overline{\Lambda}|\cdot \exp\big(-A_p(N)\cdot \varepsilon_2^2\big).\end{aligned}$$

Again we require $\epsilon_1 = \epsilon_2$,

$$\varepsilon_2 = \frac{(1-\varepsilon_1)\cdot \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} - \eta/C_p(N)}{\|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty}} - 1 = \varepsilon_1.$$

Solving the above for ε_1 we get

$$\varepsilon_1 = \frac{\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} - \eta/C_p(N)}{\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty}}.$$

If we now insert the definition of $c_0(\Lambda), d_0(\Lambda)$ from (6.3) *we can estimate ε_1 from below as:*

$$\varepsilon_1 > \frac{c_0(\Lambda) - d_0(\Lambda) - \eta\cdot \big(C_p(N)\cdot \sigma^{(M)}\big)^{-1}}{c_0(\Lambda) + d_0(\Lambda)} = \gamma_M > 0$$

Condition (6.4) *ensures that $\gamma_M > 0$ and so we can bound for any subset J of size at most $M-1$ the probability that OMP fails to pick another good atom as*

$$\mathbb{P}(\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} > \eta) < (1+|\overline{\Lambda}|)\cdot \exp\big(-A_p(N)\cdot \gamma_M^2\big).$$

In the end to be independent of the sequence of subsets that OMP finds we use a union bound over all $\mathcal{C}_M := \sum_{m=0}^{M-1} \binom{|\Lambda|}{m}$ subsets $J \subset \Lambda$ of size at most $M-1$ to get the upper estimate on the probability of failure in (6.5)*.*

Note that the union bound we take above leads to a constant $\mathcal{C}_S = 2^S$ if we want to estimate recovering the whole support. This is a considerable factor, for which there is no numerical evidence in either our simulations or the results in [3]. One of our future goals therefore is to improve the probability estimate by finding a way around taking the crude union bound.
Also note that in the proof instead of estimating ε_1 in terms of $c_0(\Lambda), d_0(\Lambda)$ we could have used any other pair of constants c, d satisfying $c \le c_0(\Lambda)$ and $d \ge d_0(\Lambda)$. While these constants result in a smaller γ_M and a stronger restriction on the noise level they may have the advantage of having a more tangible form than the original ones. Thus the next subsection is dedicated to finding new constants c, d in terms of properties of the dictionary, which lead directly to the results in the featured theorems in Section 3 when used instead of $c_0(\Lambda), d_0(\Lambda)$.

6.3 Bounds on $c_0(\Lambda)$ and $d_0(\Lambda)$

The following results estimate constants in terms of the 2-cross-Babel function $\mu_2(\Lambda) = \mu_2(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda)$, the similarity β and the (local) restricted isometry constants $\delta_\Lambda = \delta_\Lambda(\mathbf{\Phi})$.

Lemma 6.2. *For any subset $J \subsetneq \Lambda$*

$$\begin{aligned}
\|\!|\mathbf{\Phi}_J^\dagger|\!\| &\le (1-\delta_J)^{-\frac{1}{2}} \le (1-\delta_\Lambda)^{-\frac{1}{2}}, \\
\|\!|(\mathbf{\Phi}_J^* \mathbf{\Phi}_J)^{-1}|\!\| &\le (1-\delta_J)^{-1} \le (1-\delta_\Lambda)^{-1}, \\
\|\!|\mathbf{\Phi}_J^* \mathbf{\Phi}_{\Lambda\setminus J}|\!\| &\le \delta_\Lambda, \qquad (6.6) \\
\|\!|\mathbf{\Phi}_J^\dagger \mathbf{\Phi}_{\Lambda\setminus J}|\!\| &\le \frac{\delta_\Lambda}{1-\delta_J} \le \frac{\delta_\Lambda}{1-\delta_\Lambda}, \\
\sup_{\ell\in\overline{\Lambda}} \|\mathbf{\Phi}_J^\star \psi_\ell\|_2 &\le \mu_2(\Lambda), \\
\sup_{\ell\in\overline{\Lambda}} \|\mathbf{\Phi}_{\Lambda\setminus J}^\star \psi_\ell\|_2 &\le \mu_2(\Lambda), \\
\sup_{i\in\Lambda\setminus J} \|\mathbf{\Phi}_J^\star \psi_i\|_2 &\le \mu_2^{in}(\Lambda).
\end{aligned}$$

Proof. All the statements except for (6.6) essentially follow directly from Lemma 2.1 about matrix norms and the definitions of δ_Λ, $\mu_2(\Lambda)$ and $\mu_2^{in}(\Lambda)$ in Section 2. To get to (6.6) note that by definition of the restricted isometry constants we have $\|\!|\mathbf{\Phi}_\Lambda^* \mathbf{\Phi}_\Lambda - \mathbf{I}|\!\| \le \delta_\Lambda$, therefore

$$\begin{aligned}
\|\!|\mathbf{\Phi}_{\Lambda\setminus J}^\star \mathbf{\Phi}_J|\!\|^2 &= \sup_{\|a_J\|_2\le 1} \|\mathbf{\Phi}_{\Lambda\setminus J}^\star \mathbf{\Phi}_J \cdot a_J\|_2^2 \\
&\le \sup_{\|a_J\|_2\le 1} \left(\|\mathbf{\Phi}_{\Lambda\setminus J}^\star \mathbf{\Phi}_J \cdot a_J\|_2^2 + \|(\mathbf{\Phi}_J^\star \mathbf{\Phi}_J - \mathbf{I}) \cdot a_J\|_2^2 \right) \\
&= \sup_{\|a_J\|_2\le 1} \left\| \begin{pmatrix} \mathbf{\Phi}_J^\star \mathbf{\Phi}_J - \mathbf{I} & \mathbf{\Phi}_J^\star \mathbf{\Phi}_{\Lambda\setminus J} \\ \mathbf{\Phi}_{\Lambda\setminus J}^\star \mathbf{\Phi}_J & \mathbf{\Phi}_{\Lambda\setminus J}^\star \mathbf{\Phi}_{\Lambda\setminus J} - \mathbf{I} \end{pmatrix} \begin{pmatrix} a_J \\ 0 \end{pmatrix} \right\|_2^2 \\
&\le \sup_{\|a\|_2\le 1} \|(\mathbf{\Phi}_\Lambda^\star \mathbf{\Phi}_\Lambda - \mathbf{I})a\|_2^2 \le \delta_\Lambda^2.
\end{aligned}$$

