Sparsity & Dictionaries - Algorithms & Design

THÈSE N^{O} 4349 (2009)

PRÉSENTÉE LE 13 MARS 2009

À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE DE TRAITEMENT DES SIGNAUX 2

PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Karin SCHNASS

acceptée sur proposition du jury:

Prof. M. Hasler, président du jury
Prof. P. Vandergheynst, directeur de thèse
Prof. H. G. Feichtinger, rapporteur
Dr R. Gribonval, rapporteur
Prof. E. Telatar, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2009

If you don't try, you never know.

Merlin to Wart (Arthur)
'The Sword in the Stone'

Acknowledgements

The first person I want to thank is Pierre (Prof. Vandergheynst), my advisor, psychiatrist and IT-support all rolled into one, for being there whenever I had a problem and leaving me in peace the rest of the time and for giving me the chance to go wherever I wanted, both scientifically and spatially.
I want to thank the members of the jury for carefully reading the thesis and for their helpful comments and suggestions.
A big scientific thank you goes out to Rémi Gribonval for getting me into dictionary learning and to the members of LTS2 for fruitful and stupid discussions during all the coffee/juice/cake/fruit breaks, especially Laurent (fruitful) and Boris & Antonin (stupid). Thanks also to what used to be ITS for being 'the happy lab', with an extra big 'Merci!' to Gilles for a lot of old computer parts.

All science would not have been possible without the moral support of people in and outside of Lausanne. So thanks to Nino for being my initial social life and the perfect husband (flatmate), to Klaske for many romantic EPFL lunches, to Maria for hugs and coffees in the MX cafeteria, to Luigi for great dinners and supportive candy, to Jasper and Gaetano for more supportive cooking, to Lorenzo for fixing my blue Ferrari and giving me the bike virus, to all the friends in Lausanne, Serkan, Fred, Oli, Zafer, Giulia, Gabriel, Davor, Harm, Esmee, for making the Swiss experience enjoyable, to the Austrian girls for rapidly typed advice, and the boys for continuing where we stopped.
The last very special thanks goes to my mom for the 24h crisis centre (hdl), my father for retrieving recipes and my brother for being my first visitor.

Abstract

With the flood of information available today the question how to deal with high dimensional data/signals, which are cumbersome to handle, to calculate with and to store, is highly important. One approach to reducing this flood is to find sparse signal representations, as a signal that is the linear combination of a few elements from a pool of building blocks, can be reduced to the few coefficients of this representation. If these building blocks form a basis, finding the sparse representation poses no problem but unfortunately not many signal classes are sparse in a basis. Taking more building blocks, i.e. a redundant dictionary, increases the chances of having sparse representations, but actually finding them becomes very hard. This led to the development of numerous strategies and algorithms for finding sparse representations, with varying complexity and success rate.

The first part of the thesis deals with two of those algorithms, Thresholding and Matching Pursuit, from a more theoretical point of view. It is shown that both those greedy algorithms can be improved with a little trick, that does not increase their complexity, and that when considering their average instead of their worst case performance they perform quite well in comparison to more complex methods.

The second part of thesis treats questions concerning the whole dictionary and its properties. First it gives more evidence that sparsity is useful by extending the concept of compressed sensing to signals that are sparse not in a basis but in a redundant dictionary. Thus to record a sparse signal it is not necessary to make as many measurements as the dimension of the signal but only a multiple of the number of dictionary elements used to represent it.

Next we show that dictionaries cannot only provide sparse representations but that their geometric properties can also be exploited to model data structures. Here we explain how to model different subclasses of a class of signals by incoherent subspaces, present an algorithm to learn a dictionary made out of these subspaces and then use it for classification of faces.

Finally we turn back to the sparse representation problem and study the fundamental question how to find a dictionary providing sparse representations. We pick up the idea to learn a dictionary via minimisation of a continuous cost function and provide conditions, guaranteeing that the decomposition of a collection of training signals into a dictionary and a coefficient matrix constitutes a local minimum. We also analyse statistically when these conditions are fulfilled with high probability.

Keywords: sparse representation, redundant dictionary, greedy algorithms, preconditioning, average case analysis, multichannel, compressed sensing, classification, dictionary learning

Zusammenfassung

Angesichts der Informationsflut heutzutage wird die Frage, wie man mit hoch-dimensionalen Daten/Signalen, die umständlich zu handhaben, zu manipulieren und zu speichern sind, umgehen soll, immer wichtiger. Ein Ansatz zur Eindämmung dieser Flut ist es spärliche Signaldarstellungen zu finden, da ein Signal, das Linearkombination weniger Elemente eines Satzes von Bausteinen ist, auf die wenigen Koeffizienten dieser Darstellung reduziert werden kann. Bilden die Bausteine eine Basis, kann die spärliche Darstellung problemlos gefunden werden, doch leider sind nicht viele Signalklassen spärlich im Bezug auf eine Basis. Nimmt man mehr Bausteine, also ein redundantes Wörterbuch, vergrößern sich die Existenzchancen einer spärliche Darstellung, aber diese auch tatsächlich zu finden wird zu einer Herausforderung, was zur Entwicklung zahlreicher Strategien und Algorithmen zur Auffindung spärlicher Darstellungen, mit verschiedenem Aufwand und Erfolg, führte.
Der erste Teil dieser Dissertation beschäftigt sich mit zwei solchen Algorithmen, "Thresholding" und "Matching Pursuit", von einem theoretischen Gesichtspunkt aus. Es wird gezeigt, dass diese beiden gierigen Algorithmen durch einen kleinen Trick, der den Aufwand nicht erhöht, verbessert werden können, und dass, wenn das durchschnittliche statt des Verhaltens im ungünstigsten Fall herangezogen wird, sie im Vergleich zu komplizierteren Verfahren recht gut abschneiden.
Der zweite Teil der Dissertation behandelt Fragen, die das gesamte Wörterbuch und seine Eigenschaften betreffen. Zuerst wird ein weiterer Beleg gegeben, wie nützlich Spärlichkeit ist, indem das Konzept der komprimierten Abtastung auf Signale ausgeweitet wird, die spärlich in einem redundanten Wörterbuch statt einer Basis sind. So ist es zur Aufnahme eines spärlichen Signals nicht nötig, soviele Messungen wie das Signal Dimensionen hat vorzunehmen, sondern nur ein Vielfaches der Anzahl von Wörterbuchelementen, die zur Darstellung verwendet wurden. Als Nächstes zeigen wir, dass Wörterbücher nicht nur spärliche Darstellungen liefern, sondern dass ihre geometrischen Eigenschaften auch zur Modellierung von Datenstrukturen ausgenutzt werden können. Hier erklären wir die Modellierung verschiedener Unterklassen einer Signalklasse durch inkoherente Teilräume, präsentieren einen Algorithmus, um ein Wörterbuch, das aus solchen Teilräumen besteht, zu lernen und verwenden ihn zur Klassifizierung von Gesichtern.
Schließlich kehren wir zurück zu dem Problem der spärliche Darstellung und beschäftigen uns mit der grundlegenden Frage, wie man ein Wörterbuch, das spärliche Darstellungen liefert finden kann. Wir greifen die Idee auf, ein Wörterbuch durch Minimierung einer kontinuierlichen Kostenfunktion zu lernen, und erarbeiten Bedingungen, die gewährleisten, dass die Zerlegung von Trainingssignalen in ein Wörterbuch und eine Koeffizientenmatrix ein lokales Minimum darstellt. Ebenfalls untersuchen wir statistisch, wann diese Bedingungen mit hoher Wahrscheinlichkeit erfüllt sind.

Stichworte: spärliche Darstellung, redundantes Wörterbuch, gierige Algorithmen, Vorkonditionierung, Durchschnittsanalyse, mehrere Kanäle, komprimierte Abtastung, Klassifizierung, Lernen von Wörterbüchern

Contents

List of Figures

List of Tables

1 Introduction

The title of this thesis is Sparsity & Dictionaries - Algorithms & Design. So let's start with a short explanation of what sparsity and dictionaries are, why we are interested in them and what we need the algorithms and the design for. Sparsity means that something is rare and in the region of our interest the world of signals, vectors and matrices what is rare or sparse are non-zero entries. Thus a sparse vector x has only a few components $x_i \neq 0$ and likewise a sparse matrix A. The two main advantages of these sparse objects are that they are easy to store and easy to compute with. To store a vector $x \in \mathbb{R}^d$ resp. a matrix $A \in \mathbb{R}^{n \times m}$ we normally need to remember d resp. nm numbers but if it is sparse with $S \ll d, nm$ non-zeros, it is enough to remember the addresses of the non-zero components and their values, i.e. $2S$ numbers. Similarly if we want to calculate with a sparse vector. Assume we want to calculate the inner product of two vectors. Normally we would have to compute the product between all the corresponding entries and then sum these up leading to $2d-1$ operations, but if one of the vectors is sparse we just need to compute the product between the non-zeros components in the sparse vector with the corresponding ones in the other vector and sum them up, leading to $2S - 1$ operations.

The concept of sparsity we have talked about so far is, however, too restrictive to be useful. For instance take a sparse vector and multiply it with an orthonormal matrix $\mathbf{\Phi}$. The resulting vector or signal $y = \mathbf{\Phi}x$ will in the generic case not be sparse anymore, meaning most of its entries will be non-zero. Still if someone gives you many signals y_i of this type and tells you to store them, you can use the knowledge that all y_i have a sparse representation in the orthonormal basis $\mathbf{\Phi}$, i.e. $y_i = \mathbf{\Phi}x_i$, calculate $x_i = \mathbf{\Phi}^\star y_i$ and store x_i and $\mathbf{\Phi}$ instead. This technique is used in every day life when looking at a jpeg image. On the hard-drive not the image y itself is stored, but the coefficients x of the image in a wavelet basis $\mathbf{\Phi}$, which are sparse and therefore take less space, and if the picture is needed y is quickly reconstructed as $\mathbf{\Phi}x$.

The problem is that for many signal classes there is no orthonormal basis that provides sparse representations or approximations for all the signals. Thus in the next step one can consider any kind of basis $\mathbf{\Phi}$ and using the biorthogonal basis $\mathbf{\Phi}^{-1}$ can again switch easily between signal and sparse representation. Unfortunately the signal classes that have good sparse representations or approximations in a basis is still not enough and one has to turn to overcomplete representation systems or dictionaries. An overcomplete dictionary corresponds to a non square $d \times K$ matrix $\mathbf{\Phi}$ with more columns than rows, $d < K$. This means that there are more, K instead of d, vectors, i.e. columns of $\mathbf{\Phi}$, that we can sparsely superpose to build a signal. The drawback is that for every signal there is now more than one way to represent it in the dictionary, just as any underdetermined system

of d linear equations in K variables has more than one solution. Out of all these representations we are of course interested in the sparsest one. The problem is that it is not easy to find. Thus the first part of this thesis is dedicated to the study of algorithms for finding sparse representations in overcomplete dictionaries. The second part is less homogenous, featuring another reason why sparse signals are useful which is known as *compressed sensing*, an application how sparsity can be used to model subclasses of a signal class and use that for classification and finally addressing the question, how to find a dictionary that is suitable to represent a signal class. The common element of manipulating or creating a whole dictionary however justifies the title 'Design'.

1.1 Outline

In the first part we study algorithms to find sparse representations. In Chapter 2 we introduce two Greedy Algorithms, Thresholding and (Orthogonal) Matching Pursuit and their shortcomings. We derive that they can be split into two steps a sensing and a reconstruction step, and that the former will fail to identify correct building blocks if the blocks in the dictionary are too similar, i.e. the coherence of the dictionary is too high. We modify the sensing step by introducing a special sensing dictionary. The correct selection of components is then determined by the *cross coherence* which can be considerably lower than the coherence. We characterise the optimal sensing dictionary and develop a constructive method to approximate it. Finally we compare the performance of Thresholding and OMP using the original and modified algorithms.

In Chapter 3 we show that the Thresholding algorithm is more powerful than previously assumed. The worst case analysis, as in Chapter 2, suggests that it can only succeed if the signals are very sparse, meaning the number of building blocks is of the order of the square root of the ambient dimension. We perform an average analysis considering a random distribution of the signs of the building blocks and find out that with high probability Thresholding can succeed for sparsity levels up to the order of the ambient dimension. As an application of the theory we take the sensing dictionaries introduced in Chapter 2, characterise when they give optimal average performance and test them numerically.

Chapter 4 is dedicated to building not single houses out of a few building blocks but whole neighbourhoods. We generalise Thresholding and (O)MP to compute simultaneous sparse approximations of multichannel signals and analyse their behaviour assuming a random model on the coefficients of the building blocks. Again we see that with high probability we can recover sparsity levels up to the order of the ambient dimension.

In the second part of the thesis we show how to exploit the fact that signals have a sparse representation in a dictionary and finally study how to learn a dictionary.
In Chapter 5 we extend the concept of *compressed sensing*, acquiring a signal from only a small number of measurements, to signals that are not sparse in an orthonormal basis but rather in a redundant dictionary. To do this we show that a matrix, which is a composition of a random matrix of certain type and a deterministic dictionary, has small restricted isometry constants, which is a sufficient condition to recover signals sparse with respect to the dictionary from a few measurements using the Basis Pursuit Principle. We also show that Thresholding can be used as recovery algorithm for compressed sensing and provide conditions that guarantee reconstruction with high probability. Finally we compare the performance of Thresholding, (O)MP and Basis Pursuit with numerical experiments.

Chapter 6 demonstrates how to use the fact that different signals can be represented more or less well by certain atoms in a dictionary for classification. We present a signal model for classification based on a collection of low dimensional subspaces embedded into the high dimensional signal space. Each subspace is spanned by a certain number of dictionary elements which represent the signals in one class well but not the other classes. We develop an alternate projection algorithm to find such a collection and test the classification performance of our scheme in comparison to Fisher's LDA and a recent approach based on sparse approximation.

Finally one of the most import problems around dictionaries and sparse representations, namely how to actually find a dictionary that will give you sparse representations for a class of signals, is addressed in Chapter 7. Given the decomposition of a signal class into a dictionary and sparse coefficients we derive conditions on the coefficients that guarantee that locally there is no dictionary leading to sparser coefficients, when sparsity is measured by the sum of the absolute values of all coefficients. We then show that assuming a random sparse model on the coefficients these conditions will be satisfied with high probability as long as the dictionary is not too coherent and the number of training signals is large enough.

Chapter 8 concludes this thesis. We briefly discuss the main contributions and point out directions for further research.

Part I

Algorithms

Dictionary Preconditioning for Greedy Algorithms

2

In this chapter we give a short introduction to dictionaries and sparse signal representations and approximations. We present two greedy algorithms for finding sparse approximations Thresholding and (Orthogonal) Matching Pursuit. We analyse their shortcomings by splitting them into a sensing and a reconstruction step, and showing that the sensing step will fail if the 1-Babel function of the dictionary, that measure the similarity of the elements, is growing too fast. We then modify the sensing step by introducing a special *sensing dictionary*. The correct selection of components is then determined by the *1-cross-Babel function* which can be considerably lower than the 1-Babel function. We characterise the optimal sensing matrix and develop a constructive method to approximate it. Finally we compare the performance of Thresholding and OMP using the original and modified algorithms. Most of the material presented in this chapter has been published in [50].

2.1 Dictionaries & Sparse Representations

In the last years, constructing sparse signal approximations by means of redundant dictionaries has received a lot of attention, see [13, 16, 21, 55] and the references therein for a thorough introduction. In short the reason for this interest is that a sparse signal representation effectively reduces the dimensionality of the signal and thus makes it easier to store or manipulate. The use of redundant dictionaries is then simply a consequence of the fact that the existence of a sparse signal representation becomes more likely as the number of building blocks or atoms in the dictionary increases. Before we can illustrate the topic further by stating two of the typically investigated problems, we will need to introduce some vocabulary. We will be working with signals $y \in \mathbb{R}^d$. A dictionary $\mathbf{\Phi}$ is assumed to be represented by a $d \times K$ matrix, with $d \ll K$, whose columns are the atoms φ_i, $\|\varphi_i\|_2 = 1$:

$$\mathbf{\Phi} = [\varphi_1 \dots \varphi_K].$$

The ratio $R = K/d$ is called redundancy. A signal is said to have a S-sparse representation in the dictionary $\mathbf{\Phi}$ if there exists a set Λ with $|\Lambda| = S$ such that we can write

$$y = \sum_{i \in \Lambda} x_i \varphi_i = \mathbf{\Phi}_\Lambda x.$$

With a slight abuse of language we will call both the set Λ and the atoms with indices in Λ the support of y and write $\mathbf{\Phi}_\Lambda$ for the $d \times S$ matrix of all the atoms in the support. The complement of the support will be denoted by $\overline{\Lambda} = \{1 \dots K\}/\Lambda$.

Now, having all definitions in place, the first problem, concerned with finding sparse signal approximations, can be more accurately stated as:

Problem 2.1.1. Given a signal y, find its best S-sparse approximation in the dictionary $\mathbf{\Phi}$, i.e.

$$\min_{\Lambda,x} \|y - \mathbf{\Phi}_\Lambda x\|_2 \text{ s.t. } |\Lambda| = S,$$

or the converse problem given y find the sparsest ε-approximation, i.e.

$$\min_{\Lambda} |\Lambda| \text{ s.t. } \min_{x} \|y - \mathbf{\Phi}_\Lambda x\|_2 \leq \varepsilon.$$

Of course for any signal and dictionary there always exist solutions to the above problems. However, in order to justify the use of the term sparse, we obviously need to have a dictionary in which the signal has a representation where both ε and S are small, i.e. $S \ll d$. This leads to the next question:

Problem 2.1.2. Given a class of signals Y, find a dictionary $\mathbf{\Phi}$ such that all signals $y \in Y$ will have a good sparse approximation in $\mathbf{\Phi}$.

Without any further assumption on the signal or the dictionary, finding the solution to the first problem is combinatorial. Thus one would have to try the orthogonal projection of the signal on all possible S-sparse supports. To circumvent this problem people started imposing restrictions on the dictionary and/or the coefficients x. By now there exists detailed theory describing under which assumptions suboptimal algorithms like Thresholding, (Orthogonal) Matching Pursuit (OMP), or the Basis Pursuit (BP) Principle, can be proven to recover the true support, see for instance [8, 22, 55]. The property at the base of most theorems for greedy algorithms is slow growth of the 1-Babel function or cumulative coherence $\mu_1(S, \mathbf{\Phi})$ of the dictionary, which is defined as:

$$\mu_1(S, \mathbf{\Phi}) = \max_i \max_{|J|=S, i \notin J} \sum_{j \in J} |\langle \varphi_j, \varphi_i \rangle|. \tag{2.1}$$

It gives an indication of how close/far the dictionary is to/from an orthonormal basis. For compactness reasons we will omit the reference to dictionary, i.e. write $\mu_1(S)$, whenever it is clear which dictionary is meant and write μ for the coherence, i.e. $\mu := \mu_1(1)$. Using this definition a typical result for Thresholding, cp. [25], and OMP, cp. [55], reads as:

Theorem 2.1.1. *If we have a signal exactly S-sparse in $\mathbf{\Phi}$, i.e. $y = \sum_{i \in \Lambda} x_i \varphi_i$ and $|\Lambda| = K$, then Thresholding is able to recover a component φ_i of the true support if*

$$\frac{|x_i|}{\|x\|_\infty} > \mu_1(S) + \mu_1(S-1). \tag{2.2}$$

OMP is able to recover all components of the true support Λ if the exact recovery coefficient is smaller than 1, i.e.

$$\|\mathbf{\Phi}_\Lambda^\dagger \mathbf{\Phi}_{\overline{\Lambda}}\|_{1,1} < 1,$$

where $\mathbf{\Phi}_\Lambda^\dagger$ *denotes the Moore-Penrose pseudo-inverse. The above condition is always satisfied if*

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

One deduction from the theorem is that it is desirable to have a dictionary where the cumulative coherence is growing slowly. Dictionaries having minimal coherence μ are called *Grassmannian frames* and are quite well studied, see [53] and references therein, but the next step of trying to minimise the cumulative coherence seems novel. However we can give a lower bound on the cumulative coherence based on results about Grassmannian frames. The following theorem is an extension of Theorem 2.3 in [53].