□

Lemma 6.3. *Valid bounds for the constants $c_0(\Lambda), d_0(\Lambda)$ are given by*

$$c(\Lambda) := \beta - \frac{\mu_2^{in}(\Lambda)}{\sqrt{1-\delta_\Lambda}}, \quad \text{and} \quad d(\Lambda) := \frac{\mu_2(\Lambda)}{1-\delta_\Lambda}. \tag{6.7}$$

Proof 5. *First we need to show that $c(\Lambda)$ as defined above is smaller than $c_0(\Lambda) = \inf_{J\subsetneq\Lambda}\inf_{i\in\Lambda\setminus J}|\langle \mathbf{Q}_J\varphi_i,\psi_i\rangle|$. Recall the definition of the operator $\mathbf{Q}_J = \mathbf{I}-\mathbf{P}_J$. We write the projection explicitely as $\mathbf{P}_J = (\mathbf{\Phi}_J^\dagger)^\star\mathbf{\Phi}_J^\star = \mathbf{\Phi}_J(\mathbf{\Phi}_J^\star\mathbf{\Phi}_J)^{-1}\mathbf{\Phi}_J^\star$, where $\mathbf{\Phi}_J^\dagger$ denotes the pseudo-inverse of $\mathbf{\Phi}_J$. Fixing $J\subsetneq\Lambda$ for the moment we get (using self-adjointness of $\mathbf{P}_J$)*

$$\begin{aligned}\inf_{i\in\Lambda\setminus J}|\langle \mathbf{Q}_J\varphi_i,\psi_i\rangle| &= \inf_{i\in\Lambda\setminus J}|\langle (\mathbf{I}-\mathbf{P}_J)\varphi_i,\psi_i\rangle| \geq \inf_{i\in\Lambda\setminus J}(|\langle\varphi_i,\psi_i\rangle| - \|\mathbf{P}_J\psi_i\|_2\|\varphi_i\|_2)\\ &\geq \inf_{i\in\Lambda\setminus J}(|\langle\varphi_i,\psi_i\rangle| - \|(\mathbf{\Phi}_J^\dagger)^\star\mathbf{\Phi}_J^\star\psi_i\|_2) \geq \inf_{i\in\Lambda\setminus J}(|\langle\varphi_i,\psi_i\rangle| - |\!|\!|\mathbf{\Phi}_J^\dagger|\!|\!|\,\|\mathbf{\Phi}_J^\star\psi_i\|_2).\end{aligned} \tag{6.8}$$

Using Lemma 6.2 and the fact that $\inf_i|\langle\varphi_i,\psi_i\rangle|\geq\beta$ we obtain

$$\inf_{i\in\Lambda\setminus J}|\langle \mathbf{Q}_J\varphi_i,\psi_i\rangle| \;\geq\; \beta - (1-\delta_\Lambda)^{-\frac{1}{2}}\cdot\mu_2^{in}(\Lambda).$$

Since the term on the right hand side no longer depends on the subset J, the inequation is valid for the infimum over all subsets J, thus leading to the first bound in (6.7).

For the second claim we need to show that $d(\Lambda)$ as defined above is larger than $d_0(\Lambda) = \sup_{J\subsetneq\Lambda}\|\mathbf{\Psi}_\Lambda^\star\mathbf{Q}_J\mathbf{\Phi}_{\Lambda\setminus J}\|_{2,\infty}$. We again start by fixing $J\subsetneq\Lambda$.

$$\begin{aligned}\|\mathbf{\Psi}_\Lambda^\star\mathbf{Q}_J\mathbf{\Phi}_{\Lambda\setminus J}\|_{2,\infty} &= \sup_{\ell\notin\Lambda}\|\mathbf{\Phi}_{\Lambda\setminus J}^\star(\mathbf{I}-\mathbf{P}_J)\psi_\ell\|_2 \leq \sup_{\ell\notin\Lambda}(\|\mathbf{\Phi}_{\Lambda\setminus J}^\star\psi_\ell\|_2 + \|\mathbf{\Phi}_{\Lambda\setminus J}^\star(\mathbf{\Phi}_J^\dagger)^\star\mathbf{\Phi}_J^\star\psi_\ell\|_2)\\ &\leq \sup_{\ell\notin\Lambda}(\|\mathbf{\Phi}_{\Lambda\setminus J}^\star\psi_\ell\|_2 + |\!|\!|\mathbf{\Phi}_J^\dagger\mathbf{\Phi}_{\Lambda\setminus J}|\!|\!|\,\|\mathbf{\Phi}_J^\star\psi_\ell\|_2).\end{aligned} \tag{6.9}$$

Using Lemma 6.2 yields

$$\|\mathbf{\Psi}_\Lambda^\star\mathbf{Q}_J\mathbf{\Phi}_{\Lambda\setminus J}\|_{2,\infty} \;\leq\; \mu_2(\Lambda)\cdot\left(1+\frac{\delta_\Lambda}{1-\delta_\Lambda}\right) = \frac{\mu_2(\Lambda)}{1-\delta_\Lambda}.$$

Again since the bound is independent of the subset J it is valid for the supremum over all subsets and thus leads to the second part of (6.7).