Theorem 2.1.2. *Let* $\mathbf{\Phi}$ *be a dictionary of* K *atoms in dimension* d*. If* $S^2 < K - 1$ *then*

$$\mu_1(S) \geq S \cdot \sqrt{\frac{K-d}{d(K-1)}}. \tag{2.3}$$

Equality holds if and only if the dictionary is an equiangular unit norm tight frame.

The proof of the theorem is quite technical and not necessary for further developments. It can be found in the appendix of [50]. What should be noted though is that *optimal Grassmannian frames* that meet the lower bound for the coherence, i.e.

$$\mu(\mathbf{\Phi}) = \sqrt{\frac{K-d}{d(K-1)}}$$

simultaneously meet the lower bound for the cumulative coherence $\mu_1(K)$ for all S with $S^2 < K-1$.

On the other hand while a dictionary minimising the cumulative coherence might be interesting for communication applications, it will not be ideal for approximation of a specific class of signals, like for instance EEGs or music. For these purposes learned dictionaries are by definition more suited to the task, see [3, 19, 29, 30]. However these learned dictionaries will not show the desired incoherence properties, that enable us to find the approximation with suboptimal algorithms in the same degree as optimal Grassmannian frames. Assume that we have a dictionary that represents a signal class well but is unfortunately so coherent that already $\mu_1(2) + \mu_1(1) > 1$, meaning that we cannot guarantee for OMP to find even a superposition of only two atoms. Thus in order to find good approximations we would have to use a more complex algorithm. Alternatively we could circumvent the problem by trying to find a new dictionary that still represents the class well but retains small minimal cumulative coherence. For more ideas in this direction, see Chapter 7.
In this chapter we introduce the concept of sensing dictionaries and present a small alteration of the suboptimal algorithms such that they can perform well for dictionaries with high cumulative coherence. In Section 2.2, we first explain how to separate the Thresholding algorithm into a sensing and a reconstruction part. We then show that sensing with a different dictionary can lower the cumulative cross-coherence and yield better recovery results. Motivated by structural properties of optimal Grassmannian frames we propose an iterative algorithm to construct a sensing dictionary/matrix giving lower cross-coherence. After analysing its convergence properties theoretically we use it to calculate sensing matrices for various dictionaries and compare the performance of Thresholding with and without sensing dictionaries in practice. In Section 2.3 we introduce sensing dictionaries as well for (O)MP and from a worst case performance analysis derive a characterisation of the ideal sensing dictionary. Again we do some numerical simulations of how OMP performs with or without sensing matrices using the sensing dictionaries obtained with the algorithm developed in

Section 2.2. Section 2.4 discusses the theoretical and numerical limitations of the schemes so far, as well as possible extensions.

2.2 Sensing Dictionaries for Thresholding

As mentioned above Thresholding can be formally decomposed into sensing steps, where we try to identify correct atoms of the support, and reconstruction steps, see the table below.

Sensing:	find Λ that contains the indices corresponding to the S largest values of $\lvert\langle y, \varphi_k\rangle\rvert$
Reconstruction:	$a = \mathbf{\Phi}_\Lambda \mathbf{\Phi}_\Lambda^\dagger y$

Table 2.1: Thresholding

$\mathbf{\Phi}_\Lambda^\dagger$ again denotes the Moore-Penrose pseudo inverse. If the dictionary is too coherent the sensing part will fail to identify correct atoms. Our idea is to change the sensing part and instead of sensing with the dictionary, use a different sensing matrix $\mathbf{\Psi}$ that allows to identify more correct components. This sensing matrix will have as columns the same number of sensing atoms as the original dictionary had atoms, so that we have a one to one correspondence between the sensing and the original atoms. If we denote the sensing atom in $\mathbf{\Psi}$ that corresponds to the atom φ_i in the original dictionary with ψ_i schematically the new algorithm looks like:

Sensing (new):	find Λ that contains the indices corresponding to the S largest values of $\lvert\langle y, \psi_k\rangle\rvert$
Reconstruction:	$a = \mathbf{\Phi}_\Lambda \mathbf{\Phi}_\Lambda^\dagger y$

Table 2.2: Thresholding with a Sensing Matrix

This approach can be easily motivated on the following example. Assume for instance that the dictionary $\mathbf{\Phi}$ is a deformed version of a dictionary $\mathbf{\Gamma}$ with low coherence, like an optimal Grassmannian frame or even more simple an orthogonal basis, meaning $\mathbf{\Phi} = A\mathbf{\Gamma}$ where A is an invertible matrix with inverse $A^{-1} = B$. For any S-sparse signal $y = \mathbf{\Phi}x$ by applying the matrix B we can construct a new signal $z = By = B\mathbf{\Phi}x = \mathbf{\Gamma}x$. To find the sparse support Λ we could equivalently use the original signal and dictionary or solve this new problem. But since for a Grassmannian frame $\mathbf{\Gamma}$ the cumulative coherence grows more slowly - in the case of $\mathbf{\Gamma}$ being an orthogonal basis it is even zero - the second problem is obviously better conditioned:

$$\begin{aligned} y = \mathbf{\Phi}x \quad &\Leftrightarrow \quad z = \mathbf{\Gamma}x \\ \mu_K(\mathbf{\Phi}) \quad &\geq \quad \mu_K(\mathbf{\Gamma}) \end{aligned}$$

However, if we write down explicitly the sensing of z with $\mathbf{\Gamma}$ ($\mathbf{\Gamma}^\star$ denotes the transpose of $\mathbf{\Gamma}$),

$$\mathbf{\Gamma}^\star z = (B\mathbf{\Phi})^\star By = (\mathbf{\Phi}^\star B^\star B)y,$$

we see that we can actually interpret it as sensing the original signal with a sensing matrix of the form $\mathbf{\Psi} = B^\star B\mathbf{\Phi}$. In the special case where we choose B such that $B^\star B = (\mathbf{\Phi}\mathbf{\Phi}^\star)^{-1}$ we get as sensing matrix the canonical dual frame (pseudo-inverse): $\mathbf{\Psi} = (\mathbf{\Phi}\mathbf{\Phi}^\star)^{-1}\mathbf{\Phi}$, which in the even more

special case where the dictionary is a basis is just the biorthogonal basis $(\mathbf{\Phi}^{-1})^\star$.
Now in order to generalise the above idea we can investigate what happens if we do not insist on deriving the sensing matrix from a linear transformation of the problem. Instead of restricting ourselves to using sensing matrices of the form $\mathbf{\Psi} = B^\star B\mathbf{\Phi}$, we will allow any matrix of the same size as the original dictionary. To see explicitly what properties we want to infer for the sensing/measuring matrix $\mathbf{\Psi}$ we do the analogue of the analysis leading to (2.2).

2.2.1 Worst Case Analysis of Thresholding with a Sensing Dictionary

Let y be a d-dimensional signal that has a S-sparse representation in the overcomplete dictionary $\mathbf{\Phi}$, $|\mathbf{\Phi}| = N$, i.e. $y = \sum_{i\in\Lambda} x_i\varphi_i$. For Thresholding to recover a component φ_i in the support, we need the inner product of signal with the corresponding sensing atom ψ_i to be larger than the inner product with any atom in the sensing matrix whose corresponding partner is not part of the support:

$$i \in \Lambda: \quad |\langle y, \psi_i\rangle| \quad \geq \quad |\langle y, \psi_j\rangle|, \quad \forall j \notin \Lambda.$$

Writing out the inner product we can estimate:

$$\begin{aligned} i \in \Lambda: \ |\langle y, \psi_i\rangle| &\geq |x_i||\langle \varphi_i, \psi_i\rangle| - \sum_{j\in\Lambda, j\neq i} |x_j||\langle \varphi_j, \psi_i\rangle| \\ &\geq |x_i||\langle \varphi_i, \psi_i\rangle| - \|x\|_\infty \sum_{j\in\Lambda, j\neq i} |\langle \varphi_j, \psi_i\rangle|, \\ k \notin \Lambda: \ |\langle y, \psi_k\rangle| &\leq \sum_{j\in\Lambda} |x_j||\langle \varphi_j, \psi_k\rangle| \leq \|x\|_\infty \sum_{j\in\Lambda} |\langle \varphi_j, \psi_k\rangle|. \end{aligned}$$

The right most terms in the above equations show a strong similarity to the cumulative coherence. In analogy we define the 1-cross-Babel function or cumulative cross-coherence of two dictionaries $\tilde{\mu}_1(K, \mathbf{\Phi}, \mathbf{\Psi})$ as well as their minimal similarity $\beta(\mathbf{\Phi}, \mathbf{\Psi})$ as:

$$\tilde{\mu}_1(K, \mathbf{\Phi}, \mathbf{\Psi}) \quad := \quad \max_i \max_{|J|=S, i\notin J} \sum_{j\in J} |\langle \varphi_j, \psi_i\rangle|, \tag{2.4}$$

$$\beta(\mathbf{\Phi}, \mathbf{\Psi}) \quad := \quad \min_i |\langle \varphi_i, \psi_i\rangle|. \tag{2.5}$$

As before we leave out the reference to the dictionaries whenever it is clear which ones are meant. Using these definitions we can further simplify the above estimates to get:

$$\begin{aligned} i \in \Lambda: \ |\langle y, \psi_i\rangle| &\geq |x_i|\beta - \|x\|_\infty \tilde{\mu}_1(K-1) \\ k \notin \Lambda: \ |\langle y, \psi_k\rangle| &\leq \|x\|_\infty \tilde{\mu}_1(K). \end{aligned}$$

Finally the combination of these two estimates leads to the following theorem.

Theorem 2.2.1. *Let y be a signal exactly K-sparse in $\mathbf{\Phi}$, i.e. $y = \sum_{i\in\Lambda} x_i\varphi_i$. Thresholding with the sensing matrix $\mathbf{\Psi}$ is able to recover a component φ_i of the true support if*

$$\frac{|x_i|}{\|x\|_\infty} > \frac{1}{\beta}(\tilde{\mu}_1(K) + \tilde{\mu}_1(K-1)) := \nu(K, \mathbf{\Phi}, \mathbf{\Psi}). \tag{2.6}$$

This is a relaxation over the traditional recovery condition (2.2) if

$$\frac{1}{\beta}(\tilde{\mu}_1(K) + \tilde{\mu}_1(K-1)) < \mu_1(K) + \mu_1(K-1).$$

The obvious questions now are: Given a dictionary $\mathbf{\Phi}$, do there exist complementary sensing dictionaries that give a relaxed recovery condition and if yes how do we find them or rather how do we find the best. Since we want to have the new recovery condition as relaxed as possible we need to find the dictionary for which the recovery coefficient $\nu(K, \mathbf{\Phi}, \mathbf{\Psi})$ is minimal, i.e.

$$\mathbf{\Psi}_0 = \arg\min_{\mathbf{\Psi}} \nu(K, \mathbf{\Phi}, \mathbf{\Psi}). \tag{2.7}$$

Consequently, unless the minimum in the above equation is attained by the dictionary itself, there will always exist better sensing dictionaries. The next subsection is dedicated to developing an algorithm for finding one of them.

2.2.2 An Algorithm for Calculating Sensing Dictionaries

If we wanted to find the optimal sensing dictionary we would have to find the solution to Problem (2.7). This a daunting task as is more clearly demonstrated by looking at the expansion of the objective function after back-inserting the definitions:

$$\min_{\mathbf{\Psi}} \frac{1}{\min_i |\langle \varphi_i, \psi_i \rangle|} \Big(\max_{|J|=K, i \notin J} \sum_{j \in J} |\langle \varphi_j, \psi_i \rangle| + \max_{|J|=K-1, i \notin J} \sum_{j \in J} |\langle \varphi_j, \psi_i \rangle| \Big).$$

Another complication arises from the fact that we may not know the exact sparsity of our signals as this can vary but only its order of magnitude.
Our approach to solving the problem is inspired by the alternative projection method in [57] for constructing equiangular tight frames. The problem of trying to find a sensing matrix $\mathbf{\Psi}$ for the dictionary $\mathbf{\Phi}$ that gives low cumulative coherence can be reformulated as looking for the Gram type matrix $G = \mathbf{\Psi}^\star \mathbf{\Phi}$ closest to the ideal Gram matrix, which by Theorem 2.1.2 has only ones on the diagonal and all off diagonal entries of absolute value $\mu = \sqrt{\frac{K-d}{d(K-1)}}$. So if we define

$$\mathcal{G} := \{G = \mathbf{\Psi}^\star \mathbf{\Phi},\ \mathbf{\Psi} \text{ a } K \times d \text{ matrix}\}$$
$$\mathcal{H} := \{H,\ \text{a } K \times K \text{ matrix with } H_{ii} = 1 \text{ and } |H_{ij}| \leq \mu \text{ for } i \neq j\}$$

and equip the space of all $N \times N$ matrices with the Frobenius norm we can write the problem as

$$\min \|G - H\|_F \ \text{ s.t } \ G \in \mathcal{G}, H \in \mathcal{H}, \tag{2.8}$$

which can be solved via projection onto convex sets (POCS) since both sets $\mathcal{G}$ and $\mathcal{H}$ are convex, see [57] for details. In our case POCS will do the following. We fix a number of iterations, initialise $G = \mathbf{\Phi}^\star \mathbf{\Phi}$ and then in each iterative step do:

a. find $H \in \mathcal{H}$ that minimises $\|G - H\|_F$

b. find $G \in \mathcal{G}$ that minimises $\|H - G\|_F$

After the last iteration we can extract our sensing dictionary from the matrix G, which by definition is of the form $\mathbf{\Psi}^\star \mathbf{\Phi}$. Let us now find explicit expressions for the projection of a matrix A onto $\mathcal{H}$

and $\mathcal{G}$. By writing out the Frobenius norm explicitely

$$\min_{H\in\mathcal{H}} \|A-H\|_F = \min_{H\in\mathcal{H}} \big(\sum_{ij} |A_{ij}-H_{ij}|^2\big)^{\frac{1}{2}} \tag{2.9}$$

we see that the minimum is attained for the matrix H with

$$H: \begin{cases} H_{ii}=1 & \\ H_{ij}=A_{ij} & \text{if } |A_{ij}|\leq \mu \\ H_{ij}=\operatorname{sgn}(A_{ij})\mu & \text{if } |A_{ij}|>\mu \end{cases}.$$

The solution to the second minimisation problem is not much harder to find. If we write $A^\star = (a_1 \dots a_K)$ we can rewrite the problem

$$\min_{G\in\mathcal{G}} \|A-G\|_F = \min_{\mathbf{\Psi}} \|A-\mathbf{\Psi}^\star\mathbf{\Phi}\|_F = \min_{\mathbf{\Psi}} \|A^\star-\mathbf{\Phi}^\star\mathbf{\Psi}\|_F = \min_{\mathbf{\Psi}} (\sum_i \|a_i-\mathbf{\Phi}^\star\psi_i\|_2^2)^{\frac{1}{2}}.$$

From the last expression it is clear that we should choose $\psi_i = (\mathbf{\Phi}^\star)^\dagger a_i$, leading to $\mathbf{\Psi}^\star = A\mathbf{\Phi}^\dagger$ and $G = A\mathbf{\Phi}^\dagger\mathbf{\Phi}$. Before testing the algorithm numerically note that in case the dictionary was a basis we have $K=d$ resulting in $\mu=0$. The set $\mathcal{H}$ consequently only contains the identity matrix and so in one iteration the algorithm will find the best sensing dictionary - the biorthogonal basis.

2.2.3 Simulations

First we calculated sensing dictionaries for three dictionaries of different types to compare the cumulative coherences and cross-coherences. To simplify the comparison we will 'hide' β within the correlations and choose the normalisation of the atoms in $\mathbf{\Psi}$ such that $|\langle\varphi_i,\psi_i\rangle| = \beta = 1$. The first dictionary was a random dictionary, of redundancy $R=2$ in dimension $d=128$. So in every atom the entries were drawn independently from a normalised standard Gaussian distribution and then the atom was rescaled to have unit norm. The second dictionary was a Gabor dictionary made up of the time-frequency shifts of one atom φ, i.e. $\mathbf{\Phi} = (\varphi_{n,m})_{n,m}$ where $\varphi_{n,m}(k) = e^{2\pi imbk}\varphi(k-na)$. In our case this atom was a normalised standard Gaussian in dimension $d=120$ and the time and frequency shift parameters were chosen as $a=8$, $b=10$, leading to a redundancy $R=1.5$. The third dictionary was the union of two orthonormal bases, the Haar-wavelet basis and the Discrete Cosine Transform (DCT) basis in dimension $d=128$.

Looking at Figure 2.1 we see that for the random dictionary, (a), the cross coherence is significantly lower than the coherence. We already have $\mu_1(S)>1$ for $S>3$ meaning that we can only guarantee to recover super positions of up to two atoms with equal absolute coefficients. On the other hand $\tilde{\mu}_1(4)+\tilde{\mu}_1(3)<1$ meaning we can recover super-positions of up to 4 atoms. Also for the Gabor dictionary, (b), there is a slight improvement so while $\mu(3)>1$ we still have $\tilde{\mu}(3)<1$. For the Haar-DCT dictionary, (c), we still observe the slower growth of the cross-coherence but in this case the difference is not large enough to change the worst case behaviour, i.e. $1<\tilde{\mu}(2)<\mu(2)$.

As second part of the simulations we tested how the sensing dictionaries performed in average for Thresholding. For every support size varying between 1 and 30 we constructed 500 signals by choosing the atoms in the support uniformly at random and coefficients of absolute value one with random signs in the case of the real dictionaries, i.e. the random and the Haar-DCT dictionary, and uniformly random angle $e^{i\theta}$ in case of the complex Gabor dictionary. We ran Thresholding using both the original and the sensing dictionary counting how often the full support could be recovered. The results are displayed in Figure 2.2.

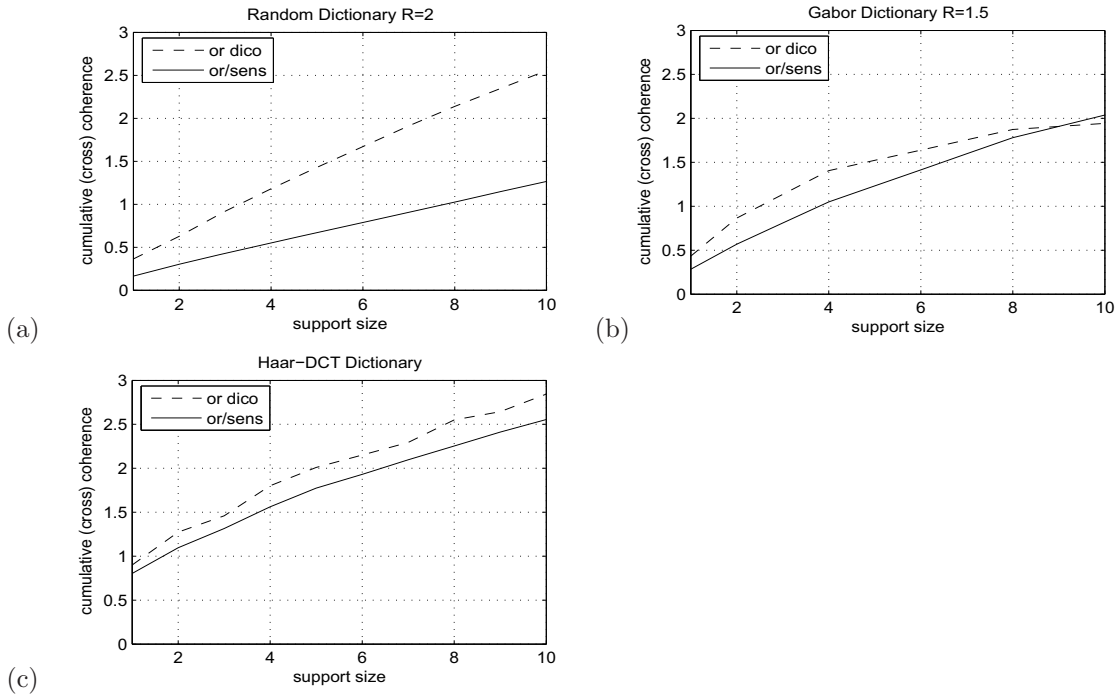

Figure 2.1: Cumulative coherence (or dico) and cross-coherence (or/dico) for various dictionaries.

As we can see while for both the random and the Gabor dictionary the recovery rates are higher when using the sensing dictionary there is no improvement for the Haar-DCT dictionary. One of the reasons might be that on average Thresholding for the Haar-DCT dictionary is already performing well. So comparing the original recovery rates of the random and the Haar-DCT dictionary, which have about the same redundancy, we observe a performance gap in favour of the Haar-DCT dictionary. However, the gap closes when using the sensing dictionary for the random matrix. Also note that in the above experiment we tested the average performance but used the sensing dictionaries that were designed to give a good worst case performance. Before discussing these issues more thoroughly in Section 2.4 let us investigate the use of sensing dictionaries for (O)MP.

2.3 Sensing Dictionaries for (O)MP

Even more clearly than Thresholding (O)MP can be decomposed into sensing and reconstruction steps. We initialise $a = 0$, $r = y$, $\Lambda = \emptyset$ and then in each step do:

Sensing:	find $i = \arg\max_j \lvert\langle r, \varphi_j\rangle\rvert$
Reconstruction:	$a = a + \langle r, \varphi_i\rangle\varphi_i$, $r = y - a$ (MP)
	$\Lambda = \Lambda \cup \{i\}$, $a = \mathbf{\Phi}_\Lambda \mathbf{\Phi}_\Lambda^\dagger y$, $r = y - a$ (OMP)

Table 2.3: (Orthogonal) Matching Pursuit

As before we can change the sensing step of the algorithm and, instead of trying to identify components of the true support with the dictionary $\mathbf{\Phi}$ itself, use a sensing dictionary $\mathbf{\Psi}$.

To determine which conditions we should impose on the sensing matrix for (O)MP we again do

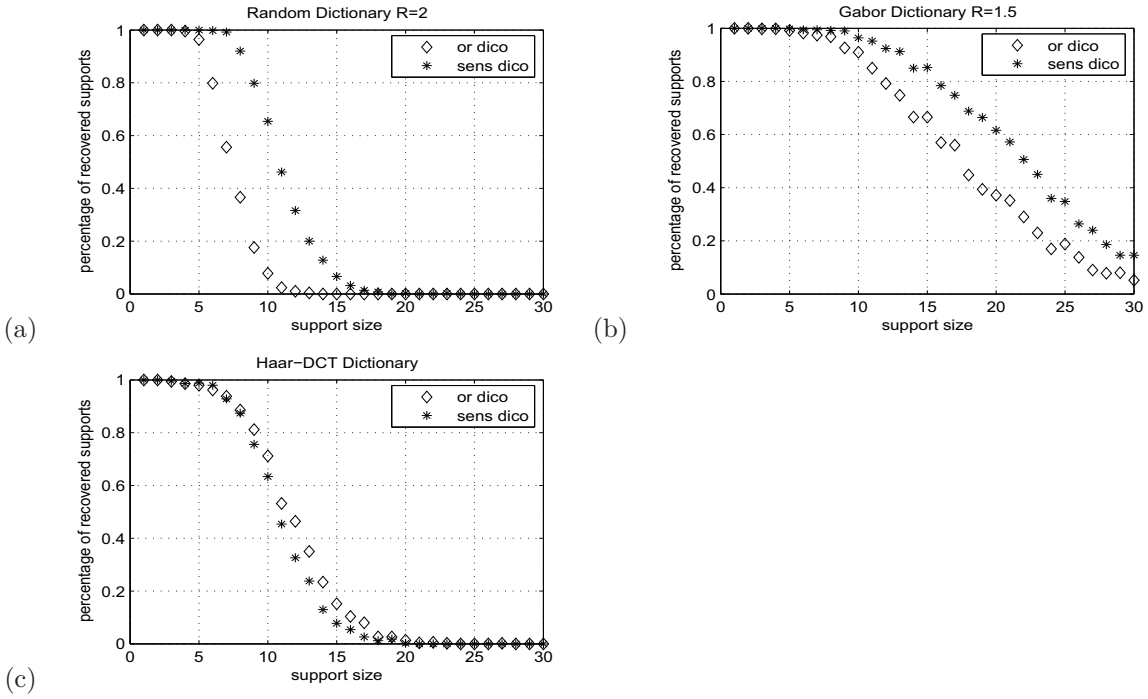

Figure 2.2: Recovery rates for Thresholding using the original dictionary (or dico) and the sensing dictionary (sens dico).

Sensing (new):	find $i = \arg\max_j \lvert\langle r, \psi_j\rangle\rvert$
Reconstruction:	$a = a + \langle r, \varphi_i\rangle\varphi_i$, $r = y - a$ (MP)
	$\Lambda = \Lambda \cup \{i\}$, $a = \mathbf{\Phi}_\Lambda \mathbf{\Phi}_\Lambda^\dagger y$, $r = y - a$ (OMP)

Table 2.4: (Orthogonal) Matching Pursuit with a Sensing Matrix

a worst case analysis.

2.3.1 Worst Case Analysis of (O)MP with a Sensing Dictionary

Theorem 2.3.1. *Let y be a signal exactly S-sparse in $\mathbf{\Phi}$, i.e. $y = \sum_{i\in\Lambda} x_i\varphi_i$. (Orthogonal) Matching Pursuing using the sensing matrix $\mathbf{\Psi}$ will always select components of the true support Λ if*

$$\|(\mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_\Lambda)^{-1}\mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_{\overline{\Lambda}}\|_{1,1} < 1 \tag{2.10}$$

which is always satisfied if

$$\tilde{\mu}_1(S) + \tilde{\mu}_1(S-1) < \beta. \tag{2.11}$$

Proof: Basically we just need to rewrite Tropp's proof for *Exact Recovery for OMP* in [55]. As long as we have only selected correct atoms we know that the residual r is still a linear combination of the atoms in the true support, i.e.

$$r = \sum_{i\in\Lambda} c_i\varphi_i = \mathbf{\Phi}_\Lambda c.$$

(O)MP will again select a correct atom at the next step if the maximal correlation of the residual with an atom in the support $\max_{i\in\Lambda}|\langle r,\psi_i\rangle|$ is larger than the maximal correlation with an atom outside the support $\max_{k\in\overline{\Lambda}}|\langle r,\psi_k\rangle|$. So we have to make sure that the quotient satisfies

$$\frac{\max_{k\in\overline{\Lambda}}|\langle r,\psi_k\rangle|}{\max_{i\in\Lambda}|\langle r,\psi_i\rangle|}=\frac{\|\boldsymbol{\Psi}_{\overline{\Lambda}}^{\star}r\|_\infty}{\|\boldsymbol{\Psi}_{\Lambda}^{\star}r\|_\infty}<1. \tag{2.12}$$

For further simplification we need to make use of p,q-matrix norms for $1\le p,q\le\infty$, defined as $\|A\|_{p,q}=\max_{\|x\|_p=1}\|Ax\|_q$. Inserting $r=\boldsymbol{\Phi}_\Lambda c$ into expression (2.12) and assuming that the matrix $\boldsymbol{\Psi}_\Lambda^\star\boldsymbol{\Phi}_\Lambda$ is invertible so that we can write $z=\boldsymbol{\Psi}_\Lambda^\star\boldsymbol{\Phi}_\Lambda c$, we can bound it as

$$\frac{\|\boldsymbol{\Psi}_{\overline{\Lambda}}^{\star}\boldsymbol{\Phi}_\Lambda c\|_\infty}{\|\boldsymbol{\Psi}_{\Lambda}^{\star}\boldsymbol{\Phi}_\Lambda c\|_\infty}=\frac{\|\boldsymbol{\Psi}_{\overline{\Lambda}}^{\star}\boldsymbol{\Phi}_\Lambda(\boldsymbol{\Psi}_\Lambda^\star\boldsymbol{\Phi}_\Lambda)^{-1}z\|_\infty}{\|z\|_\infty}\le\|\boldsymbol{\Psi}_{\overline{\Lambda}}^{\star}\boldsymbol{\Phi}_\Lambda(\boldsymbol{\Psi}_\Lambda^\star\boldsymbol{\Phi}_\Lambda)^{-1}\|_{\infty,\infty}.$$

Finally we note that $\|\boldsymbol{\Psi}_{\overline{\Lambda}}^{\star}\boldsymbol{\Phi}_\Lambda(\boldsymbol{\Psi}_\Lambda^\star\boldsymbol{\Phi}_\Lambda)^{-1}\|_{\infty,\infty}=\|(\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_\Lambda)^{-1}\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_{\overline{\Lambda}}\|_{1,1}$ which by condition (2.10) is smaller than one as required.
For the second part of the proof we just have to show that condition (2.11) implies condition (2.10). First we can estimate

$$\|(\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_\Lambda)^{-1}\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_{\overline{\Lambda}}\|_{1,1}\le\|(\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_\Lambda)^{-1}\|_{1,1}\|\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_{\overline{\Lambda}}\|_{1,1}.$$

The second term in the above can easily be bounded with the cross-coherence,

$$\|\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_{\overline{\Lambda}}\|_{1,1}=\max_{k\in\overline{\Lambda}}\sum_{i\in\Lambda}|\langle\varphi_i,\psi_k\rangle|\le\tilde{\mu}_1(K).$$

To bound the first term we use the fact that whenever $\|A\|_{1,1}<1$ we have $\|(\mathbf{I}+A)^{-1}\|_{1,1}<(1-\|A\|_{1,1})^{-1}$. Set $A=\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_\Lambda-\mathbf{I}$, then

$$\|A\|_{1,1}=\max_{i\in\Lambda}\left(|\langle\varphi_i,\psi_i\rangle-1|+\sum_{j\neq i}|\langle\varphi_i,\psi_j\rangle|\right)\le1-\beta+\tilde{\mu}_1(K-1),$$

and consequently

$$\|(\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_\Lambda)^{-1}\|_{1,1}\le(1-(1-\beta+\tilde{\mu}_1(K-1))^{-1}\le(\beta-\tilde{\mu}_1(K-1))^{-1}.$$

If we now combine these two estimates with condition (2.11) we get the desired bound

$$\|(\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_\Lambda)^{-1}\boldsymbol{\Phi}_\Lambda^\star\boldsymbol{\Psi}_{\overline{\Lambda}}\|_{1,1}\le\frac{\tilde{\mu}_1(K)}{\beta-\tilde{\mu}_1(K-1)}<1.$$

□

The theorem above is applicable to both MP and OMP as we only used that in each step the residual is a linear combination of the atoms in the support. Note, however, that picking a correct atom does not mean picking a new correct atom. Indeed since the sensing atoms corresponding to already found atoms are not orthogonal to the residual not even OMP can be guaranteed to find the full support in S steps.
As a consequence to Theorem 2.3.1 we get a characterisation of the optimal sensing dictionary for (O)MP. Given a dictionary $\boldsymbol{\Phi}$ and a sparsity level S, the best sensing dictionary $\boldsymbol{\Psi}_0$ is the solution

to:

$$\mathbf{\Psi}_0 = \arg\min_{\mathbf{\Psi}} \max_{|\Lambda|=S} \|(\mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_\Lambda)^{-1} \mathbf{\Phi}_\Lambda^\star \mathbf{\Psi}_{\overline{\Lambda}}\|_{1,1}. \tag{2.13}$$

Unfortunately solving this problem is even harder than solving the original problem of finding the best sensing dictionary for Thresholding in (2.7), as in addition to the maximum over all subsets of size K we also have to consider the inverse of a pseudo Gram matrix. However we still have the sufficient condition (2.11) for recovery success in terms of the cross coherence. Thus if we take a sensing dictionary calculated with the algorithm developed in Section 2.2.2 that has cross-coherence smaller than the coherence we can at least guarantee recovery for signals with higher sparsity. Finally what remains to be done is to check wether these sensing dictionaries also improve the average case performance of OMP.

2.3.2 Simulations for OMP

For our simulations we used the same three dictionaries and sensing dictionaries as for tresholding and the same set up. So for every support size varying between 10 and 40 we constructed 500 signals in the same way as for Thresholding. Then we ran OMP using both the original and the sensing dictionary counting how often the full support could be recovered. The results are displayed in Figure 2.3.

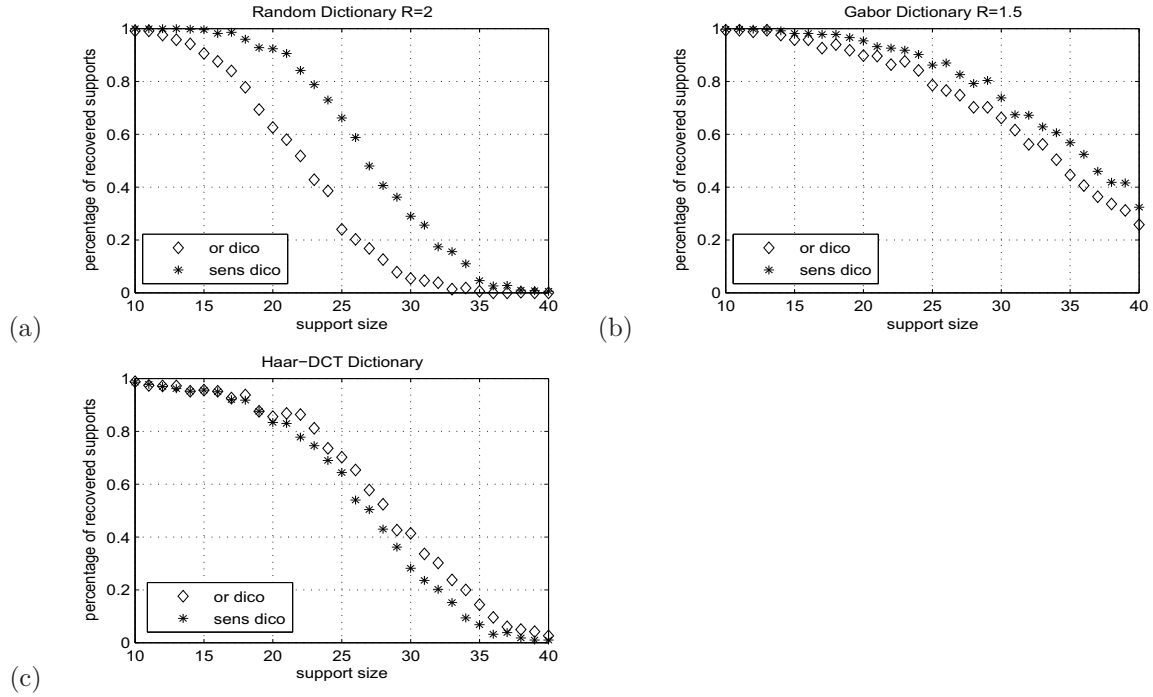

Figure 2.3: Recovery Rates for OMP using the original dictionary (or dico) and the sensing dictionary (sens dico).

Surprisingly even though the sensing matrices are derived from optimising only a sufficient worst case condition we can observe the same trends as for Thresholding. So for both the random and the Gabor dictionary the recovery rates are higher when using the sensing dictionary but there is no improvement for the Haar-DCT dictionary. Comparing the original recovery rates of the random and the Haar-DCT dictionary we observe the same performance gap in favour of the Haar-DCT

dictionary as for Thresholding. Again the gap closes when using the sensing dictionary for the random matrix.

2.4 Discussion

We have seen that using sensing dictionaries the performance of Thresholding and OMP can be improved, while maintaining the same computational complexity. The analysis of the worst case behaviour of both algorithms when using a sensing dictionary led to a characterisations of the optimal sensing dictionaries for worst case performance and with the developed algorithm we could even find good sensing dictionaries, i.e. with lower cumulative cross coherence than coherence, even though this difference is not always sufficiently large to guarantee a higher recovery rate. However with the numerical simulations we did not test the worst case but the average performance of both algorithms. The question is why in some cases the sensing dictionaries for good worst case performance also improve the average performance. There is a simple heuristic argument why the recovery rates increased for the random and the Gabor dictionary but not for the Haar-DCT dictionary. So for the random and the Gabor dictionary lowering the extreme correlations that are contributing to the cumulative coherence went together with lowering all the correlations, while for the Haar-DCT dictionary lowering the extremal correlations came at the price of increasing some of the a priori small correlations. Figure 2.4 showing the Gram matrices $\mathbf{\Phi}^\star\mathbf{\Phi}$ and pseudo Gram matrices $\mathbf{\Psi}^\star\mathbf{\Phi}$ nicely illustrates this effect. For the Gabor dictionary the first off-diagonal band corresponding to the highest correlations is lower for the pseudo-Gram matrix, which in turn has larger correlations on the second to fourth off-diagonal band. Also for the Haar-DCT Dictionary we see that the correlations in the upper right and lower left corner of the pseudo Gram matrix are lower than in the Gram matrix but that as price to pay there are non zero-correlations in the upper left and lower right part of $\mathbf{\Psi}^\star\mathbf{\Phi}$ that do not appear in $\mathbf{\Phi}^\star\mathbf{\Phi}$.

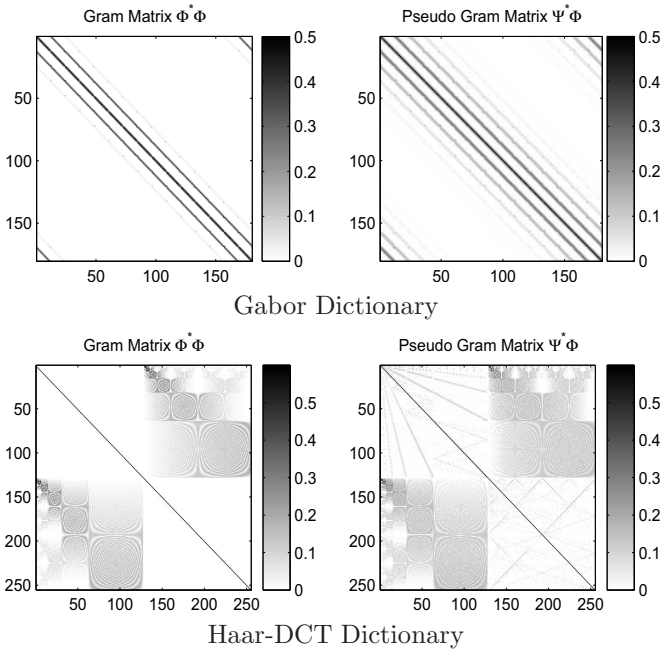

Figure 2.4: Gram and Pseudo Gram Matrices.

The next chapter will shed more light on the question how Thresholding performs on average and how to find good average sensing dictionaries.

Average Performance Analysis for Thresholding

3

This chapter shows that with high probability Thresholding can recover signals that are sparse in a redundant dictionary as long as the *2-Babel function* is growing slowly. This implies that it can succeed for sparsity levels up to the order of the ambient dimension. The theoretical bounds are illustrated with numerical simulations. As an application of the theory sensing dictionaries for optimal average performance are characterised and their performance is tested numerically. The major part of the findings presented in this chapter has been published in [49].