Based on the estimates for $c(\Lambda)$ and $d(\Lambda)$ as they appear above we can now give proofs for the featured theorems in Section 3.

6.4 Proof of Theorem 3.4

All we need to do is replace $c_0(\Lambda), d_0(\Lambda)$ in Theorem 6.1 by the bounds derived in the lemma above. However to make the formulas less ugly we further estimate

$$c_0(\Lambda) \geq \beta - \frac{\mu_2^{in}(\Lambda)}{\sqrt{1-\delta_\Lambda}} \geq \beta - \frac{\mu_2^{in}(\Lambda)}{1-\delta_\Lambda} := c(\Lambda).$$

To finally arrive at Theorem 3.4 simply note that whenever $\mathbf{\Psi} = \mathbf{\Phi}$ we have $\beta = 1$ and because of the assumption that E is orthogonal to the atoms in Λ the noise level reduces to $\eta = \|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty}$.

6.5 Proof of Theorem 3.5

The only missing ingredient we need for this proof is the following lemma, providing further bounds for the constants $c_0(\Lambda), d_0(\Lambda)$ to be used instead in Theorem 6.1.

Lemma 6.4. *Suppose that $\mathbf{\Psi} = \mathbf{\Phi}$, and let S be the cardinality of Λ. Then we can bound $c_0(\Lambda), d_0(\Lambda)$ by*

$$c_S := 1 - \frac{\delta_{S+1}}{\sqrt{1-\delta_S}} \qquad \text{and} \qquad d_S := \frac{\delta_{S+1}}{1-\delta_S}.$$

Proof 6. *We first show that for any S we have $\mu_2(\mathbf{\Phi}, \mathbf{\Phi}, S) \leq \delta_{S+1}$. For $\ell \notin J$ we define $\Lambda = J \cup \{\ell\}$ and obtain from (6.6) that $\|\mathbf{\Phi}_J^\star \varphi_\ell\|_2 = \|\!|\mathbf{\Phi}_J^\star \varphi_\ell|\!\| \leq \delta_{J\cup\{\ell\}}$. Therefore $\mu_2(\mathbf{\Phi}, \mathbf{\Phi}, J) \leq \sup_{\ell \notin J} \delta_{J\cup\{\ell\}}$ and*

$$\mu_2(\mathbf{\Phi}, \mathbf{\Phi}, S) = \sup_{|J|\leq S} \mu_2(\mathbf{\Phi}, \mathbf{\Phi}, J) \leq \sup_{|J|\leq S} \sup_{\ell\notin J} \delta_{J\cup\{\ell\}} = \delta_{S+1}.$$

Combing this estimate with Lemma 6.3 then leads to

$$\begin{aligned} c_0(\Lambda) &\geq 1 - \frac{\mu_2^{in}(\Lambda)}{\sqrt{1-\delta_\Lambda}} \geq 1 - \frac{\mu_2(S)}{\sqrt{1-\delta_S}} \geq 1 - \frac{\delta_{S+1}}{\sqrt{1-\delta_S}}, \\ d_0(\Lambda) &\leq \frac{\mu_2(\Lambda)}{1-\delta_\Lambda} \leq \frac{\mu_2(S)}{1-\delta_S} \leq \frac{\delta_{S+1}}{1-\delta_S}. \end{aligned}$$

Again to prove the theorem we replace $c_0(\Lambda), d_0(\Lambda)$ by c_S, d_S in Theorem 6.1 and then need the noise level η to satisfy

$$\eta \leq C_1(N) \cdot \sigma_{\min} \cdot \big(c_S - d_S\big) = \sqrt{\frac{2}{\pi}} N \cdot \sigma_{\min} \cdot \big(1 - \delta_{S+1} \cdot \frac{\sqrt{1-\delta_S}+1}{1-\delta_S}\big).$$

The above condition is ensured by $\eta < \sqrt{\frac{2}{\pi}} N \cdot \sigma_{\min} \cdot (1 - 3\delta_{S+1})$ since for $\delta_{S+1} < 1/3$ the fraction in the expression above is smaller than 3 (it is always larger than 2) and so by Theorem 6.1 the probability of failure is smaller than

$$(1 + K - S) 2^S \exp(-A_p(N)\gamma_S^2) \qquad \text{with} \qquad \gamma_S = \frac{c_S - d_S - \eta \cdot (\sqrt{\frac{2}{\pi}} N \cdot \sigma_{\min})^{-1}}{c_S + d_S}.$$

Inserting the explicit values for c_S, d_S and $\delta_{S+1} < 1/3$ we get from a lengthy but uninteresting calculation that $\gamma_S > 1 - 3\delta_{S+1} - \eta \cdot (\frac{N}{\pi} \cdot \sigma_{\min})^{-1} = \gamma$. Together with the observation that for $p = 1$ we have $A_p(N) = N/\pi$ this leads to the final bound for failure featured in Theorem 3.5.

$$\mathbb{P}(\text{failure of 1-OMP}) \leq K \cdot 2^S \cdot \exp(-N\gamma^2/\pi).$$

6.6 Proof of Theorem 3.6

In order to prove the second main theorem we need Joel Tropp's result that for a random support set Λ the local isometry constants δ_Λ are well behaved provided the coherence μ is small. The following statement is [27, Theorem B] rewritten.

Theorem 6.5. *Suppose Λ is selected uniformly at random among all subsets of $\{1,\dots,K\}$ of size $S \geq 3$. If $c\delta - \|\mathbf{\Phi}\|^2 S/K > 0$ then*

$$\mathbb{P}\left(\delta_\Lambda > \delta\right) < 2\exp\left(-\left(\frac{c\delta - \|\mathbf{\Phi}\|^2 S/K}{\mu\sqrt{S}}\right)^2\right),$$

where the constant c is not smaller than 0.0818.