3.1 Why Average Performance?

In the last chapter we introduced two greedy algorithms for finding sparse approximation, Thresholding and (O)MP, which together with the Basis Pursuit Principle, see Table 3.1 and [21] for more details, are among the most popular in the signal processing community. However, while they are successfully employed to find sparse approximations in practice, the theoretical analysis of these algorithms was so far limited to studying their worst case performance. We also did a worst case analysis to study the performance of sensing dictionaries. The problem with the resulting worst case bounds for recoverable sparsity levels is that they are over-pessimistic and quite in contrast to the much better performance in practice. So the worst case analysis tells us that we can recover superpositions of S atoms as long as:

$$S \lesssim \mu^{-1} \approx \sqrt{d},$$

while in practice it is usually possible to recover supports sizes of the order of d. This phenomenon could also be seen in the simulation results in Subsection 2.2.3 of the last chapter. From a worst case point of view we were for instance able to recover super-positions of 2 atoms in the Gabor dictionary but the numerical simulations, testing the average performance, showed that it was always possible to recover 5 atoms and in more than 90% of the cases even up to 10 atoms.
Motivated by the desire to better understand and capture the performance of an algorithm together with a dictionary people have started to analyse the average case performance. In a recent paper, [56], Tropp was able to show that random subdictionaries of a general dictionary are very likely to be well conditioned as long as their size is of the order of $\mu^{-2} \approx d$ (see Theorem B, [56]). As an application of this result it is shown that a signal constructed from a random superposition of S atoms with coefficients drawn from a continuous distribution has almost surely no sparser

representation (see Theorem 12, [56]). If additionally the signs of the coefficients are drawn from a uniform distribution then this representation is with high probability recoverable via Basis Pursuit, compare Table 3.1.

Replace the problem
$$\mathrm{P}(0) \quad \min \|x\|_0 \text{ s.t } \|y - \mathbf{\Phi}\mathbf{x}\|_2 \leq \varepsilon$$
which is not convex because $\|\cdot\|_0$ counting the number of non-zero entries is not convex with the convex problem
$$\mathrm{P}(1) \quad \min \|x\|_1 \text{ s.t } \|y - \mathbf{\Phi}\mathbf{x}\|_2 \leq \varepsilon$$
and hope that the solutions coincide.

Table 3.1: Basis Pursuit (Denoising if $\varepsilon > 0$)

Theorem 3.1.1 (Theorem 13 in [56])**.** *Assume that $\mathbf{\Phi}_\Lambda$ has least singular value $\sigma_{\min}(\mathbf{\Phi}_\Lambda) \geq \sqrt{1/2}$ and that the signal $y = \mathbf{\Phi}_\Lambda \mathbf{x}_\Lambda$ is synthesised from a coefficient sequence x_Λ whose signs form a Steinhaus sequence, i.e. $\sigma_i = x_i/|x_i|$, $i \in \Lambda$ are independent realisations of the random variable e^{iX} with X uniformly distributed on $(0, 2\pi)$. Then the probability that Basis Pursuit fails to recover $\mathbf{x}_\Lambda$ from y satisfies*

$$\mathbb{P}(BP \text{ fails}) \leq 2K \exp\Big(-\frac{1}{8\mu^2 S}\Big) \tag{3.1}$$

One of the conclusions of the above results is that Basis Pursuit is able to recover sparse signal representations even when the sparsity level is higher than the worst case barrier of $\sqrt{d}$. However the problem is that in practice Basis Pursuit is simply too complex. Consider for instance image compression, a small picture of size 64×64 already results in $d = 4096$. Taking a dictionary with reasonable redundancy 2 means that we have to solve a convex optimisation problem in $\mathbb{R}^{8192}$. On the other hand one would typically be happy to recover the 100 most important components of the signal. Unfortunately this is still more than $64 = \sqrt{d}$ signifying the worst case performance bottleneck for simpler algorithms like thresholding or the Matching Pursuits. In the following we will therefore analyse the average behaviour of thresholding to find out that also here the recoverable sparsity scales with the ambient dimension. Again the result will be in terms of the coherence μ or rather the 2-Babel function μ_2, defined as

$$\mu_2(\Lambda, k) = \big(\sum_{i\in\Lambda} |\langle \varphi_i, \varphi_k\rangle|^2\big)^{\frac{1}{2}}, \qquad \mu_2(\Lambda) = \max_{k\notin\Lambda} \mu_2(\Lambda, k), \qquad \mu_2(S) = \max_{|\Lambda|=S} \mu_2(\Lambda). \tag{3.2}$$

3.2 Theoretical Analysis

To do an average analysis we first need to introduce the probabilistic model we assume for our signals y.

Signal Model:

$$y = \mathbf{\Phi}_\Lambda \mathbf{x}_\Lambda = \sum_{i\in\Lambda} \mathbf{x}_i \varphi_i, \qquad \mathbf{x}_i = \sigma_i |\mathbf{x}_i|, \quad \forall i \in \Lambda,$$

where $\mathbf{\Phi}$ is a dictionary of K normalised atoms and $\mathbf{\Phi}_\Lambda$ a subdictionary of all atoms with indices in Λ and $|\Lambda| = S$. While the support Λ and the absolute magnitude of the coefficients are considered to

be arbitrary, the signs σ_i form either a Steinhaus sequence or a Rademacher sequence, i.e. $\sigma_i = \pm 1$ with equal probability.

Theorem 3.2.1. *Let's abbreviate the event 'Thresholding fails to recover the component φ_i' as '$\circledcirc_i$' and 'Thresholding fails to recover all components' as '$\circledcirc$'. Under the above signal model*

$$a) \qquad \mathbb{P}(\circledcirc_i) < 2(K-S+1)\exp\Big(-\frac{|\mathbf{x}_i|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big)$$

$$b) \qquad \mathbb{P}(\circledcirc) < 2K\exp\Big(-\frac{|\mathbf{x}_{\min}|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big)$$

where $c = 1$ for Steinhaus and $c = 1/16$ for Rademacher sequences.

The proof is a straightforward application of the following large deviation inequalities.

Theorem 3.2.2. *Let α be a real/complex vector and σ a Rademacher/Steinhaus sequence. Then for all $t > 0$*

$$\mathbb{P}(|\sum_i \sigma_i\alpha_i| > t) \leq 2e^{-c_0t^2/\|\alpha\|_2^2}$$

where $c_0 = 1/32$ for Rademacher and $c_0 = 1/2$ for Steinhaus sequences.

For a proof for Steinhaus sequences see [56] and references therein. The proof for Rademacher sequences can be found in Section 4 of [31].

Proof: [Theorem 3.2.1] We can bound the probability of not recovering φ_i by the probability that its inner product with the signal is lower than a threshold p or the inner product of an atom not in the support is higher than the threshold.

$$\begin{aligned}\mathbb{P}(\circledcirc_i) &\leq \mathbb{P}\big(|\langle y,\varphi_i\rangle| < \max_{k\in\overline{\Lambda}}|\langle y,\varphi_k\rangle|\big)\\ &\leq \mathbb{P}\big(|\langle y,\varphi_i\rangle| < p\big) + \mathbb{P}\big(\max_{k\in\overline{\Lambda}}|\langle y,\varphi_k\rangle| > p\big)\\ &\leq \mathbb{P}\big(|\langle y,\varphi_i\rangle| < p\big) + \mathbb{P}\big(\bigcup_{k\in\overline{\Lambda}}|\langle y,\varphi_k\rangle| > p\big) \leq \mathbb{P}\big(|\langle y,\varphi_i\rangle| < p\big) + \sum_{k\in\overline{\Lambda}}\mathbb{P}\big(|\langle y,\varphi_k\rangle| > p\big)\end{aligned}$$

The probability of the correlation of the signal with φ_i being smaller than the threshold can be further bounded as,

$$\begin{aligned}\mathbb{P}\big(|\langle y,\varphi_i\rangle| < p\big) &= \mathbb{P}\big(|\sum_{j\in\Lambda}\mathbf{x}_j\langle\varphi_j,\varphi_i\rangle| < p\big)\\ &= \mathbb{P}\big(|\mathbf{x}_i + \sum_{j\neq i}\mathbf{x}_j\langle\varphi_j,\varphi_i\rangle| < p\big) \leq \mathbb{P}\big(|\sum_{j\neq i}\mathbf{x}_j\langle\varphi_j,\varphi_i\rangle| > |\mathbf{x}_i| - p\big).\end{aligned}$$

Choosing the threshold as $p = |\mathbf{x}_i|/2$ and using Theorem 3.2.2 we arrive at,

$$\begin{aligned}\mathbb{P}\big(|\langle y,\varphi_i\rangle| \leq p\big) &< \mathbb{P}\big(|\sum_{j\neq i}\sigma_j|\mathbf{x}_j|\langle\varphi_j,\varphi_i\rangle| > \tfrac{1}{2}|\mathbf{x}_i|\big)\\ &\leq 2\exp\Big(-\frac{c_0}{4}\frac{|\mathbf{x}_i|^2}{\sum_{j\neq i}|\mathbf{x}_j|^2|\langle\varphi_j,\varphi_i\rangle|^2}\Big) \leq 2\exp\Big(-\frac{|\mathbf{x}_i|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S-1)}\Big).\end{aligned}$$

Similarly we can estimate the probability of the correlation of an atom not in the support being larger than the threshold,

$$\begin{aligned}\mathbb{P}\big(|\langle y,\varphi_k\rangle|>p\big) &\le \mathbb{P}\big(|\sum_{j\in\Lambda}\sigma_j|\mathbf{x}_j|\langle\varphi_j,\varphi_k\rangle|>\tfrac{1}{2}|\mathbf{x}_i|\big)\\ &\le 2\exp\Big(-\frac{c_0}{4}\frac{|\mathbf{x}_i|^2}{\sum_{j\in\Lambda}|\mathbf{x}_j|^2|\langle\varphi_j,\varphi_k\rangle|^2}\Big)\le 2\exp\Big(-\frac{|\mathbf{x}_i|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big).\end{aligned}$$

Putting it all together we finally arrive at,

$$\begin{aligned}\mathbb{P}(\circledcirc_i) &\le 2\exp\Big(-\frac{|\mathbf{x}_i|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S-1)}\Big)+|\overline{\Lambda}|2\exp\Big(-\frac{|\mathbf{x}_i|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big)\\ &\le 2(K-S+1)\exp\Big(-\frac{|\mathbf{x}_i|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big).\end{aligned}$$

To estimate the probability of thresholding failing to recover all components we can proceed in the same fashion. Essentially we just need to adapt the choice of the threshold p.

$$\mathbb{P}(\circledcirc)=\mathbb{P}\big(\min_{i\in\Lambda}|\langle y,\varphi_i\rangle|<\max_{k\in\overline{\Lambda}}|\langle y,\varphi_k\rangle|\big)\le\mathbb{P}\big(\min_{i\in\Lambda}|\langle y,\varphi_i\rangle|<p\big)+\mathbb{P}\big(\max_{k\in\overline{\Lambda}}|\langle y,\varphi_k\rangle|>p\big).$$

The first probability can be expanded as

$$\begin{aligned}\mathbb{P}\big(\min_{i\in\Lambda}|\langle y,\varphi_i\rangle|<p\big) &\le \mathbb{P}\big(\min_{i\in\Lambda}|x_i+\sum_{j\neq i}\mathbf{x}_j\langle\varphi_j,\varphi_i\rangle|<p\big)\\ &\le \mathbb{P}\big(\min_{i\in\Lambda}(|x_{\min}|-|\sum_{j\neq i}\mathbf{x}_j\langle\varphi_j,\varphi_i\rangle|)<p\big)\\ &\le \mathbb{P}\big(\max_{i\in\Lambda}|\sum_{j\neq i}\mathbf{x}_j\langle\varphi_j,\varphi_i\rangle|>|x_{\min}|-p\big)\\ &\le \sum_{i\in\Lambda}\mathbb{P}(|\sum_{j\neq i}\mathbf{x}_j\langle\varphi_j,\varphi_i\rangle|>|x_{\min}|-p)\end{aligned}$$

Now we choose as threshold $p=|\mathbf{x}_{\min}|/2$ and using Theorem 3.2.2 get the bound:

$$\mathbb{P}\big(\min_{i\in\Lambda}|\langle y,\varphi_i\rangle|<p\big)\le 2S\exp\Big(-\frac{|\mathbf{x}_{\min}|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S-1)}\Big).$$

Repeating the steps above we can estimate the probability of an atom not in the support having higher correlation than the threshold as

$$\mathbb{P}\big(\max_{k\in\overline{\Lambda}}|\langle y,\varphi_k\rangle|>p\big)\le 2(K-S)\exp\Big(-\frac{|\mathbf{x}_{\min}|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big).$$

In combination this leads to the final bound:

$$\mathbb{P}(\circledcirc)<2K\exp\Big(-\frac{|\mathbf{x}_{\min}|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\mu_2^2(S)}\Big).$$

□

Comparing the above result for Steinhaus sequences to Theorem 3.1.1 we see that the essential

difference in the failure probability bound for the two algorithms is the additional coefficient $\frac{\mathbf{x}_{\min}^2}{\|\mathbf{x}\|_\infty^2}$ in the exponent for thresholding. This means that for coefficients of constant absolute magnitude the two algorithms should perform comparably. Also it promises a good behaviour of thresholding as long as the coefficients are reasonably well balanced and in that case makes it an interesting low complexity alternative to BP.

3.3 Applications & Numerical Simulations

3.3.1 An Experiment with Dimensions

To show numerically how the recovery rates of thresholding scale with the dimension we conducted the following experiment. In dimensions 2^p, $p = 8 \dots 12$ a dictionary made up of the Dirac and the Discrete Cosine Transform bases was constructed. The coherence of these dictionaries is $\mu = \sqrt{2/d}$ and the 2-Babel function behaves approximately like $\mu_2(S) \approx \sqrt{S/d}$. For each dimension and relative sparsity level S/d, 1000 signals were constructed by randomly choosing a support and coefficients with constant absolute value one and random signs, $\mathbf{x}_i = \pm 1$ with equal probability. Then we counted how often thresholding was able to recover the full support.

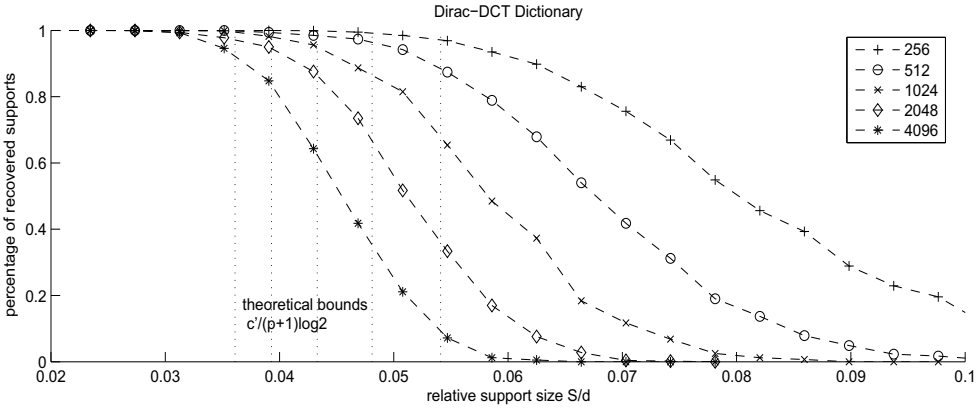

Figure 3.1: Comparison of Numerical Recovery Rates and Theoretical Recovery Bounds

From the theorem we know that thresholding will fail with small probability as long as

$$\mu_2^2(S) \lesssim \frac{c'}{\log(2K)} \qquad \Rightarrow \qquad \frac{S}{d} \lesssim \frac{c'}{(p+1)\log 2}.$$

If we compare these theoretical bounds to the simulation results displayed in Figure 3.1 we see that they reflect the average behaviour quite well. For the bounds as plotted in the figure we chose $c' = 0.3$ which is somewhat better than the theorem suggests ($c \approx \frac{1}{128}$).

3.3.2 An Application

As an application of Theorem 3.2.1 we will construct a sensing dictionary to improve the average performance of a dictionary for Thresholding, as promised at the end of the last chapter. The average performance of thresholding with a sensing dictionary can be analysed as before. We only need to adjust the definition of the 2-Babel function to describe the pseudo Gram matrix $\mathbf{\Psi}^\star\mathbf{\Phi}$ instead of the Gram matrix.

$$\tilde{\mu}_2(\Lambda,k)=\big(\sum_{i\in\Lambda}|\langle\varphi_i,\psi_k\rangle|^2\big)^{\frac{1}{2}},\qquad \tilde{\mu}_2(\Lambda)=\max_{k\notin\Lambda}\tilde{\mu}_2(\Lambda,k),\qquad \tilde{\mu}_2(S)=\max_{|\Lambda|=S}\tilde{\mu}_2(\Lambda). \tag{3.3}$$

The analogue of part *b*) of Theorem 3.2.1 now reads:

Theorem 3.3.1. *Under the same assumptions on the signal model as in the previous section we can bound the probability that thresholding with the sensing matrix $\mathbf{\Psi}$ fails as*

$$\mathbb{P}(\circledcirc)<2K\exp\Big(-\frac{|\mathbf{x}_{\min}|^2}{\|\mathbf{x}\|_\infty^2}\frac{c}{8\tilde{\mu}_2^2(S)}\Big).$$

Proof: Follow the proof of Theorem 3.2.1 mutatis mutandis. □

One deduction from the Theorem is that a sensing matrix for good average performance should minimise the 2-Babel function. However if we also assume that the support Λ is picked at random we see that all the squared off-diagonal entries are equally likely to contribute to the final bound. A simplified but sensible approach would therefore be to find the sensing dictionary that minimises the Frobenius norm of the pseudo-Gram matrix.

$$\mathbf{\Psi}_0=\underset{\langle\psi_i,\varphi_i\rangle=1}{\arg\min}\|\mathbf{\Psi}^\star\mathbf{\Phi}\|_F=\underset{\langle\psi_i,\varphi_i\rangle=1}{\arg\min}\big(\sum_i\sum_j|\langle\varphi_i,\psi_j\rangle|^2\big)^{\frac{1}{2}}.$$

The advantage of the problem as formulated above is that there exists an analytic solution, that can be easily derived using Lagrange multipliers. To make our lives easier we consider the square of the objective function $\|\mathbf{\Psi}^\star\mathbf{\Phi}\|_F^2$.

$$\begin{aligned}\frac{d}{d\psi_j}\|\mathbf{\Psi}^\star\mathbf{\Phi}\|_F^2&=\sum_i 2\langle\varphi_i,\psi_j\rangle\varphi_i=2\mathbf{\Phi}\mathbf{\Phi}^\star\psi_j\\ \frac{d}{d\psi_j}\langle\varphi_j,\psi_j\rangle&=\varphi_j\\ 2\mathbf{\Phi}\mathbf{\Phi}^\star\psi_j=c_j\varphi_j\qquad&\Rightarrow\psi_j=\tfrac{c_j}{2}(\mathbf{\Phi}\mathbf{\Phi}^\star)^{-1}\varphi_j.\end{aligned}$$

If we choose the constants c_j appropriately to ensure $\langle\varphi_j,\psi_j\rangle=1$ and collect them in the diagonal matrix D, we see that the optimal sensing matrix is just the rescaled transpose of the Moore Penrose pseudo inverse,

$$\mathbf{\Psi}_0=(\mathbf{\Phi}\mathbf{\Phi}^\star)^{-1}\mathbf{\Phi}D=(\mathbf{\Phi}^\dagger)^\star D.$$

To test the performance of an average sensing matrix we did the following small experiment. We built a dictionary of 256 atoms that are randomly distributed on the sphere in $\mathbb{R}^{128}$. For each support size between 1 and 20 we constructed 1000 signals by choosing the support set uniformly at random and coefficients of absolute value one but with random signs, i.e. $x_i=\pm1$ with equal probability. We then compared how often thresholding could recover the full support when using the original dictionary, the worst case sensing matrix, see [50], and the average case sensing matrix. The results are displayed in Figure 3.2

The improvement already gained by using the worst case sensing matrix is further increased by using the average case sensing matrix. The performance differences are also well reflected by the Frobenius norms of the (pseudo-) Gram matrices in Table 3.3.2.

So there is a large decrease in norm between the original dictionary and the worst case sensing

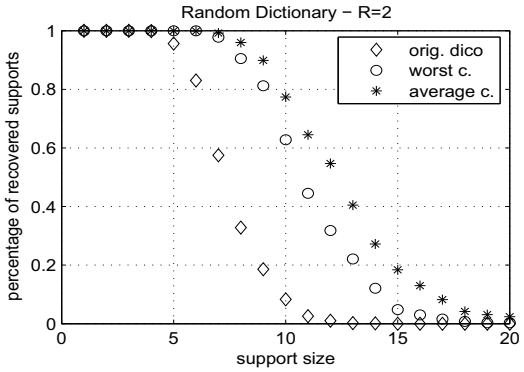

Figure 3.2: Recovery Rates for Different Sensing Dictionaries

dictionary	original	worst case	average case
$\|\mathbf{\Psi}^\star\mathbf{\Phi}\|_F$	27.7217	23.8902	22.6743

Table 3.2: Frobenius norms of (pseudo-) Gram matrices

matrix accounting for the large performance gap and a smaller decrease between the worst case and the average case sensing matrix reflecting a smaller improvement.
Considering that from a worst case point of view (O)MP is a more powerful algorithm that Thresholding, we would expect that also its average performance is better than that of Thresholding, allowing us to recover super-positions of a number of atoms scaling with the dimension. Unfortunately in the case of a single signal we do not have a comparable result. However, in the next chapter we will see that in case we want to sparsely approximate not one signal but several signals at the same time we can show that not only Thresholding but also (O)MP can on average recover support sizes of the order of the ambient dimension.

Average Case Analysis of Multi-Channel Greedy Algorithms

4

In this chapter we generalise the Thresholding and OMP algorithms to find simultaneous sparse approximations of multichannel signals and using a random model analyse when they are likely to succeed with high probability. All the results presented in this chapter and more have been published in [25].

4.1 Multi-Channel Greedy Algorithms

In the first chapters we studied two greedy algorithms to calculate sparse signal representations in a redundant dictionary. Here we will generalise both of them to calculate simultaneous sparse approximations for multi-channel signals. First let us explain what a multi-channel signal is and why we would be interested in a simultaneous sparse approximation. Assume that we have a network of sensors monitoring a common phenomenon. Let's give a not so serious example, for a more serious one see for instance [33]. We give a banana to a monkey and just from observing his EEG want to be able to say 'oh the monkey is thinking about bananas'. The idea behind this is the following. The stimulus, the banana, activates several parts of the monkey's brain. The visual centre sends out a waveform saying yellow and another one saying long, the tactile centre generates the impulse for smooth, the taste centre sends out its 'yummy' signal and from somewhere in the memory there comes a waveform saying peel. All five waveforms now start propagating through the skull to the EEG electrode cap the monkey is wearing. However because of the distance of the various regions to the skull and thus different travel paths and electric properties of the brain these waveforms arrive at the different electrodes with varying magnitudes and signs. Also at the electrode closest to the tactile centre a faint waveform saying ticklish and referring to the EEG cap will arrive and at other electrodes similar noise waveforms from secondary thought processes. So while at each electrode we receive a different signal, they all consist of a superposition of the waveforms for yellow, long, smooth, yummy and peel with varying magnitudes plus some noise. We can easily translate this into the language of sparseness and dictionaries. An impulse or waveform sent out from one part of the brain can be modelled as one of the K elements φ_k of the dictionary $\mathbf{\Phi}$, of which the banana triggers only S ones in the support Λ. The fact that each impulse φ_k arrives at different electrodes with a different magnitude can be modelled by weighting its contribution to the EEG signal at electrode n with a coefficient $x_n(k)$. If we collect all the random thoughts arriving at electrode n but not

related to the banana in the noise impulse e_n we can write the received EEG signals as

$$y_n = \sum x_n(k)\varphi_k = \mathbf{\Phi}_\Lambda x_n + e_n,\ n = 1, ..., N. \tag{4.1}$$

For simplicity we will assume that the noise components are orthogonal to the banana part of the signal. If we collect the N signals y_n as columns in the signal matrix $Y = (y_1 \dots y_N)$, similarly the coefficients in $X = (x_1 \dots x_N)$ and the noise in $E = (e_1 \dots e_N)$, we can write compactly

$$Y = \mathbf{\Phi}_\Lambda X + E.$$

Following this model it is easy to find out what the monkey is thinking. We just need to identify the five waveforms in the support Λ that can be used to build the main part of all the signals, i.e. give us the best simultanous approximations to all signals. Once we have identified these impulses as yellow, long, smooth, yummy and peel it is easy to conclude banana. We also see that the more electrodes or channels we have the easier it should be to detect the main components instead of noise. So the electrode close to the tactile centre might just receive long smooth and ticklish quite strongly but yellow, yummy and peel only faintly. Similarly other electrodes will miss long or smooth or any subset of the five important impulses. Still if we collect enough signals at different electrode positions the best 5 atoms to approximate all signals together will be yellow, long, smooth, yummy and peel.
While the above example might seem far-fetched a similar model is actually used to detect EEG micro-states that help diagnose schizophrenia, see [54], and while one might not care whether the monkey is thinking about bananas or peanuts in this real case it becomes very important to find algorithms that correctly identify these micro-states, indicating schizophrenia. In this chapter we generalise the two greedy algorithms we met in the previous chapters to find simultaneous sparse approximations and analyse when they succeed in identifying the sparsest simultaneous approximation.

Both single channel greedy algorithms Thresholding and (O)MP were relying on the inner products between the signal/residual to approximate and the elements of the dictionary or a sensing dictionary. In analogy simultaneous greedy algorithms should rely on the inner products of the elements of the (sensing) dictionary with the signals in all channels.

single channel:		multi-channel:
$\langle y, \psi_k \rangle$	$\Rightarrow$	$\psi_i^\star Y = \begin{pmatrix} \langle y_1, \psi_k \rangle \\ \vdots \\ \langle y_N, \psi_k \rangle \end{pmatrix}$

To get a criterion which atom to choose we need to combine the entries in the correlation vector $\psi_i^\star Y$, for instance by taking a norm. In the following we will consider the p-norms,

$$\|\psi_k^\star Y\|_p := \left(\sum_{n=1}^{N} |\langle \psi_k, y_n \rangle|^p \right)^{1/p}, \tag{4.2}$$

where $p \geq 1$ and with the standard modification for $p = \infty$. With this definition p-Thresholding and p-Simultaneous Orthogonal Matching Pursuit (p-SOMP) can be derived directly from their single channel counterparts just by replacing $|\langle y, \psi_k \rangle|$ with $\|\psi_k^\star Y\|_p$ and the vectors a, r, y with the matrices A, R, Y. For a summary see Tables 4.1 and 4.1. The parameter p reflects how much we expect the contribution of the atoms across channels to be correlated. For $p = 1$ we expect high correlation and an atom will only be selected if it triggers a strong response averaged across all channels. For

$p = \infty$ we do not expect much correlation, an the atom that gives the strongest response in any channel will be selected. The choice of p thus depends on the user's a priori information about the signals or the application.

Sensing:	find Λ_M that contains the indices corresponding to the M largest values of $\|\psi_k^\star Y\|_p$
Reconstruction:	$A_M = \mathbf{\Phi}_\Lambda \mathbf{\Phi}_\Lambda^\dagger Y$

Table 4.1: p-Thresholding

Initialisation:	$R_0 = Y$, $A_0 = 0$, $\Lambda_0 = \emptyset$
Sensing:	find $k_M = \arg\max_k \|\psi_k^\star R_M\|_p$
Reconstruction:	$\Lambda_M = \Lambda_{M-1} \cup \{k\}$,
	$A_M = Y - \mathbf{\Phi}_{\Lambda_M} \mathbf{\Phi}_{\Lambda_M}^\dagger Y := \mathbf{P}_M Y$,
	$R_M = Y - A_M = (I - \mathbf{P}_M) Y$
where $\mathbf{P}_{\Lambda_M} = \mathbf{\Phi}_{\Lambda_M} \mathbf{\Phi}_{\Lambda_M}^\dagger$ is the orthogonal projection onto the linear span of the selected atoms.	

Table 4.2: p-Simultaneous Orthogonal Matching Pursuit

The question now is when the two algorithms are successful in finding the sparsest simultaneous approximation. As for the single channel versions this is equivalent to recovering the right support, i.e. when we set $M = |\Lambda| = S$, 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. $\Lambda_M \subset \Lambda$.

As in the single channel case we could start with a worst case analysis, to derive conditions under which both algorithms are sure to succeed. However, looking back to Chapter 2 we see that worst case analyses are not very exciting. Deriving the multi-channel results by generalising the arguments of the single channel analysis is a straight forward exercise and the calculations can for instance be found in [25]. Indeed the worst case result does not improve with the number of channels, which is counter intuitive. Also in Chapter 3 we have already seen that the results of worst case analyses rarely reflect the behaviour of an algorithm in practice well as they tend to be too pessimistic. Therefore we refer the interested readers to the above mentioned paper and here go directly to an average case analysis based on a random model of the sparse coefficient matrix.

Random Model

We will assume that in every channel the sparse coefficients follow a Gaussian distribution, i.e. the components $x_n(i)$, $i \in \Lambda$, of the random vectors x_n are independent Gaussian variables of variance α_i. This freedom in choosing the variances allows us to model the different strength of certain atoms when averaged across channels. The assumption that the coefficients are Gaussian is probably not necessary - a Bernoulli distribution or any other symmetric distribution having certain concentration properties, as described in Subsection 4.4.2, would likely give the same results - but will make the analysis easier and clearer. In order to keep the notational mess to a minimum we will translate the above definition in terms of vectors into matrices. If we let U be a $S \times N$ random matrix with independent standard gaussian entries and let D be a $S \times S$ diagonal matrix whose diagonal entries α_i^2 are positive real numbers our model can be written in the compact form:

$$Y = \mathbf{\Phi}_\Lambda \cdot D^{\frac{1}{2}} \cdot U + E, \tag{4.3}$$

With this random model we are almost ready to start. So in the next section we will introduce some notation and give reminders on how to deal with matrix norms, Babel functions and isometry constants. In Section 4.3 we present the main results, which of course should not prevent you from reading on but motivate you. Sections 4.4 and 4.5 contain the proofs of the main theorems, always starting with the idea before going into mathematical detail, and the last section is dedicated to some discussion.

4.2 Technical Tools and Notations

This section provides the main tools and notations necessary to state and prove our results.

4.2.1 Matrix Norms

For a neat analysis of the algorithms it will be convenient to redefine the matrix norms $\|\cdot\|_{p,\infty}$ for this chapter. Let A be a $n \times m$-matrix with rows $(A_i)_{1\dots n}$ then we define

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

To denote the operator norm which is normally denoted like this we will use the notation $\|\!|A|\!\|_{p\to\infty}$. For general $1 \le p, q \le \infty$ this operator norm is defined as:

$$\|\!|A|\!\|_{p\to q} = \max_{\|x\|_p=1} \|Ax\|_q. \tag{4.4}$$

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

$$\|A\|_{p,\infty} = \|\!|A|\!\|_{p'\to\infty}. \tag{4.5}$$

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,

$$\|\!|A|\!\|_{2\to 2} = \lambda_{\max}(A) = \text{largest singular value of } A. \tag{4.6}$$

Also we will write for shortness $\|\!|\cdot|\!\| := \|\!|\cdot|\!\|_{2\to 2}$. The following lemma collects two useful properties of operator norms. Proofs can be found in any standard linear algebra text book, e.g. [27].

Lemma 4.2.1. *a. For two matrices A, B we have*

$$\|\!|AB|\!\|_{p\to q} \le \|\!|B|\!\|_{p\to s} \|\!|A|\!\|_{s\to q}. \tag{4.7}$$

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

$$\|\!|A^\dagger|\!\|_{2\to 2} = \frac{1}{\lambda_{\min}(A)}, \tag{4.8}$$

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 chapter.

Corollary 4.2.2. *For two matrices A, B we have*

$$\frac{\|AB\|_{p,\infty}}{\|B\|_{p,\infty}} \leq \|\!|A|\!\|_{\infty\to\infty} = \|A\|_{1,\infty} = \max_{i=1..n} \sum_{j=1}^{m} |A_{ij}|. \tag{4.9}$$

4.2.2 Babel Functions and Isometry Constants

Even though we have already met the 1/2 (cross) Babel functions in the previous two chapters, we repeat the definition here to be on one hand more general and on the other hand more precise.

p-Babel functions.
For a pair of dictionaries $(\mathbf{\Phi}, \mathbf{\Psi})$ containing the same number of unit norm atoms and a support set Λ we define the p-Babel function

$$\mu_p(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda) := \sup_{\ell \notin \Lambda} \big(\sum_{j\in\Lambda} |\langle \varphi_j, \psi_\ell \rangle|^p \big)^{\frac{1}{p}} \tag{4.10}$$

which measures the amount of correlation between sensing atoms ψ_ℓ *outside* the support Λ and modeling atoms φ_j *inside* the support Λ. To capture also the amount of correlation between atoms *inside* the support Λ we define additionally

$$\mu_p^{in}(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda) := \sup_{i\in\Lambda} \mu_p(\mathbf{\Phi}_\Lambda, \mathbf{\Psi}_\Lambda, \Lambda\backslash\{i\}). \tag{4.11}$$

For the cases when we do not care to be very precise we again take the supremum over all possible subsets of size at most S to get the definition of the p-Babel function for an integer S as

$$\mu_p(\mathbf{\Phi}, \mathbf{\Psi}, S) := \sup_{|\Lambda|\leq S} \mu_p(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda). \tag{4.12}$$

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

$$\mu_p^{in}(\mathbf{\Phi}, \mathbf{\Psi}, S) \leq \mu_p(\mathbf{\Phi}, \mathbf{\Psi}, S-1). \tag{4.13}$$

Most interesting for us are the cases $p = 1$ and $p = 2$. In the rest of this chapter 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 $\mu_p^{in}(S)$.
Thinking back to Chapter 2 if we are dealing with a sensing dictionary different from the approximation dictionary we also need to consider the similarity between corresponding atoms in the two dictionaries. We define

$$\beta_k(\mathbf{\Phi}, \mathbf{\Psi}) := \langle \varphi_k, \psi_k \rangle > 0, \qquad \beta(\mathbf{\Phi}, \mathbf{\Psi}, \Lambda) := \min_{i\in\Lambda} \beta_i, \qquad \beta(\mathbf{\Phi}, \mathbf{\Psi}) := \min_k \beta_k. \tag{4.14}$$

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. Again we will omit the reference to the dictionary pair unless it is necessary.

Isometry constants have not been introduced before but are an important tool to characterise the conditioning of a subdictionary. We will meet them again in the next chapter when discussing

Compressed Sensing.

Isometry constants.
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

$$(1 - \delta_\Lambda) \cdot \|x\|_2^2 \leq \|\mathbf{\Phi}_\Lambda x\|_2^2 \leq (1 + \delta_\Lambda) \cdot \|x\|_2^2 \qquad \forall x \neq 0. \tag{4.15}$$

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

$$\lambda_{\min}(\mathbf{\Phi}_\Lambda) \geq \sqrt{1 - \delta_\Lambda} \qquad \text{and} \qquad \lambda_{\max}(\mathbf{\Phi}_\Lambda) \leq \sqrt{1 + \delta_\Lambda}, \tag{4.16}$$

where the first one is of course only valid if $\delta_\Lambda \leq 1$. 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

$$\delta_S := \sup_{|\Lambda|=S} \delta_\Lambda. \tag{4.17}$$

If for a dictionary the global (restricted) isometry constant is small, i.e. $\delta_S \ll 1$, we say that the dictionary satisfies a uniform uncertainty principle, cp. [9]. It is easy to check that δ_S is a non-decreasing function of S. Restricted isometry constants were introduced by Candès, Romberg and Tao in [8, 9] in order to study recovery by Basis Pursuit (ℓ_1) in the context of Compressed Sensing and we will meet them again in the next chapter. Good estimates of these numbers were obtained for random Gaussian and Bernoulli $d \times K$ matrices $\mathbf{\Phi}$: If

$$S \leq C_\delta \frac{d}{\log\left(\frac{K}{S\epsilon}\right)} \tag{4.18}$$

then with probability at least $1 - \epsilon$ the restricted isometry constant of $\mathbf{\Phi}$ satisfies $\delta_S \leq \delta$, see e.g. [4, 9, 46]. 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 [9, 45, 48].

4.3 Main Results

The analyses of both p-Thresholding and p-SOMP follow a similar route. First, we provide sufficient conditions which guarantee that the considered algorithm (partially) recovers the desired support and then state some theorems describing when these sufficient conditions are satisfied with high probability if the signals follow our model. To give a more worldly flavour to the theoretical results, we will highlight them with the example of a dictionary composed of the union of the Dirac and DCT bases or short the Dirac-DCT dictionary. More precisely, $\mathbf{\Phi}_{\text{DDCT}}$ 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:

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

with $\Omega_k = 1/\sqrt{2}$ 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 $\overline{\Lambda} := \{1 \le k \le K, k \notin \Lambda\}$. By the triangle inequalities

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

and

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

we get the recovery condition

$$\|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} < \min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda X\|_p - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{\Phi}_\Lambda X\|_{p,\infty}. \tag{4.19}$$

Recovery conditions for p-SOMP.
p-SOMP partially recovers the good support Λ after M steps if the set Λ_M only contains "good" atoms, i.e. 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 guaranteed if for the residual R_M we have $\|\mathbf{\Psi}_\Lambda^\star R_M\|_{p,\infty} > \|\mathbf{\Psi}_{\overline{\Lambda}}^\star R_M\|_{p,\infty}$. 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 $\mathbf{Q}_\emptyset = \mathbf{I}$), and using the triangle inequalities

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

and

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

we get the recovery condition

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_{\Lambda_M} E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_{\Lambda_M} E\|_{p,\infty} < \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty}. \tag{4.20}$$

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

$$\|\mathbf{\Psi}_\Lambda^\star E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star E\|_{p,\infty} < \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_{\Lambda_M} \mathbf{\Phi}_\Lambda X\|_{p,\infty}. \tag{4.21}$$

Remark 4.3.1. 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. However from an approximation theory perspective, E typically represents the error of best approximation of Y using the atoms in Λ, i.e. $E = Y - \mathbf{\Phi}_\Lambda X$ 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 (4.19) and (4.21) 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.
In the following theorems we formulate conditions that untangle the role of the different objects we are manipulating and show when the two algorithms will succeed with high probability.