With this theorem we can now estimate the probability that 1-OMP fails as:

$$\mathbb{P}(\text{1-OMP } fails) \leq \mathbb{P}(\text{1-OMP } fails|\delta_\Lambda < 1/3) + \mathbb{P}(\delta_\Lambda > 1/3)$$

To estimate the first term on the right hand side we can proceed as before. Because of Lemma 6.3 and $\mu_2(S-1) \leq \mu_2(S)$ we can replace $c_0(\Lambda), d_0(\Lambda)$ by

$$c_S = 1 - \frac{\mu_2(S)}{\sqrt{1-\delta_\Lambda}} \qquad \text{and} \qquad d_S = \frac{\mu_2(S)}{1-\delta_\Lambda}.$$

We then need the noise η to satisfy

$$\eta \leq C_1(N)\cdot\sigma_{\min}\cdot\big(c_S - d_S\big) = \sqrt{\frac{2}{\pi}}N\cdot\sigma_{\min}\cdot\big(1-\mu_2(S)\cdot\frac{\sqrt{1-\delta_\Lambda}+1}{1-\delta_\Lambda}\big),$$

which is again ensured by $\delta_\Lambda < 1/3$ and $\eta < \sqrt{\frac{2}{\pi}}N\cdot\sigma_{\min}\cdot(1-3\mu_2(S))$. Inserting all the values, i.e $\delta_\Lambda < 1/3$ and $\mu_2(S) < 1/3$ (as a consequence of the condition on the noise), into the formula for γ_S leads to the estimate $\gamma_S > 0.9(1-3\mu_2(S) - \eta\cdot(\frac{N}{\pi}\cdot\sigma_{\min})^{-1}) = \gamma$ and we get the bound,

$$\mathbb{P}(\text{1-OMP } fails|\delta_\Lambda < 1/3) \leq K\cdot 2^S\cdot\exp(-N\gamma^2/\pi).$$

Finally to bound the probability that $\mathbb{P}(\delta_\Lambda > 1/3)$ we simply note that $c/3 > 1/37$ and that for a tight frame we have $\|\mathbf{\Phi}\|^2 = K/d$. Thus whenever $S < d/37$ the condition of Theorem 6.5 is satisfied and

$$\mathbb{P}\left(\delta_\Lambda > 1/3\right) < 2\exp\left(-\left(\frac{1/37 - S/d}{\mu\sqrt{S}}\right)^2\right).$$

7 Robustness with respect to the dictionary

In some applications the sparsity inducing dictionary $\mathbf{\Phi}$ might not be known exactly and one actually uses a slightly different dictionary $\mathbf{\Psi}$ in the algorithms instead. This is the case in particular in blind source separation where the equivalent of the dictionary is a mixing matrix [12], which is unknown but estimated from the observed data. The success

of sparsity based blind source separation shows that even an approximate knowledge of the dictionary (the mixing matrix) still makes it possible to obtain a "reasonable" estimate of the coefficients (the sources) [16], but the question arises under which conditions it is possible to robustly recover the support of a signal.

Concerning thresholding one can actually directly apply the results for worst case and average case analysis obtained in Sections 4 and 5. Indeed, one can treat $\mathbf{\Phi}$ and $\mathbf{\Psi}$ in the same way as they were treated there. The fact that we do not know the "synthesis" dictionary $\mathbf{\Phi}$ precisely does not affect the analysis. The only difference is that in the final projection step we use the dictionary $\mathbf{\Psi}_\Lambda$ instead of $\mathbf{\Phi}_\Lambda$. However, this just slightly changes the reconstructed coefficients but not the reconstructed support Λ (see also the statement on the reconstruction error in Theorem 7.1 below).

For an analysis of OMP slightly more effort has to be invested. Indeed, when we update the new residual at each step of OMP we project onto the span of the assumed atoms $(\psi_i)_{i\in J}$ instead of the original atoms $(\varphi_i)_{i\in J}$. Thus, the residual has the form

$$R_J = (\mathbf{I} - \mathbf{\Psi}_J \mathbf{\Psi}_J^\dagger) Y.$$

Hence, we have to work now with $\tilde{\mathbf{Q}}_J = (\mathbf{I} - \mathbf{\Psi}_J \mathbf{\Psi}_J^\dagger)$ instead of $\mathbf{Q}_J = (\mathbf{I} - \mathbf{\Phi}_J \mathbf{\Phi}_J^\dagger)$.

Concerning the worst case analysis, one can simply write

$$Y = \mathbf{\Phi}_\Lambda X = \mathbf{\Psi}_\Lambda X + (\mathbf{\Phi}_\Lambda - \mathbf{\Psi}_\Lambda) X.$$

Setting $E = (\mathbf{\Phi}_\Lambda - \mathbf{\Psi}_\Lambda)X$ and assuming $\|\!|\mathbf{\Phi}_\Lambda - \mathbf{\Psi}_\Lambda|\!\|$ to be small it is straightforward to apply the results of Section 4 also to the robustness problem.