Theorem 4.3.1 (Average case analysis for 1-Thresholding)**.** *Let $p = 1$ and $S = |\Lambda|$. Assume*

that $Y = \mathbf{\Phi}_\Lambda \, D^{\frac{1}{2}} \, U + E$ *with* U *a* $S \times N$ *matrix of standard Gaussian random variables and* $D = \operatorname{diag}(\alpha_i^2)_{i\in\Lambda}$*, and suppose that*

$$\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} + \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty} < \sqrt{\frac{2}{\pi}} N \cdot \left(\min_{i\in\Lambda} \alpha_i - \max_{i\in\Lambda} \alpha_i \cdot \mu_2(S) \right). \tag{4.22}$$

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*

$$\gamma := \frac{\min_{i\in\Lambda} \alpha_i - \max_{i\in\Lambda} \alpha_i \cdot \mu_2(S) - \sqrt{\frac{\pi}{2}} N^{-1} \cdot \left(\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} + \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty} \right)}{\min_{i\in\Lambda} \alpha_i + \max_{i\in\Lambda} \alpha_i \cdot \mu_2(S)}. \tag{4.23}$$

Similar results hold for $1 < p \leq \infty$ where $\sqrt{\frac{2}{\pi}} N$ is replaced with a constant $C_p(N)$. To allow for the largest possible noise we should maximise the r.h.s of (4.22). First of all in order to be larger than zero for any fixed number of channels N, this implies

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

The most favourable situation is reached when all components of Λ have the same strength, i.e when the ratio on the l.h.s gets close to one. The range of allowed sparsity is then constrained by the 2-Babel function $\mu_2(S) < 1$, meaning we can recover up to roughly $S = \mu^{-2}$ atoms with high probability, which is much higher than predicted by the worst case analysis in [25] predicting the recoverability of only up to $S = \mu^{-1}$ atoms. When the number of channels N grows, condition (4.22) demands that the average noise per channel $N^{-1}(\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} + \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty})$ 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 quite promising the constraint that each component of the support be equally important remains quite a limitation to the algorithm. This motivates turning our attention to p-SOMP in the hope that this more complex algorithm will perform well under relaxed conditions.

Theorem 4.3.2. *Let* $p = 1$*,* $S := |\Lambda|$ *and* $Y = \mathbf{\Phi}_\Lambda \, D^{\frac{1}{2}} \, U + E$ *with* U *a* $S \times N$ *matrix of standard Gaussian random variables,* $D = \operatorname{diag}(\alpha_i^2)_{i\in\Lambda}$*, 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

$$\|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty} < \sqrt{\frac{2}{\pi}} N \kappa \min_{i\in\Lambda} \alpha_i. \tag{4.24}$$

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*

$$\gamma := \frac{\kappa - \left(\sqrt{\frac{2}{\pi}} N \cdot \min_{i\in\Lambda} \alpha_i \right)^{-1} \cdot \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty}}{\kappa}. \tag{4.25}$$

The theorem gives a characterisation of all index sets Λ that can be recovered with high probability. The main requirement embodied by (4.24) is that the approximation error is sufficiently small compared to the correlations of atoms on the support and correlations of the support with the

rest of the dictionary, measured by the 2-Babel function. Essentially we are asking that:

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

If that is the case, and the average approximation error per channel $N^{-1} \cdot \|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty}$ 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 [9], meaning 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 4.3.3 (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*

$$\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} < \sqrt{\frac{2}{\pi}} N \cdot \min_{i \in \Lambda} \alpha_i \cdot (1 - 3\delta_{S+1}). \tag{4.26}$$

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

$$\gamma := 1 - 3\delta_{S+1} - \left(\sqrt{\frac{2}{\pi}} N \cdot \min_{i \in \Lambda} \alpha_i\right)^{-1} \cdot \|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty}. \tag{4.27}$$

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 (4.18) 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 [5].

Note that because of the term 2^S in the probability bound above, which also appears in Theorem 4.3.2, 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, i.e. $\mathbf{\Phi}\mathbf{\Phi}^\star = I$, 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 4.3.4. *Assume $\mathbf{\Phi}$ to be a tight frame. Let $Y = \mathbf{\Phi}_\Lambda D^{\frac{1}{2}} U$ 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*

$$\|\mathbf{\Phi}_\Lambda^\star E\|_{1,\infty} < \sqrt{\frac{2}{\pi}} N \cdot \min_{i \in \Lambda} \alpha_i \cdot (1 - 3\mu_2(S)) \qquad \text{and} \qquad S < d/37. \tag{4.28}$$

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

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

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

Before proceeding to the technical core of this chapter, let us illustrate our findings using the Dirac-DCT dictionary introduced above. Since in that case we have $\mu_q(S) = S^{1/q}\sqrt{2/d}$, for $q = 1, 2$, the worst case intuition or the analysis in [25] 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 coefficients have equal variances asserts that the probability of recovering supports of size $S \approx d$ rapidly approaches one as the number of channels grows. The same theoretical conclusions for 1-SOMP as for 1-Thresholding can be reached by inspecting equation (4.28).

These theoretical findings are also supported by simulations of the performance of 2-thresholding with $\mathbf{\Psi} = \mathbf{\Phi}$ when the dictionary is made of the Dirac and Fourier basis, $\mathbf{\Phi} = (\mathbf{I}_d, \mathcal{F}_d)$, in dimension $d = 1024$, which has coherence $\mu = 1/\sqrt{d}$. For each number of channels N, varying from 1 to 128, and support size, varying from 1 to 1024 in steps of 16, we created 180 signals by choosing a support Λ uniformly at random and independent Gaussian coefficients with variances $\alpha_i = 1$ and calculated the percentage of thresholding being able to recover the full support. The results can be seen in Figure 4.1.

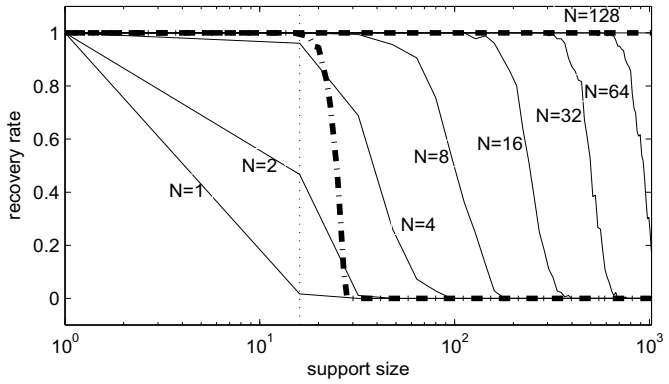

Figure 4.1: Thresholding Recovery Rates for Varying Support Size and Number of Channels.

As reference we also calculated how many out of 200 randomly chosen supports of a given size satisfy the worst case recovery condition $\mu_1(\Lambda) + \sup_{i\in\Lambda} \mu_1(\Lambda\backslash\{i\}) < 1$, derived in [25]. This is indicated by the dash dotted line and can be seen to drop rapidly once the theoretical limit $|\Lambda| = 16$ is reached. Since $\mu = 1/\sqrt{d}$ the average recovery condition $\mu_2(\Lambda) < 1$, indicated by the dashed line, is always satisfied. We can see that as predicted by Theorem 4.3.1 with an increasing number of channels we get closer to the average case bound, which is actually attained once $N = 128$.

Together with the experiment above the average case results confirm 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 and 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. Possibly a different technique similar to the one used in the last chapter when analysing single channel Thresholding could lead to even more beneficial results.

4.4 Average Case Analysis for Thresholding

In this section we will study the average performances of simultaneous p-Thresholding under the multi-channel Gaussian signal model $X = D^{\frac{1}{2}}U$. We first sketch the main arguments so the busy

readers can get enough insight and intuition to go directly to Theorem 4.4.2, which can be simplified to get Theorem 4.3.1, and skip its proof.

4.4.1 Spirit of the Proof

If we want Thresholding to succeed we need to show that

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

The main idea of the proof is based on concentration of measure phenomenon 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

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

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

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

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 necessary machinery to estimate the typicality and precision of the approximation $\|\psi_j^\star \mathbf{\Phi}_\Lambda D^{\frac{1}{2}} U\|_p \approx C_p(N) \cdot \|\psi_j^\star \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_2$ in order to give a fully detailed proof.

4.4.2 Concentration of Measure

As mentioned above the corner stone on which both the average case analysis of Thresholding and of SOMP rely are the following concentration of measure inequalities. Their actual proofs in all gory mathematical detail are awaiting the very motivated reader in the appendix of [25].

Theorem 4.4.1. *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 $0 < \varepsilon_2 < 1$,*

$$\mathbb{P}\Big(\|v_k^\star U\|_p \geq (1+\varepsilon_1) C_p(N) \|v_k\|_2\Big) \leq \exp(-\varepsilon_1^2 A_p(N)) \tag{4.29}$$

$$\mathbb{P}\Big(\|v_k^\star U\|_p \leq (1-\varepsilon_2) C_p(N) \|v_k\|_2\Big) \leq \exp(-\varepsilon_2^2 A_p(N)) \tag{4.30}$$

for each vector v_k, and

$$\mathbb{P}\Big(\max_{k\in\Omega} \|v_k^\star U\|_p \geq (1+\varepsilon_1) C_p(N) \max_{k\in\Omega} \|v_k\|_2\Big) \leq |\Omega| \cdot \exp(-\varepsilon_1^2 A_p(N)) \tag{4.31}$$

$$\mathbb{P}\Big(\max_{k\in\Omega} \|v_k^\star U\|_p \leq (1-\varepsilon_2) C_p(N) \max_{k\in\Omega} \|v_k\|_2\Big) \leq \exp(-\varepsilon_2^2 A_p(N)) \tag{4.32}$$

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

$$\mathbb{P}\Big(\min_{k\in\Omega} \|v_k^\star U\|_p \leq (1-\varepsilon_2) C_p(N) \min_{k\in\Omega} \|v_k\|_2\Big) \leq |\Omega| \cdot \exp(-\varepsilon_2^2 A_p(N)). \tag{4.33}$$

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

Table 4.3: Constants $A_p(N)$ and $C_p(N)$

4.4.3 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

$$\eta := \|\mathbf{\Psi}_{\overline{\Lambda}}^{\star} E\|_{p,\infty} + \|\mathbf{\Psi}_{\Lambda}^{\star} E\|_{p,\infty}, \tag{4.34}$$

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

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

Theorem 4.4.2. *Assume that the noise level η is sufficiently small, i.e.*

$$\eta < C_p(N) \cdot \big(\beta \cdot \alpha_{\min} - \mu_2(\Lambda) \cdot \alpha_{\max}\big). \tag{4.35}$$

Then, under the multichannel Gaussian signal model $X = D^{\frac{1}{2}}U$, 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

$$\gamma := \frac{\beta \cdot \alpha_{\min} - \mu_2(\Lambda) \cdot \alpha_{\max} - \eta/C_p(N)}{\beta \cdot \alpha_{\min} + \mu_2(\Lambda) \cdot \alpha_{\max}} \tag{4.36}$$

Proof: We can bound the probability that Thresholding fails with the following trick,

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

Motivated by the concentration of measure results we set

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

where we choose ε_1 later. Using (4.33) we can bound the first probability in the above as:

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

To bound the second probability we have to work a little bit more before applying (4.31).

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

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 $\varepsilon_2 = \varepsilon_1$,

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

Solving the equation above for ε_1 we get

$$\varepsilon_1 := \frac{\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_2 - \max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_2 - \eta/C_p(N)}{\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_2 + \max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_2}. \tag{4.37}$$

To see that $\varepsilon_1 > 0$ observe that

$$\begin{aligned}
\min_{i\in\Lambda} \|\psi_i^\star \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_2^2 &= \min_{i\in\Lambda} \sum_{k\in\Lambda} |\alpha_k|^2 |\langle \varphi_k, \psi_i\rangle|^2 \geq \alpha_{\min}^2 \cdot \min_{i\in\Lambda} (|\langle \psi_i, \varphi_i\rangle|^2 + \|\mathbf{\Phi}_{\Lambda/i}^\star \psi_i\|_2^2) \geq \alpha_{\min}^2 \cdot \beta^2 \\
\max_{\ell\in\overline{\Lambda}} \|\psi_\ell^\star \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_2^2 &= \max_{\ell\in\overline{\Lambda}} \sum_{k\in\Lambda} |\alpha_k|^2 |\langle \varphi_k, \psi_\ell\rangle|^2 \leq \alpha_{\max}^2 \cdot \max_{\ell\in\overline{\Lambda}} \sum_{k\in\Lambda} |\alpha_k|^2 |\langle \varphi_k, \psi_\ell\rangle|^2 \leq \alpha_{\max}^2 \cdot \mu_2^2(\Lambda).
\end{aligned}$$

Thus we can estimate ε_1 from below as,

$$\varepsilon_1 > \frac{\beta\cdot \alpha_{\min} - \mu_2(\Lambda)\cdot \alpha_{\max} - \eta/C_p(N)}{\beta\cdot \alpha_{\min} + \mu_2(\Lambda)\cdot \alpha_{\max}} =: \gamma. \tag{4.38}$$

This is larger than zero by condition (4.35) and we get as final bound for the probability that Thresholding fails,

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

□

To get from the above theorem to Featured Theorem 4.3.1 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.

4.5 Average Case Analysis of SOMP

In the previous section we have seen that Thresholding requires balanced coefficient variances in order to ensure viable recovery results. This is quite a strong limitation. Motivated by the fact that in the single channel case OMP enables us to overcome this restriction we will now analyse the average performance of SOMP.

4.5.1 Spirit of the Proof

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

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda D^{\frac{1}{2}} U\|_{p,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda D^{\frac{1}{2}} U\|_{p,\infty} > \|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J E\|_{p,\infty}. \tag{4.39}$$

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 (4.39) should be satisfied with high probability as long as

$$\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_{2,\infty} - \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_{2,\infty} > \frac{\|\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J E\|_{p,\infty} + \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J E\|_{p,\infty}}{C_p(N)}. \tag{4.40}$$

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 (4.40) can be estimated from below as

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

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

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

The combination of these two bounds leads to

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

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

$$c_0(\Lambda) = \inf_{J\subsetneq\Lambda} \inf_{i\in\Lambda\setminus J} |\langle \mathbf{Q}_J \varphi_i, \psi_i\rangle|, \quad \text{and} \quad d_0(\Lambda) = \sup_{J\subsetneq\Lambda} \|\mathbf{\Psi}_{\overline{\Lambda}}^\star \mathbf{Q}_J \mathbf{\Phi}_{\Lambda\setminus J}\|_{2,\infty} \tag{4.41}$$

we can finally lower bound the left hand side in (4.40) as

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

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

4.5.2 A General Recovery Result

Theorem 4.5.1. *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.*

$$\eta < \big(c_0(\Lambda) - d_0(\Lambda)\big) \cdot C_p(N) \cdot \alpha^{(M)}. \tag{4.42}$$

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

$$\mathbb{P}(\textit{p-OMP fails after at most } M \textit{ steps}) \leq (1 + |\overline{\Lambda}|) \cdot \mathcal{C}_M \cdot \exp\big(-A_p(N) \cdot \gamma_M^2\big) \tag{4.43}$$

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

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

Proof: We have to show that for any subset J of size at most $M-1$ equation (4.39) 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

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

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.

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

We choose $C = (1-\varepsilon_1) \cdot C_p(N) \cdot |\mathbf{\Psi}_\Lambda^\star \mathbf{Q}_J \mathbf{\Phi}_\Lambda D^{\frac{1}{2}}\|_{2,\infty}$ and use concentration inequality (4.32) to bound the first probability as

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

To bound the second probability we proceed as for Thresholding and use inequality (4.31),

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

Again we require $\epsilon_1 = \epsilon_2$,

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

Solving the above for ε_1 we get

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

If we now insert the definition of $c_0(\Lambda), d_0(\Lambda)$ from (4.41) we can estimate ε_1 from below as:

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

Condition (4.42) 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

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

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 (4.43). □

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 [5]. A future goal therefore would be 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 $c \leq c_0(\Lambda)$ and $d \geq d_0(\Lambda)$. 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. The proofs of the featured theorems of Section 4.3 in the next subsections will rely on such alternatives bounds $c_0(\Lambda), d_0(\Lambda)$.

4.5.3 Proof of Theorem 4.3.2

All we need to do is replace $c_0(\Lambda), d_0(\Lambda)$ in Theorem 4.5.1 by the bounds derived in the following lemma, whose proof can be found in [25].

Lemma 4.5.2. *Valid bounds for the constants $c_0(\Lambda), d_0(\Lambda)$ are given by*

$$c(\Lambda) := \beta - \frac{\mu_2^{in}(\Lambda)}{\sqrt{1-\delta_\Lambda}}, \quad \textit{and} \quad d(\Lambda) := \frac{\mu_2(\Lambda)}{1-\delta_\Lambda}. \tag{4.44}$$

However to make the formulas less ugly we further estimate

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

To finally arrive at Theorem 4.3.2 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 $\eta = \|\mathbf{\Phi}_{\overline{\Lambda}}^\star E\|_{1,\infty}$.

4.5.4 Proof of Theorem 4.3.3

Again the only missing ingredient we need for this proof is a lemma, providing further bounds for the constants $c_0(\Lambda), d_0(\Lambda)$ to be used instead in Theorem 4.5.1.

Lemma 4.5.3. *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*

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

The proof can be found in [25]. To finally prove the theorem we replace $c_0(\Lambda), d_0(\Lambda)$ by c_S, d_S in Theorem 4.5.1 and then need the noise level η to satisfy

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

The above condition is ensured by $\eta < \sqrt{\frac{2}{\pi}} N \cdot \alpha_{\min} \cdot (1 - 3\delta_{S+1})$ 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 4.5.1 the probability of failure is smaller than

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

Inserting the explicit values for c_S, d_S and $\delta_{S+1} < 1/3$ we get from a lengthy but uninteresting calculation that $\gamma_S > 1 - 3\delta_{S+1} - \eta \cdot (\frac{N}{\pi} \cdot \alpha_{\min})^{-1} = \gamma$. 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 4.3.3.

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

4.5.5 Proof of Theorem 4.3.4

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 [56, Theorem B] rewritten.

Theorem 4.5.4. *Suppose* Λ *is selected uniformly at random among all subsets of* $\{1, \ldots, K\}$ *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:

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

To estimate the first term on the right hand side we can proceed as before. Because of Lemma 4.5.2 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

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

which is again ensured by $\delta_\Lambda < 1/3$ and $\eta < \sqrt{\frac{2}{\pi}} N \cdot \alpha_{\min} \cdot (1 - 3\mu_2(S))$. 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 $\gamma_S > 0.9(1 - 3\mu_2(S) - \eta \cdot (\frac{N}{\pi} \cdot \alpha_{\min})^{-1}) = \gamma$ and we get the bound,

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

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 4.5.4 is satisfied and

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

4.6 Discussion

We have seen that in the multi channel case not only the average behaviour of Thresholding but also that of OMP are much better than could be expected from the worst case analysis in [25]. Nevertheless, our results are far from being the final answer. While for Thresholding we have already seen in the last chapter that the average behaviour is also good in the single channel case, we are not aware of comparable results for OMP. Indeed a similar average case analysis in the single channel case would be a major breakthrough. The hitch in our theorems on p-SOMP is the factor resulting from the pachydermal union bounds in the proofs which in consequence necessitates many channels to reach practical success probabilities. Solving this issue with finer arguments would lead to further bridging the gap between theory and practice.

Part II

Design

Compressed Sensing and Redundant Dictionaries

5

In this chapter we extend the concept of *compressed sensing* to signals that are not sparse in an orthonormal basis but rather in a redundant dictionary. We show that a matrix, which is a composition of a random matrix of a certain type and a deterministic dictionary, has small restricted isometry constants. Thus, signals that are sparse with respect to the dictionary can be recovered via Basis Pursuit from a small number of random measurements. Further, we investigate Thresholding as recovery algorithm for compressed sensing and provide conditions that guarantee reconstruction with high probability. The different schemes are compared by numerical experiments. Most of the material presented in this chapter has been published in [46].

5.1 Compressed Sensing

Recently there has been a growing interest in recovering sparse signals from their projection onto a small number of random vectors [6, 8, 9, 14, 24, 39, 40, 44, 48]. The word most often used in this context is *compressed sensing*. It originates from the idea that it is not necessary to invest a lot of power into observing the entries of a sparse signal in all coordinates when most of them are zero anyway. Rather it should be possible to collect only a small number of measurements that still allow for reconstruction. This is potentially useful in applications where one cannot afford to collect or transmit a lot of measurements but has rich resources at the decoder.

Until now the theory of compressed sensing has only been developed for classes of signals that have a very sparse representation in an orthonormal basis (ONB). This is a rather stringent restriction. Indeed as we have seen in the last three chapters, allowing the signal to be sparse with respect to a redundant dictionary adds a lot of flexibility and significantly extends the range of applicability. Already the use of two ONBs instead of just one dramatically increases the class of signals that can be modelled in this way. A more practical example would be a dictionary made up of damped sinusoids which is used for NMR spectroscopy, see [18], a dictionary of translated pulses as used in [38] or a dictionary produced from a localisation grid used for target localization in sensor networks in [11].

There are two main questions in compressed sensing which are of course not independent. How many and what kind of measurements should we take and how can we (stably) reconstruct the signal? Since the measurements are supposed to be very simple they are modelled as an inner product of the sparse signal $x \in \mathbb{R}^d$ with a sampling vector in $\mathbb{R}^d$. Taking n of these linear non-

adaptive measurements, which are stored in the n-dimensional measurement vector s, can then be simply written as multiplying the signal with the $n \times d$ matrix $\mathbf{\Psi}$ which has all the sampling vectors as its rows, i.e. $s = \mathbf{\Psi}x$. To reconstruct the sparse signal from the measurements we have to solve the problem:

$$\text{find a sparse vector } x \text{ satisfying } s = \mathbf{\Psi}x \tag{5.1}$$

Anybody having read the last three chapters should now see that the reconstruction problem is essentially equivalent to finding a sparse representation or, if we assume that the samples are contaminated with noise, a sparse approximation of the samples s in the dictionary $\mathbf{\Psi}$. Thus we can use all the techniques we have seen so far, like combinatorial brute force, greedy algorithms, or BP, but with the additional advantage that the measurement matrix or dictionary is not predefined but can be designed to ensure that the chosen algorithm will succeed.

Candès, Romberg and Tao [8, 9] observed that successful recovery by BP is guaranteed whenever $\mathbf{\Psi}$ has small global restricted isometry constants, meaning it obeys a uniform uncertainty principle, compare Subsection 4.2.2. Based on this concept, Candès, Romberg and Tao proved the following recovery theorem for BP in [8, Theorem 1].

Theorem 5.1.1. *Assume that $\mathbf{\Psi}$ satisfies*

$$\delta_{3S}(\mathbf{\Psi}) + 3\delta_{4S}(\mathbf{\Psi}) < 2$$

for some $S \in \mathbb{N}$. Let x be an S-sparse vector and assume we are given noisy data $y = \mathbf{\Psi}x + e$ with $\|e\|_2 \leq \varepsilon$. Then the solution $x^\#$ calculated via BP, i.e. the solution to the problem $(P1)$ in Table 3.1 satisfies

$$\|x^\# - x\|_2 \leq C\varepsilon. \tag{5.2}$$

The constant C depends only on δ_{3S} and δ_{4S}. If $\delta_{4S} \leq 1/3$ then $C \leq 15.41$.

In particular, if no noise is present, i.e., $\varepsilon = 0$, then under the stated condition BP recovers x exactly. Note that a slight variation of the above theorem holds also in the case that x is not sparse in a strict sense, but can be well-approximated by an S-sparse vector [8, Theorem 2]. The discovery of the restricted isometry constants has triggered a huge interest in compressed sensing and by now there are proofs that several other simpler techniques like the Matching Pursuit variants Regularised Orthogonal MP (ROMP), [40], and Compressed Sensing MP (CoSaMP), [39], or Iterative Hard Thresholding, [6], also guarantee stable recovery if the measurement matrix/dictionary satisfies a uniform uncertainty principle.

However all this theory would be quite useless unless we could actually find measuring matrices having low restricted isometry constants. So what makes the above theorem useful is the fact that for instance an $n \times d$ random matrix with entries drawn from a standard Gaussian distribution (or some other distribution showing certain concentration properties, see below) will have small restricted isometry constants δ_S with overwhelming probability as long as

$$n = \mathcal{O}(S \log(d/S)), \tag{5.3}$$

see [4, 8, 9, 48] for details. 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 [9, 45, 48]. We note that, even though there are deterministically constructed matrices that together with other reconstruction techniques work well for compressed sensing, [28], so far no deterministic construction of measurement matrices obeying the

uniform uncertainty principle for reasonably small n, i.e. comparable to (5.3) is known.

Here we want to address the question whether the techniques described above can be extended to signals y that are not sparse in an ONB but rather in a redundant dictionary $\mathbf{\Phi} \in \mathbb{R}^{d\times K}$ with $K > d$. So now $y = \mathbf{\Phi}x$, where x has only few non-zero components. Again the goal is to reconstruct y from few measurements. More formally, given a suitable measurement matrix $A \in \mathbb{R}^{n\times d}$ we want to recover y from $s = Ay = A\mathbf{\Phi}x$. The key idea then is to use the sparse representation in $\mathbf{\Phi}$ to drive the reconstruction procedure, i.e. try to identify the sparse coefficient sequence x and from that reconstruct y. Clearly, we may represent $s = \mathbf{\Psi}x$ with

$$\mathbf{\Psi} = A\mathbf{\Phi} \in \mathbb{R}^{n\times K}.$$

In particular, we can apply all of the reconstruction methods described above by using this particular matrix $\mathbf{\Psi}$. Of course, the remaining question is whether for a fixed dictionary $\mathbf{\Phi} \in \mathbb{R}^{d\times K}$ one can find a suitable matrix $A \in \mathbb{R}^{n\times d}$ such that the composed matrix $\mathbf{\Psi} = A\mathbf{\Phi}$ allows for reconstruction of vectors having only a small number of non-zero entries. Again the strategy is to choose a random matrix A, for instance with independent standard Gaussian entries, and investigate under which conditions on $\mathbf{\Phi}$, n and S recovery is successful with high probability.

Note that already Donoho considered extensions from orthonormal bases to (redundant) tight frames $\mathbf{\Phi}$ in [14]. There it is assumed that the analysis coefficients $x' = \mathbf{\Phi}^\star y = \mathbf{\Phi}^\star\mathbf{\Phi}x$ are sparse. For redundant frames, however, this assumption does not seem very realistic as even for sparse vectors x the coefficient vector $x' = \mathbf{\Phi}^\star\mathbf{\Phi}x$ is usually fully populated.

Another motivation for investigating the applicability of Compressed Sensing for signals sparse in a dictionary is computational efficiency. If we compare the original problem of finding x from y to the new one of finding x from s we see that instead of the $d \times K$ matrix $\mathbf{\Phi}$ we now have the much smaller $n \times K$ matrix $\mathbf{\Psi}$. Considering that Matching Pursuits and Thresholding, as well as iterative solvers for BP, rely on inner products between the signal and the dictionary elements, we can thus reduce the number of flops per iteration from $\mathcal{O}(dK)$ to $\mathcal{O}(nK)$, where typically $n = \mathcal{O}(S\log(K/S))$, cf. Corollary 5.2.4. Of course this does not make sense when the dictionary has a special structure that allows for fast computation of inner products, e.g. a Gabor dictionary, as the random projections will destroy this structure. However, it has great potential when using for instance a learned and thus unstructured dictionary, cp. [3].

In the following section we will investigate under which conditions on the deterministic dictionary $\mathbf{\Phi}$ its combination with a random measurement matrix will have small isometry constants. By Theorem 5.1.1 this determines how many measurements n will be typically required for BP to succeed in reconstructing all signals of sparsity S with respect to the given dictionary and the interested reader can formulate analogue results for the algorithms in [6, 39, 40]. In Section 5.3 we will analyse the performance of Thresholding, which actually has not yet been considered as a reconstruction algorithm in compressed sensing because of its simplicity and hence resulting limitations. The last section is dedicated to numerical simulations showing the performance of compressed sensing for dictionaries in practice and comparing it to the situation where sparsity is induced by an ONB. So far we are not aware of a proof guaranteeing the success of OMP, however, since it tends to outperform ROMP in practice and unlike CoSaMP is already familiar to all readers we will include it as representative for the Matching Pursuits in the simulations.

5.2 Isometry Constants for $A\Phi$

In order to determine the isometry constants for a matrix of the type $\mathbf{\Psi} = A\mathbf{\Phi}$, where A is an $n \times d$ measurement matrix and $\mathbf{\Phi}$ is a $d \times K$ dictionary, we will follow the approach taken in [4], which was inspired by proofs for the Johnson-Lindenstrauss lemma [1]. We will not discuss this connection further but use as starting point concentration of measure for random variables. This describes the phenomenon that in high dimensions the probability mass of certain random variables concentrates strongly around their expectation.

In the following we will assume that A is an $n \times d$ random matrix that satisfies

$$\mathbb{P}\left(\left|\|Av\|^2 - \|v\|^2\right| \geq \varepsilon\|v\|^2\right) \leq 2e^{-c\frac{n}{2}\varepsilon^2}, \qquad \varepsilon \in (0, 1/3) \tag{5.4}$$

for all $v \in \mathbb{R}^d$ and some constant $c > 0$. Let us list some examples of random matrices that satisfy the above condition.

- **Gaussian ensemble:** If the entries of A are independent normal variables with mean zero and variance n^{-1} then

$$\mathbb{P}(\left|\|Av\|^2 - \|v\|^2\right| \geq \varepsilon\|v\|^2) \leq 2e^{-\frac{n}{2}(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3})}, \qquad \varepsilon \in (0, 1), \tag{5.5}$$

  see e.g. [1, 4]. In particular, (5.4) holds with $c = 1/2 - 1/9 = 7/18$.

- **Bernoulli ensemble:** Choose the entries of A as independent realisations of $\pm 1/\sqrt{n}$ random variables. Then again (5.5) is valid, see [1, 4]. In particular (5.4) holds with $c = 7/18$.

- **Isotropic subgaussian ensembles:** In generalisation of the two examples above, we can choose the rows of A as $\frac{1}{\sqrt{n}}$-scaled independent copies of a random vector $Y \in \mathbb{R}^d$ that satisfies $\mathbb{E}|\langle Y, v\rangle|^2 = \|v\|^2$ for all $v \in \mathbb{R}^d$ and has subgaussian tail behaviour. See [37, eq. (3.2)] for details.

- **Basis transformation:** If we take any valid random matrix A and a (deterministic) orthogonal $d \times d$ matrix U then it is easy to see that also AU satisfies the concentration inequality (5.4). In particular, this applies to the Bernoulli ensemble although in general AU and A have different probability distributions.

Using the concentration inequality (5.4) we can now investigate the local and subsequently the global restricted isometry constants of the $n \times K$ matrix $A\mathbf{\Phi}$.

Lemma 5.2.1. *Let A be a random matrix of size $n \times d$ drawn from a distribution that satisfies the concentration inequality* (5.4)*. Extract from the $d \times K$ dictionary $\mathbf{\Phi}$ any sub-dictionary $\mathbf{\Phi}_\Lambda$ of size S, i.e. $|\Lambda| = S$ with (local) isometry constant $\delta_\Lambda = \delta_\Lambda(\mathbf{\Phi})$. For $0 < \delta < 1$ we set*

$$\nu := \delta_\Lambda + \delta + \delta_\Lambda\delta. \tag{5.6}$$

Then

$$(1 - \nu)\|x\|^2 \leq \|A\mathbf{\Phi}_\Lambda x\|^2 \leq \|x\|^2(1 + \nu) \tag{5.7}$$

with probability exceeding

$$1 - 2\left(1 + \frac{12}{\delta}\right)^S e^{-\frac{c}{9}\delta^2 n}. \tag{5.8}$$

Proof: First we choose a finite ε_1-covering of the unit sphere in $\mathbb{R}^S$, i.e. a set of points Q, with $\|q\| = 1$ for all $q \in Q$, such that for all $\|x\| = 1$

$$\min_{q \in Q} \|x - q\| \leq \varepsilon_1$$

for some $\varepsilon_1 \in (0, 1)$. According to Lemma 2.2 in [37] there exists such a Q with $|Q| \leq (1 + 2/\varepsilon_1)^S$. Applying the measure concentration in (5.4) with $\varepsilon_2 < 1/3$ to all the points $\mathbf{\Phi}_\Lambda q$ and taking the union bound we get

$$(1 - \varepsilon_2)\|\mathbf{\Phi}_\Lambda q\|^2 \leq \|A\mathbf{\Phi}_\Lambda q\|^2 \leq (1 + \varepsilon_2)\|\mathbf{\Phi}_\Lambda q\|^2$$

for all $q \in Q$ with probability larger than

$$1 - 2\left(1 + \frac{2}{\varepsilon_1}\right)^S e^{-cn\varepsilon_2^2}.$$

Define ν as the smallest number such that

$$\|A\mathbf{\Phi}_\Lambda x\|^2 \leq (1 + \nu)\|x\|^2, \tag{5.9}$$

for all x supported on Λ.
Now we estimate ν in terms of $\varepsilon_1, \varepsilon_2$. We know that for all x with $\|x\| = 1$ we can choose a q such that $\|x - q\| \leq \varepsilon_1$ and get

$$\begin{aligned}
\|A\mathbf{\Phi}_\Lambda x\| &\leq \|A\mathbf{\Phi}_\Lambda q\| + \|A\mathbf{\Phi}_\Lambda (x - q)\| \\
&\leq (1 + \varepsilon_2)^{\frac{1}{2}} \|\mathbf{\Phi}_\Lambda q\| + \|A\mathbf{\Phi}_\Lambda (x - q)\| \\
&\leq (1 + \varepsilon_2)^{\frac{1}{2}} (1 + \delta_\Lambda)^{\frac{1}{2}} + (1 + \nu)^{\frac{1}{2}} \varepsilon_1.
\end{aligned}$$

Since ν is the smallest possible constant for which (5.9) holds it also has to satisfy

$$\sqrt{1 + \nu} \quad \leq \quad \sqrt{1 + \varepsilon_2}\sqrt{1 + \delta_\Lambda} + \varepsilon_1 \sqrt{1 + \nu}.$$

Simplifying the above equation yields

$$(1 + \nu) \leq \frac{1 + \varepsilon_2}{(1 - \varepsilon_1)^2}(1 + \delta_\Lambda).$$

Now we choose $\varepsilon_1 = \delta/6$ and $\varepsilon_2 = \delta/3 < 1/3$. Then

$$\frac{1 + \varepsilon_2}{(1 - \varepsilon_1)^2} = \frac{1 + \delta/3}{(1 - \delta/6)^2} = \frac{1 + \delta/3}{1 - \delta/3 + \delta^2/36} < \frac{1 + \delta/3}{1 - \delta/3} = 1 + \frac{2\delta/3}{1 - \delta/3} < 1 + \delta.$$

Thus,

$$\nu \quad < \quad \delta + \delta_\Lambda (1 + \delta).$$

To get the lower bound we operate in a similar fashion.

$$\|A\mathbf{\Phi}_\Lambda x\| \geq \|A\mathbf{\Phi}_\Lambda q\| - \|A\mathbf{\Phi}_\Lambda (x - q)\| \geq (1 - \varepsilon_2)^{\frac{1}{2}} (1 - \delta_\Lambda)^{\frac{1}{2}} - (1 + \nu)^{\frac{1}{2}} \varepsilon_1.$$

Now square both sides and observe that $\nu < 1$ (otherwise we have nothing to show). Then we finally

arrive at

$$\begin{aligned}\|A\mathbf{\Phi}_\Lambda x\|^2 &\geq \big((1-\varepsilon_2)^{\frac{1}{2}}(1-\delta_\Lambda)^{1/2} - \varepsilon_1\sqrt{2}\big)^2 \\ &\geq \cdots \geq 1-\delta_\Lambda-\varepsilon_2-2\varepsilon_1\sqrt{2} \geq 1-\delta_\Lambda-\delta \geq 1-\nu.\end{aligned}$$

This completes the proof. □

Note that the choice of ε_1 and ε_2 in the previous proof is not the only one possible. While our choice has the advantage of resulting in an appealing form of ν in (5.6), others might actually yield better constants. Based on the previous theorem it is easy to derive an estimation of the global restricted isometry constants of the composed matrix $\mathbf{\Psi} = A\mathbf{\Phi}$.

Theorem 5.2.2. *Let $\mathbf{\Phi} \in \mathbb{R}^{d\times K}$ be a dictionary of size K in $\mathbb{R}^d$ with restricted isometry constant $\delta_S(\mathbf{\Phi})$, $S \in \mathbb{N}$. Let $A \in \mathbb{R}^{n\times d}$ be a random matrix satisfying (5.4) and assume*

$$n \geq C\delta^{-2}\Big(S\log(K/S) + \log(2e(1+12/\delta)) + t\Big) \tag{5.10}$$

for some $\delta \in (0,1)$ and $t > 0$. Then with probability at least $1-e^{-t}$ the composed matrix $\mathbf{\Psi} = A\mathbf{\Phi}$ has restricted isometry constant

$$\delta_S(A\mathbf{\Phi}) \leq \delta_S(\mathbf{\Phi}) + \delta(1+\delta_S(\mathbf{\Phi})). \tag{5.11}$$

The constant satisfies $C \leq 9/c$.

Proof: By Lemma 5.2.1 we can estimate the probability that a sub-dictionary $\mathbf{\Psi}_\Lambda = (A\mathbf{\Phi})_\Lambda = A\mathbf{\Phi}_\Lambda$, $\Lambda \subset \{1,\dots,K\}$ fails to have (local) isometry constants $\delta_\Lambda(\mathbf{\Psi}) \leq \delta_\Lambda(\mathbf{\Phi}) + \delta + \delta_\Lambda(\mathbf{\Phi})\delta$ by

$$\mathbb{P}\big(\delta_\Lambda(\mathbf{\Psi}) > \delta_\Lambda(\mathbf{\Phi}) + \delta + \delta_\Lambda(\mathbf{\Phi})\delta\big) \leq 2\big(1+\frac{12}{\delta}\big)^S e^{-\frac{c}{9}\delta^2 n}.$$

By taking the union bound over all $\binom{K}{S}$ possible sub-dictionaries of size S we can estimate the probability of $\delta_S(\mathbf{\Psi}) = \sup_{\Lambda\subset\{1,\dots,K\},|\Lambda|=S}\delta_\Lambda(\mathbf{\Psi})$ *not* satisfying (5.11) by

$$\mathbb{P}\big(\delta_S(\mathbf{\Psi}) > \delta_S(\mathbf{\Phi}) + \delta(1+\delta_S(\mathbf{\Phi}))\big) \leq 2\binom{K}{S}\left(1+\frac{12}{\delta}\right)^S e^{-\frac{c}{9}\delta^2 n}.$$

Using $\binom{K}{S} \leq (eK/S)^S$ (Stirling's formula) and requiring that the above term is less than e^{-t} shows the claim. □

Note that for fixed δ and t condition (5.10) can be expressed in the more compact form

$$n \geq CS\log(K/S).$$

Moreover, if the dictionary $\mathbf{\Phi}$ is an orthonormal basis then $\delta(\mathbf{\Phi}) = 0$ and we recover essentially the previously known estimates of the isometry constants for a random matrix A, see e.g. [4, Theorem 5.2].

Now that we have established how the isometry constants of a deterministic dictionary $\mathbf{\Phi}$ are affected by multiplication with a random measurement matrix, we could in theory go on and apply the result to compressed sensing of signals that are sparse in $\mathbf{\Phi}$. In practice, though, it is not easy to evaluate $\delta_S(\mathbf{\Phi})$ and so need some more initial information about $\mathbf{\Phi}$ first. The following little lemma gives a very crude estimate of the isometry constants of $\mathbf{\Phi}$ in terms of its coherence μ or

Babel function $\mu_1(k)$, compare Equation (2.1) or Subsection 4.2.2 of the last chapter.

Lemma 5.2.3. *For a dictionary with coherence μ and Babel function $\mu_1(k)$ we can bound the restricted isometry constants by*

$$\delta_S \leq \mu_1(S-1) \leq (S-1)\mu. \tag{5.12}$$

Proof: Essentially this can be derived from the proof of Lemma 2.3 in [55]. □

Combining this Lemma with Theorem 5.2.2 provides the following estimate of the isometry constants of the composed matrix $\mathbf{\Psi} = A\mathbf{\Phi}$.