To deal with the average case analysis, the simple trick which consists in considering the modeling error as noise cannot be used directly since our average case analysis of p-SOMP is based on a deterministic model of the noise and *a probabilistic and independent* model of the coefficients X. Here, the noise and the coefficients would be two highly dependent random variables, and the previous results could not be used in a straightforward fashion. Instead, for the average case analysis we will use essentially the same techniques as in Section 6. In the noiseless case $Y = \mathbf{\Phi}_\Lambda X$, we want to show that

$$\|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} - \|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}} U\|_{p,\infty} > 0, \tag{7.1}$$

and concentration of measure tells us that this will be satisfied with very high probability as long as

$$\|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} - \|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} > 0. \tag{7.2}$$

Again to ensure the condition above we need to find a lower bound for the left hand side that does not depend on J itself but only on its size, $|J| \leq M-1$.
The first term on the left hand side can be estimated analogously to before as

$$\|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Phi}_\Lambda \Sigma\|_{2,\infty} \geq \sup_{i\in\Lambda\setminus J} \sigma_i \cdot \inf_{i\in\Lambda\setminus J} |\langle \tilde{\mathbf{Q}}_J \varphi_i, \psi_i\rangle|.$$

To get an upper bound on the second term we need to pay more attention because the contribution of the atoms in $\mathbf{\Phi}_J$ is no longer annihilated by the projection on the assumed atoms in $\mathbf{\Psi}_J$. To circumvent the difficulties arising from that we will use the following trick

$$\|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Phi}_\Lambda \Sigma\|_{2,\infty} \;\leq\; \|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Psi}_\Lambda \Sigma\|_{2,\infty} + \|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J (\mathbf{\Phi}_\Lambda - \mathbf{\Psi}_\Lambda)\Sigma\|_{2,\infty}.$$

The first norm in this expression can again be estimated as

$$\|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Psi}_\Lambda \Sigma\|_{2,\infty} \quad \leq \quad \sup_{i\in\Lambda\setminus J} \sigma_i \ \cdot \|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Psi}_{\Lambda\setminus J}\|_{2,\infty}.$$

To estimate the second norm we use $|\!|\!| Q_J |\!|\!| = \|\psi_k\|_2 = 1$ and $\|\phi_i - \psi_i\|_2^2 \leq 2(1-\beta)$,

$$\begin{aligned}
\|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J (\mathbf{\Phi}_\Lambda - \mathbf{\Psi}_\Lambda)\Sigma\|_{2,\infty}^2 &= \sup_{k\notin\Lambda} \sum_{i\in\Lambda} |\sigma_i|^2 |\langle \tilde{\mathbf{Q}}_J(\varphi_i - \psi_i), \psi_k\rangle|^2 \\
&\leq \sup_{k\notin\Lambda} \sum_{i\in\Lambda} |\sigma_i|^2 |\!|\!| \tilde{\mathbf{Q}}_J |\!|\!|^2 \|\varphi_i - \psi_i\|_2^2 \|\psi_k\|_2^2 \\
&\leq \sum_{i\in\Lambda} |\sigma_i|^2 2(1-\beta) \leq \|\sigma\|_2^2 \cdot 2(1-\beta) \leq S \cdot \sigma_{\max}^2 \cdot 2(1-\beta).
\end{aligned}$$

If we now define the coefficients $c_0'(\Lambda), d_0'(\Lambda)$

$$c_0'(\Lambda) = \inf_{J\subsetneq\Lambda} \inf_{i\in\Lambda\setminus J} |\langle \tilde{\mathbf{Q}}_J \varphi_i, \psi_i\rangle|, \quad d_0'(\Lambda) = \sup_{J\subsetneq\Lambda} \|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Psi}_{\Lambda\setminus J}\|_{2,\infty} + \frac{\sigma_{\max}}{\sigma_{\min}} \cdot \sqrt{2S(1-\beta)} \tag{7.3}$$

we can finally lower bound the left hand side in (7.2) as

$$\begin{aligned}
&\|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} - \|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Phi}_\Lambda \Sigma^{\frac{1}{2}}\|_{2,\infty} \\
&> \sup_{i\in\Lambda\setminus J} \sigma_i \cdot \Big(\inf_{i\in\Lambda\setminus J} |\langle \tilde{\mathbf{Q}}_J \varphi_i, \psi_i\rangle| - \|\mathbf{\Psi}_\Lambda^\star \tilde{\mathbf{Q}}_J \mathbf{\Psi}_{\Lambda\setminus J}\|_{2,\infty} - \frac{\|\sigma\|_2}{\sup_{i\in\Lambda\setminus J}\sigma_i} \sqrt{2(1-\beta)} \Big) \\
&> \sigma^{(M)} \cdot \big(c_0'(\Lambda) - d_0'(\Lambda) \big).
\end{aligned}$$

Using the notation above we could now formulate and proof the analogue of Theorem 6.1. However since that would basically mean rewriting Section 6.2 replacing $c_0(\Lambda), d_0(\Lambda)$ by $c_0'(\Lambda), d_0'(\Lambda)$ we will directly formulate the analogue of Theorem 3.4.

Theorem 7.1. *Let $p = 1$ and $Y = \mathbf{\Phi}_\Lambda \, \Sigma^{\frac{1}{2}} U$ with U a $S \times N$ matrix of standard Gaussian random variables, $\Sigma = \mathrm{diag}(\sigma_i^2)_{i\in\Lambda}$. Suppose in addition the following condition on the dynamic range is satisfied*

$$\frac{\sigma_{\max}}{\sigma_{\min}} < \frac{1}{\sqrt{2S(1-\beta)}} \cdot \big(\beta - \frac{\mu_2^{in}(\mathbf{\Psi}, \mathbf{\Psi}, \Lambda) + \mu_2(\mathbf{\Psi}, \mathbf{\Psi}, \Lambda)}{1 - \delta_\Lambda(\mathbf{\Psi})} \big). \tag{7.4}$$

Then the probability that S steps of 1-SOMP with $\mathbf{\Psi}$ instead of $\mathbf{\Phi}$ fail to exactly recover the support Λ does not exceed $K \cdot 2^S \cdot \exp(-N\gamma^2/\pi)$ with K the number of atoms in $\mathbf{\Phi}$ and

$$\gamma := \frac{\beta - \frac{\mu_2^{in}(\Lambda) + \mu_2(\Lambda)}{1-\delta_\Lambda} - \frac{\sigma_{\max}}{\sigma_{\min}} \cdot \sqrt{2S(1-\beta)}}{\beta - \frac{\mu_2^{in}(\Lambda) - \mu_2(\Lambda)}{1-\delta_\Lambda} + \frac{\sigma_{\max}}{\sigma_{\min}} \cdot \sqrt{2S(1-\beta)}}. \tag{7.5}$$

Proof 7. *Follow the same line of argument as in the proof of Theorem 3.4 and replace $c_0'(\Lambda), d_0'(\Lambda)$ by their bounds*

$$c'(\Lambda) = \beta - \frac{\mu_2^{in}(\mathbf{\Psi}, \mathbf{\Psi}, \Lambda)}{1 - \delta_\Lambda(\mathbf{\Psi})} \quad \textit{and} \quad d'(\Lambda) = \frac{\mu_2(\mathbf{\Psi}, \mathbf{\Psi}, \Lambda)}{1 - \delta_\Lambda(\mathbf{\Psi})} + \frac{\sigma_{\max}}{\sigma_{\min}} \cdot \sqrt{2S(1-\beta)}$$

Both estimates can be derived as in the proof of Lemma 6.3 simply by reversing the roles of $\mathbf{\Phi}$ and $\mathbf{\Psi}$. For the first one simply note that this works because $\tilde{\mathbf{Q}}_J$ is self-adjoint and so we have $\inf_{J\subsetneq\Lambda} \inf_{i\in\Lambda\setminus J} |\langle \tilde{\mathbf{Q}}_J \varphi_i, \psi_i\rangle| = \inf_{J\subsetneq\Lambda} \inf_{i\in\Lambda\setminus J} |\langle \tilde{\mathbf{Q}}_J \psi_i, \varphi_i\rangle|$.

Contrary to Theorem 3.4, the dynamic range $\sigma_{\max}/\sigma_{\min}$ is now constrained, which is more similar to the behaviour of p-thresholding. With p-thresholding the condition on the dynamic range to allow recovery (in the noiseless case) is

$$\frac{\sigma_{\max}}{\sigma_{\min}} < \frac{1}{\mu_2(S)}.$$

This is much less stringent than the condition (7.4) when the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are very different, that is to say when β is small. However if $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are very similar (β is close to one) then the condition on the dynamic range essentially vanishes and one can check that we recover the noiseless version of Theorem 3.4. This suggests that it might be preferable to choose the decomposition algorithm depending on the available precision of the estimate of $\mathbf{\Phi}$.

8 Conclusions and Outlook

Sparse approximations of signals over redundant dictionaries is an emerging methodology that has attracted researchers from a remarkably broad community, from signal processing practitioners to mathematicians. Despite remarkable practical success, there has always been quite a gap between the performances predicted by theory and those achieved in practice. Clearly, the weak element in theory was the prominent role of worst case analysis, casting overly pessimistic shadows on achievable results. In this paper we shed new light on the problem by turning to average case analysis, showing that greedy algorithms perform much better than the worst case prediction in most cases.

Nevertheless, our results are far from being the final answer. First, we had to restrict ourselves to the multichannel case where we could take advantage of the collective behaviour of atoms across channels. A similar average case analysis in the single channel case would be a major breakthrough. Advances have been reported for the simple thresholding algorithm [21], but success for iterative greedy algorithms remains elusive. Second, some of our theorems, most notably in the case of p-SOMP, use pachydermal union bounds that seem to require many channels in order to reach practical success probabilities. Solving this issue with finer arguments would also lead to further bridging the gap between theory and practice.

Acknowledgment

Some of the ideas that led to this paper first arose in a discussion of two of the authors with Morten Nielsen, whom we like to thank, at the Erwin Schrödinger International Institute for Mathematical Physics (ESI) in Vienna.

A Proof of Theorem 5.1

The proof of Theorem 5.1 relies heavily on the following standard result, see e.g. [14, eq. (2.35)] or [15, eq. (1.6)].

Theorem A.1. *Let f be a Lipschitz function on $\mathbb{R}^N$, i.e., $|f(x)-f(y)| \leq L\|x-y\|_2$ for all $x, y \in \mathbb{R}^N$. Further assume that $Z = (Z_1, \ldots, Z_N)$ is a vector of independent standard Gaussian random variables. Then*

$$\mathbb{P}(f(Z) \geq \mathbb{E}[f(Z)] + t) \leq \exp\left(-\frac{t^2}{2L^2}\right) \quad \text{and} \quad \mathbb{P}(f(Z) \leq \mathbb{E}[f(Z)] - t) \leq \exp\left(-\frac{t^2}{2L^2}\right).$$

Let us specialize this theorem to the p-norm (with the usual modification for $p = \infty$). To this end we let

$$C_p(N) := \mathbb{E}\left[\|Z\|_p\right] = \mathbb{E}\left(\sum_{n=1}^{N} |Z_n|^p\right)^{1/p}.$$

Further, we let $L_p(N)$ be the smallest constant such that

$$\|x\|_p \leq L_p(N)\|x\|_2 \quad \text{for all } x \in \mathbb{R}^N.$$

Further, we define

$$A_p(N) := \frac{C_p(N)^2}{2L_p(N)^2}.$$

We will later on give estimates of these constants for the most interesting cases, i.e $p = 1$, 2, ∞. Theorem A.1 thus leads to the following.

Corollary A.2. *Let $1 \leq p \leq \infty$. Suppose $Z = (Z_1, \ldots, Z_N)$ is a vector of independent standard Gaussians. Then*

$$\mathbb{P}(\|Z\|_p \geq (1+\epsilon)C_p(N)) \leq \exp\left(-\epsilon^2 A_p(N)\right) \tag{A.1}$$

and

$$\mathbb{P}(\|Z\|_p \leq (1-\epsilon)C_p(N)) \leq \exp\left(-\epsilon^2 A_p(N)\right). \tag{A.2}$$

Proof 8. *The Lipschitz constant of the function $f(x) = \|x\|_p$ can be estimated as*

$$|\|x\|_p - \|y\|_p| \leq \|x-y\|_p \leq L_p(N)\|x-y\|_2. \tag{A.3}$$

Taking $y = 0$ shows that this estimation is sharp. Applying Theorem A.1 with $t = \epsilon C_p(N)$ and using the definition of $A_p(N)$ yields the statement.

Remark A.1. We could even worked with $0 < p < 1$. Then one has to replace $L_p(N)$ by $2^{1/p-1}L_p(N)$, and hence $A_p(N)$ by $4^{1-1/p}A_p(N)$. Indeed, though $\|.\|_p$ is not a norm for $p < 1$, we have the quasi-triangle inequality $\|x+y\|_p \leq 2^{1/p-1}(\|x\|_p + \|y\|_p)$, see e.g. [8]. This would then be used in the first inequality in (A.3) instead of the usual triangle inequality.

Proof 9. *Consider the vector $v_k^\star U \in \mathbb{R}^N$. Its entries are given by $\langle v_k, U_n\rangle$, $\quad n = 1, \ldots, N$ where $U_n = (U_{n1}, \ldots, U_{nS})$ is a vector of independent standard Gaussians. Observe that the inner products $\langle v_k, U_n\rangle, n = 1, \ldots, N$ are stochastically independent with the same distribution as the (univariate) scaled Gaussian $\|v_k\|_2 U_{n1}$. Denoting $Z = (U_{11}, \ldots, U_{N1})$, Corollary A.2 yields*

$$\begin{aligned}\mathbb{P}\Big(\|v_k^\star U\|_p \geq (1+\epsilon_1)C_p(N)\|v_k\|_2\Big) &= \mathbb{P}\Big(\big\|(\|v_k\|_2 U_{n1})_{n=1}^N\big\|_p \geq (1+\epsilon_1)C_p(N)\|v_k\|_2\Big) \\ &= \mathbb{P}\left(\|Z\|_p \geq (1+\epsilon_1)C_p(N)\right) \leq \exp(-\epsilon_1^2 A_p(N)).\end{aligned}$$

In the same fashion we obtain the second inequality. Now by a union bound

$$\mathbb{P}\left(\max_{k\in\Omega}\|v_k^\star U\|_p \geq (1+\epsilon_1)C_p(N)\max_{k\in\Omega}\|v_k\|_2\right) \leq \sum_{k\in\Omega}\mathbb{P}\left(\|v_k^\star U\|_p \geq (1+\epsilon_1)C_p(N)\max_{k'\in\Omega}\|v_{k'}\|_2\right)$$
$$\leq \sum_{k\in\Omega}\mathbb{P}\Big(\|v_k^\star U\|_p \geq (1+\epsilon_1)C_p(N)\|v_k\|_2\Big) \leq |\Omega|\cdot\exp(-\epsilon_1^2 A_p(N)) \tag{A.4}$$

and, denoting $k_0\in\Omega$ *such that* $\|v_{k_0}\|_2 = \max_{k'\in\Omega}\|v_{k'}\|_2$

$$\mathbb{P}\left(\max_{k\in\Omega}\|v_k^\star U\|_p \leq (1-\epsilon_2)C_p(N)\max_{k\in\Omega}\|v_k\|_2\right) \leq \min_{k\in\Omega}\mathbb{P}\Big(\|v_k^\star U\|_p \leq (1-\epsilon_2)C_p(N)\|v_{k_0}\|_2\Big)$$
$$\leq \mathbb{P}\Big(\|v_{k_0}U\|_p \leq (1-\epsilon_2)C_p(N)\|v_{k_0}\|_2\Big) \leq \exp(-\epsilon_2^2 A_p(N)). \tag{A.5}$$

Similar techniques yield the last two estimates.

We could actually slightly improve the probability bound in the previous lemma. Indeed, in inequality (A.4) we were a bit crude when replacing $\max_{k'\in\Omega}\|v_{k'}\|_2$ with $\|v_k\|_2$ for each k. However, the resulting estimates improving (5.3) and (5.5) would be much more complicated, and in particular, if all the norms $\|v_k\|_2$ were roughly the same the gain would be marginal (which might be expected when $v_k = \Sigma\mathbf{\Phi}_\Lambda^\star\psi_k$ as used below). So we preferred to state the result in the current form. We thus sacrificed a little bit of precision to gain a much simpler looking result.

B Computation of $A_p(N)$ and $C_p(N)$

Let us now determine $A_p(N)$ and $C_p(N)$ for the important cases $p = 1, 2, \infty$.

Lemma B.1. *(a) For* $p=1$ *it holds* $C_1(N) = \sqrt{\frac{2}{\pi}}N$, $L_1(N) = \sqrt{N}$ *and*

$$A_1(N) \;=\; \frac{N}{\pi}.$$

(b) For $p=2$ *we have* $C_2(N) = \sqrt{2}\frac{\Gamma(N/2)}{\Gamma((N-1)/2)} \sim \sqrt{N}$, *where* Γ *denotes the* Γ *function. Hence,*

$$A_2(N) \;=\; \frac{\Gamma^2(N/2)}{\Gamma^2((N-1)/2)} \;\sim\; N/2.$$

(c) For $p=\infty$ *there is a constant* D *such that* $D^{-1}\sqrt{\log(N)} \leq C_\infty(N) \leq D\sqrt{\log(N)}$ *and hence*

$$A_\infty(N) \asymp \log(N).$$

Proof. Case (a) ($p=1$) is obvious. For $p=\infty$, case (c), we have $L_\infty(N)=1$ and [15, eq. (3.14)] tells us that there exists a constant K such that

$$K^{-1}\sqrt{\log(N)} \leq \mathbb{E}\|Z\|_\infty \leq K\sqrt{\log(N)}.$$

Hence,

$$A_\infty(N) \asymp \log(N).$$

Concerning $p = 2$, case (b), we clearly have $L_2(N) = 1$. The claim on $C_2(N) = \mathbb{E}\|Z\|_2$ is proved as follows.

The random variable

$$Y := \sum_{n=1}^{N} Z_n^2$$

has the $\chi^2(N-1)$ distribution (see e.g. [13]), that is, its probability density is given by

$$f(x) \;=\; \frac{1}{\Gamma((N-1)/2)}(1/2)^{(N-1)/2}x^{(N-1)/2-1}e^{-x/2}, \quad x \geq 0$$

Hence,

$$\begin{aligned}\mathbb{E}\|Z\|_2 \;&=\; \mathbb{E}\sqrt{Y} \;=\; \int_0^\infty x^{1/2}f(x)dx \;=\; \frac{1}{\Gamma((N-1)/2)}(1/2)^{(N-1)/2}\int_0^\infty x^{N/2-1}e^{-x/2}dx \\ &=\; \frac{1}{\Gamma((N-1)/2)}(1/2)^{(N-1)/2}2^{N/2}\int_0^\infty x^{N/2-1}e^{-x}dx \;=\; \sqrt{2}\frac{\Gamma(N/2)}{\Gamma((N-1)/2)}.\end{aligned}$$

Here, we used a substitution in the integral and the definition of the Γ-function, $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$. Further, using Stirling's formula, $\Gamma(z) \sim \sqrt{2\pi}z^{z-1/2}e^{-z}$, we obtain

$$\begin{aligned}\frac{\Gamma(N/2)}{\Gamma((N-1)/2)} \;&\sim\; \frac{(N/2)^{(N-1)/2}e^{-N/2}}{((N-1)/2)^{(N-2)/2}e^{-(N-1)2}} \;=\; \sqrt{\frac{1}{2e}}\frac{N^{(N-1)/2}}{(N-1)^{(N-2)/2}} \\ &=\; \sqrt{\frac{1}{2e}}\sqrt{N}\left[(1-1/N)^{N+2}\right]^{-1/2} \;\sim\; \sqrt{\frac{N}{2}}.\end{aligned}$$

The claim for $A_2(N)$ follows immediately. □

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *preprint*, 2007.

[2] D. Baron, M. Duarte, S. Sarvotham, M. Wakin, and R. Baraniuk. An information-theoretic approach to distributed compressed sensing. In *Proc. 45rd Conference on Communication, Control, and Computing*, 2005.

[3] D. Baron, M. Wakin, M. Duarte, S. Sarvotham, and R. Baraniuk. Distributed Compressed Sensing. *preprint*, 2005.

[4] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[5] E. Candes and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies ? *IEEE Transactions on Information Theory*, 52(12):5406–5425, December 2006.

[6] J. Chen and X. Huo. Sparse representations for multiple measurement vectors (MMV) in an over-complete dictionary. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP-2005)*, 2005.

[7] J. Chen and X. Huo. Theoretical results on sparse representations of multiple measurement vectors. *IEEE Transactions on Signal Processing*, 54(12):4634–4643, December 2006.

[8] R. DeVore and G. Lorentz. *Constructive approximation.* Springer-Verlag, 1993.

[9] D. Donoho and M. Elad. Maximal sparsity representation via l^1 minimization. *Proc. Nat. Acad. Sci.*, 100(4):369–388, 2003.

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

[11] D. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies. Data compression and harmonic analysis. *IEEE Transactions on Information Theory*, 44:391–432, August 1998.

[12] R. Gribonval and M. Nielsen. Beyond sparsity : recovering structured representations by ℓ^1-minimization and greedy algorithms. *Advances in Computational Mathematics*, 2006. accepted.

[13] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes.* Oxford University Press, 2001.

[14] M. Ledoux. *The Concentration of Measure Phenomenon.* AMS, 2001.

[15] M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes.* Springer-Verlag, Berlin, Heidelberg, NewYork, 1991.

[16] S. Lesage, S. Krstulovic, and R. Gribonval. Under-determined source separation: comparison of two approaches based on sparse decompositions. In J. P. J. Rosca, D. Erdogmus and S. Haykin, editors, *Proc. of the Int'l. Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2006)*, volume 3889 of *LNCS Series*, pages 633–640, Charleston, South Carolina, USA, Mar. 2006. Springer.

[17] Z. Luo, M. Gaspar, J. Liu, and A. Swami. Distributed signal processing in sensor networks. *IEEE Signal processing magazine*, 23(4):14–15, July 2006.

[18] H. Rauhut. Stability results for random sampling of sparse trigonometric polynomials. *preprint*, 2006.

[19] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *preprint*, 2006.

[20] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In *Proc. CISS 2006 (40th Annual Conference on Information Sciences and Systems)*. others, 2006.

[21] K. Schnass and P. Vandergheynst. Average Performance Analysis for Thresholding. *accepted to IEEE Signal Processing Letters*, 2007.

[22] K. Schnass and P. Vandergheynst. Dictionary preconditiong for greedy algorithms. *submitted to IEEE Transactions on Signal Processing*, 2007.

[23] D. Taubman and W. Marcellin. *JPEG2000: Image compression fundamentals, standards, and practice*. Springer, 2002.

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

[25] J. Tropp. *Topics in sparse approximation*. PhD thesis, University of Texas at Austin, 2004.

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

[27] J. Tropp. Random subdictionaries of general dictionaries. *preprint*, 2006.

[28] J. Tropp, A. Gilbert, and M. Strauss. Algorithms for simultaneous sparse approximations. Part I: Greedy pursuit. *Signal Processing, special issue "Sparse approximations in signal and image processing"*, 86:572–588, 2006.