Corollary 5.2.4. *Let $\mathbf{\Phi} \in \mathbb{R}^{d\times K}$ be a dictionary with coherence μ. Assume that*

$$S - 1 \leq \frac{1}{16}\mu^{-1}. \tag{5.13}$$

Let $A \in \mathbb{R}^{n\times d}$ be a random matrix satisfying (5.4). Assume that

$$n \geq C_1(S\log(K/S) + C_2 + t).$$

Then with probability at least $1 - e^{-t}$ the composed matrix $A\mathbf{\Phi}$ has restricted isometry constant

$$\delta_S(\mathbf{\Psi}) \leq 1/3. \tag{5.14}$$

The constants satisfy $C_1 \leq 138.51\, c^{-1}$ and $C_2 \leq \log(1250/13) + 1 \approx 5.57$. In particular, for the Gaussian and Bernoulli ensemble $C_1 \leq 356.18$.

Proof: By Lemma 5.2.3 the restricted isometry constant of $\mathbf{\Phi}$ satisfies

$$\delta_S(\mathbf{\Phi}) \leq (S-1)\mu \leq 1/16.$$

Hence, choosing $\delta = 13/(3\cdot 17)$ yields

$$\delta(A\mathbf{\Phi}) \leq \delta_S(\mathbf{\Phi}) + \delta(1+\delta_S(\mathbf{\Phi})) \leq \frac{1}{16} + \frac{13}{3\cdot 17}(1+\frac{1}{16}) = 1/3.$$

Plugging this particular choice of δ into Theorem 5.2.2 yields the assertion. □

Of course, the numbers 1/16 and 1/3 in (5.13) and (5.14) were just arbitrarily chosen. Other choices will only result in different constants C_1, C_2. Combining the previous result with Theorem 5.1.1 yields a result on stable recovery by Basis Pursuit of sparse signals in a redundant dictionary. We leave the straightforward task of formulating the precise statement to the interested reader. We just want to point out that this recovery result is uniform in the sense that a single matrix A can ensure recovery of *all* sparse signals.

The constants C_1 and C_2 of Corollary 5.2.4 are certainly not optimal; however, we did not further pursue the task of improving them. In the case of a Gaussian ensemble A and an orthonormal basis $\mathbf{\Phi}$ recovery conditions for BP with quite small constants were obtained in [48] and precise asymptotic results can be found in [17]. One might raise the objection that the condition $S - 1 \leq \frac{1}{16\mu}$ in Corollary 5.2.4 is too weak for practial applications. We have already seen that a lower bound on

the coherence in terms of the dictionary size is

$$\mu > \sqrt{\frac{K-d}{d(K-1)}}$$

and that for reasonable dictionaries we can usually expect the coherence to be of the order $\mu \sim C/\sqrt{d}$. The restriction on the sparsity thus is $S < \sqrt{d}/C$. However, compressed sensing is only useful if indeed the sparsity is rather small compared to the dimension d, so this restriction is actually not severe. Moreover, if it is already impossible to recover the support from complete information on the original signal we cannot expect to do this with even less information.

To illustrate the theorem let us have a look at an example where the dictionary is the union of two ONBs.

Example 5.2.5 (Dirac-DCT)**.** *Assume that our dictionary is the union of the Dirac and the Discrete Cosine Transform bases in $\mathbb{R}^d$ for $d = 2^{2p+1}$. The coherence in this case is $\mu = \sqrt{2/d} = 2^{-p}$ and the number of atoms $K = 2^{2p+2}$. If we assume the sparsity of the signal to be smaller than 2^{p-6} we get the following crude estimate for the number of necessary samples to have $\delta_{4S}(A\mathbf{\Phi}) < 1/3$ as recommended for recovery by BP in Theorem 5.1.1,*

$$n \geq C_1(4S(2p\log 2 - \log S) + C_2 + t)$$

with the constants $C_1 \approx 138.51\, c^{-1}$ and $C_2 \approx 5.57$ from Corollary 5.2.4.

In comparison if the signal is sparse in just the Dirac basis we can estimate the necessary number of samples to have $\delta_{4S}(A) < 1/3$ with Theorem 5.2.2 as

$$n \geq C_1'(4S(2p\log 2 - \log 2S) + C_2' + t)$$

with $C_1' = \left(\frac{13}{17}\right)^2 C_1$ and $C_2' \approx 5.3$, thus implying an improvement of roughly the factor $(\frac{17}{13})^2 \approx 1.71$.

5.3 Recovery by Thresholding

In this section we investigate recovery from random measurements by Thresholding. Since Thresholding works by comparing inner products of the signal with the atoms an essential ingredient will be stability of inner products under multiplication with a random matrix A, i.e.

$$\langle Ax, Ay\rangle \approx \langle x, y\rangle.$$

The exact result that we will use is summarised in the following lemma.

Lemma 5.3.1. *Let $x, y \in \mathbb{R}^d$ with $\|x\|_2, \|y\|_2 \leq 1$. Assume that A is an $n \times d$ random matrix with independent $\mathcal{N}(0, n^{-1})$ entries (independent of x, y). Then for all $t > 0$*

$$\mathbb{P}\big(|\langle Ax, Ay\rangle - \langle x, y\rangle| \geq t\big) \leq 2\exp\left(-n\frac{t^2}{C_1 + C_2 t}\right), \tag{5.15}$$

with $C_1 = \frac{8e}{\sqrt{6\pi}} \approx 5.0088$ and $C_2 = \sqrt{8}e \approx 7.6885$. The analogue statement holds for a random matrix A with independent $\pm 1/\sqrt{n}$ Bernoulli entries. In this case the constants are $C_1 = \frac{4e}{\sqrt{6\pi}} \approx 2.5044$ and $C_2 = 2e \approx 5.4366$.

Note that taking $x = y$ in the lemma provides the concentration inequality (5.4) for Gaussian and Bernoulli matrices (with non-optimal constants however). The proof of the lemma is rather

technical and can be found in [46]. However armed with it, we can now investigate the stability of recovery via Thresholding.

Theorem 5.3.2. *Let $\mathbf{\Phi}$ be a $d \times K$ dictionary. Assume that the support x of a signal $y = \mathbf{\Phi}x$, normalised to have $\|y\|_2 = 1$, could be recovered by Thresholding with a margin ε, i.e.*

$$\min_{i\in\Lambda} |\langle y, \varphi_i\rangle| > \max_{k\in\overline{\Lambda}} |\langle y, \varphi_k\rangle| + \varepsilon.$$

Let A be an $n \times d$ random matrix satisfying one of the two probability models of the previous lemma. Then with probability exceeding $1 - e^{-t}$ the support and thus the signal can be reconstructed via Thresholding from the n-dimensional measurement vector $s = Ay = A\mathbf{\Phi}x$ as long as

$$n \geq C(\varepsilon)(\log(2K) + t).$$

where $C(\varepsilon) = 4C_1\varepsilon^{-2} + 2C_2\varepsilon^{-1}$ and C_1, C_2 are the constants from Lemma 5.3.1. In particular,

$$C(\varepsilon) \leq C_3\varepsilon^{-2}$$

with $C_3 \leq 4C_1 + 2C_2 \leq 35.42$ for the Gaussian case and $C_3 \leq 20.90$ in the Bernoulli case.

Proof: Thresholding will succeed if we have

$$\min_{i\in\Lambda} |\langle Ay, A\varphi_i\rangle| > \max_{k\in\overline{\Lambda}} |\langle Ay, A\varphi_k\rangle|.$$

So let us estimate the probability that the above inequality does *not* hold,

$$\begin{aligned}
&\mathbb{P}(\min_{i\in\Lambda} |\langle Ay, A\varphi_i\rangle| \leq \max_{k\in\overline{\Lambda}} |\langle Ay, A\varphi_k\rangle|) \\
&\quad \leq \mathbb{P}(\min_{i\in\Lambda} |\langle Ay, A\varphi_i\rangle| \leq \min_{i\in\Lambda} |\langle y, \varphi_i\rangle| - \frac{\varepsilon}{2}) + \mathbb{P}(\max_{k\in\overline{\Lambda}} |\langle Ay, A\varphi_k\rangle| \geq \max_{k\in\overline{\Lambda}} |\langle y, \varphi_k\rangle| + \frac{\varepsilon}{2})
\end{aligned}$$

The probability of the good components having responses lower than the threshold can be further estimated as

$$\begin{aligned}
\mathbb{P}(\min_{i\in\Lambda} |\langle Ay, A\varphi_i\rangle| \leq \min_{i\in\Lambda} |\langle y, \varphi_i\rangle| - \frac{\varepsilon}{2}) &\leq \mathbb{P}\left(\bigcup_{i\in\Lambda} \{|\langle Ay, A\varphi_i\rangle| \leq |\langle y, \varphi_i\rangle| - \frac{\varepsilon}{2}\}\right) \\
&\leq \sum_{i\in\Lambda} \mathbb{P}\left(|\langle y, \varphi_i\rangle - \langle Ay, A\varphi_i\rangle| \geq \frac{\varepsilon}{2}\right) \\
&\leq 2|\Lambda| \exp\left(-n\frac{\varepsilon^2/4}{C_1 + C_2\varepsilon/2}\right).
\end{aligned}$$

Similarly we can bound the probability of the bad components being higher than the threshold,

$$\begin{aligned}
\mathbb{P}(\max_{k\in\overline{\Lambda}} |\langle Ay, A\varphi_k\rangle| \geq \max_{k\in\overline{\Lambda}} |\langle y, \varphi_k\rangle| + \frac{\varepsilon}{2}) &\leq \mathbb{P}(\bigcup_{k\in\overline{\Lambda}} \{|\langle Ay, A\varphi_k\rangle| \geq |\langle y, \varphi_k\rangle| + \frac{\varepsilon}{2}\}) \\
&\leq \sum_{k\in\overline{\Lambda}} \mathbb{P}(|\langle Ay, A\varphi_k\rangle - \langle y, \varphi_k\rangle| \geq \frac{\varepsilon}{2}) \\
&\leq 2|\overline{\Lambda}| \exp\left(-n\frac{\varepsilon^2/4}{C_1 + C_2\varepsilon/2}\right).
\end{aligned}$$

Combining these two estimates we see that the probability of success for Thresholding is exceeding

$$1 - 2K \exp\left(-n \frac{\varepsilon^2/4}{C_1 + C_2\varepsilon/2}\right).$$

The lemma finally follows from requiring this probability to be higher than $1-e^{-t}$ and solving for n. □

The result above may appear surprising because the number of measurements seems to be independent of the sparsity. The dependence, however, is quite well hidden in the margin ε and the normalisation $\|y\|_2 = 1$. For clarification we will estimate ε given the coefficients and the coherence of the dictionary.

Corollary 5.3.3. *Let $\mathbf{\Phi}$ be an $d \times K$ dictionary with Babel function $\mu_1(k)$. Assume a signal $y = \mathbf{\Phi}_\Lambda x$ with $|\Lambda| = S$ satisfies the sufficient recovery condition for Thresholding,*

$$\frac{|x_{\min}|}{\|x\|_\infty} > \mu_1(S) + \mu_1(S-1), \tag{5.16}$$

where $|x_{\min}| = \min_{i\in\Lambda} |x_i|$. If A is an $n \times d$ random matrix according to one of the probability models in Lemma 5.3.1 then with probability at least $1-e^{-t}$ Thresholding can recover x (and hence y) from $s = Ay = A\mathbf{\Phi}x$ as long as

$$n \geq C_3 S(1+\mu_1(S-1))(\log(2K)+t) \cdot \left(\frac{|x_{\min}|}{\|x\|_\infty} - \mu_1(S) - \mu_1(S-1)\right)^{-2}. \tag{5.17}$$

Here, C_3 is the constant from Theorem 5.3.2. In the special case that the dictionary is an ONB the signal always satisfies the recovery condition and the bound for the necessary number of samples reduces to

$$n > C_3 S \left(\frac{\|x\|_\infty}{|x_{\min}|}\right)^2 (\log(2K)+t). \tag{5.18}$$

Proof: The best possible value for ε in Theorem 5.3.2 is quite obviously

$$\begin{aligned}
\varepsilon &= \min_{i\in\Lambda} |\langle y/\|y\|_2, \varphi_i\rangle| - \max_{k\in\overline{\Lambda}} |\langle y/\|y\|_2, \varphi_k\rangle| \\
&= \frac{1}{\|y\|_2}\big(|\min_{i\in\Lambda} \sum_{j\in\Lambda} x_j \langle\varphi_j, \varphi_i\rangle| - \max_{k\in\overline{\Lambda}} |\sum_{j\in\Lambda} x_j\langle\varphi_j,\varphi_k\rangle|\big) \\
&\geq \frac{1}{\|y\|_2}\left(|x_{\min}| - \|x\|_\infty \mu_1(S-1) - \|x\|_\infty \mu_1(S)\right).
\end{aligned}$$

Therefore, we can bound the factor $C(\varepsilon)$ in Theorem 5.3.2 as

$$C(\varepsilon) \leq C_3 \varepsilon^{-2} \leq C_3 \frac{\|y\|_2^2}{\|x\|_\infty^2} \cdot \big(\frac{|x_{\min}|}{\|x\|_\infty} - \mu_1(S) - \mu_1(S-1)\big)^{-2}.$$

To get to the final estimate observe that by Lemma 5.2.3

$$\frac{\|y\|_2^2}{\|x\|_\infty^2} = \frac{\|\mathbf{\Phi}_\Lambda x\|_2^2}{\|x\|_\infty^2} \leq (1+\mu_1(S-1)) \frac{\|x\|_2^2}{\|x\|_\infty^2} \leq (1+\mu_1(S-1))S.$$

The case of an ONB simply follows from $\mu_1(S) = 0$. □

The previous results tell us that as for BP we can choose the number n of samples linear in the

sparsity S. However, for Thresholding successful recovery additionally depends on the ratio of the largest to the smallest coefficient. Also, in contrast to BP the result is no longer uniform, meaning that the stated success probability is only valid for the given signal x. It does not imply that a single matrix A can ensure recovery for all sparse signals. Indeed, in the case of a Gaussian matrix A and an orthonormal basis $\mathbf{\Phi}$ it is known that once A is randomly chosen then with high probability there exists a sparse signal x (depending on A) such that Thresholding fails on x unless the number of samples n is quadratic in the sparsity S, see e.g. [15, Section 7]. This fact seems to generalise to redundant $\mathbf{\Phi}$.

Example 5.3.4 (Dirac-DCT)**.** *Assume again that our dictionary is the union of the Dirac and the Discrete Cosine Transform bases in $\mathbb{R}^d$ for $d = 2^{2p+1}$. The coherence is again $\mu = 2^{-p}$ and the number of atoms $K = 2^{2p+1}$. If we assume the sparsity $S \leq 2^{p-2}$ and balanced coefficients, i.e. $|x_i| = 1$, we get the following crude estimate for the number of necessary samples*

$$n \geq 6C_3\, S(\log(2)(2p+2) + t).$$

If we just allow the use of one of the two ONBs to build the signal, the number of necessary samples reduces to

$$n \geq C_3\, S(\log(2)(2p+1) + t).$$

Again we see that whenever the sparsity $S \lesssim \sqrt{d}$ the results for ONBs and general dictionaries are comparable. At this point it would be nice to have a similar result for OMP. This task seems rather difficult due to stochastic dependency issues and so, unfortunately, we have not been able to do this analysis yet.

5.4 Numerical Simulations

In order to give a quantitative illustration of the results in Theorem 5.2.2 and Theorem 5.3.2 we will run numerical simulations using the dictionary, we already know from the examples, i.e. the combination of the Dirac and the Discrete Cosine Transform bases in $\mathbb{R}^d$, $d = 256$, with coherence $\mu = \sqrt{1/128} \approx 0.0884$, cp. Lemma 5.2.3 for the resulting bound on the isometry constants.

We drew six measurement matrices of size $n \times d$, with n varying between 64 and 224 in steps of 32, by choosing each entry as independent realisation of a centered Gaussian random variable with variance $\sigma^2 = n^{-1}$. Then for every sparsity level S, varying between 4 and 64 in steps of 4, respectively between 2 and 32 in steps of 2 for Thresholding, we constructed 100 signals. The support Λ was chosen uniformly at random among all $\binom{K}{S}$ possible supports of the given sparsity S. For BP and OMP the coefficients $(x_i)_{i\in\Lambda}$ of the corresponding entries were drawn from a normalised standard Gaussian distribution while for Thresholding we chose them of absolute value one with random signs. Then for each of the algorithms we counted how often the correct support could be recovered. For comparison the same setup was repeated replacing the dictionary with the canonical (Dirac) basis. The results are displayed in Figures 5.1, 5.2 and 5.3.

As predicted by the theorems the necessary number of measurements is higher if the sparsity inducing dictionary is not an ONB. If we compare the three recovery schemes we see that Thresholding gives the weakest results as expected. The improvement in performance of BP over OMP is not that significant, which is especially interesting considering that in practice BP is a lot more computationally intensive than OMP. Still, however, the transition from 'failure' to 'success' is sharper for BP than for OMP.

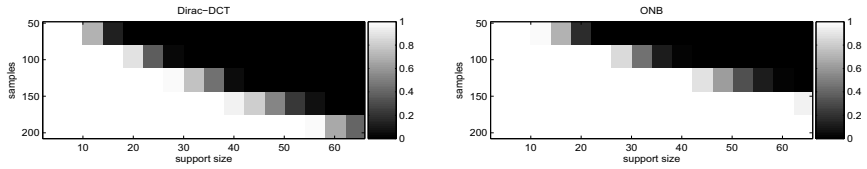

Figure 5.1: Recovery Rates for BP as a Function of the Support and Sample Sizes

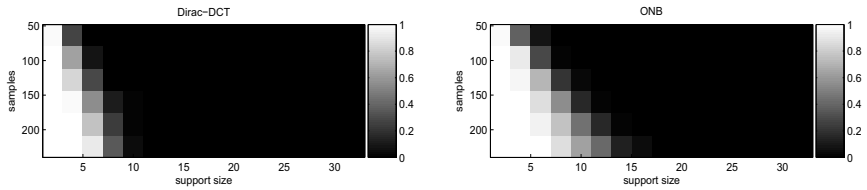

Figure 5.2: Recovery Rates for Thresholding as a Function of the Support and Sample Sizes

5.5 Discussion

We have shown that compressed sensing can also be applied to signals that are sparse in a redundant dictionary. The spirit is that whenever the support can be reconstructed from the signal itself it can also be reconstructed from a small number of random samples with high probability. We have shown that this kind of stability is valid for reconstruction by Basis Pursuit as well as for the simple Thresholding algorithm. Thresholding has the advantage of being much faster and easier to implement than BP. However, it has the slight drawback that the number of required samples depends on the ratio of the largest to the smallest coefficient, and recovery is only guaranteed with high probability for a given signal and not uniformly for all signals in contrast to BP. While we are not aware of a proof guaranteeing the success of OMP if the measurement matrix has small restricted isometry constants our theory that the combination of a deterministic dictionary and a random sensing matrix has well behaved isometry constants can be used to guarantee recovery by the MP variants, ROMP and CoSaMP. In practice however Orthogonal Matching Pursuit seems to indeed work well. In particular, it is still faster than BP and the required number of samples does not seem to depend on the ratio of the largest to the smallest coefficient.
Note that we have a quite strict incoherence assumption on the dictionary, which is a result of asking to be able to reconstruct all signals of a certain sparsity. If on the other hand we just wanted to recover most typical signals we could again expect to get less restrictive conditions based on the 2- instead of the 1-Babel function, e.g. in the case of Basis Pursuit combine Theorem 5.2.2 and Theorem B in [56].
A interesting open question is for which dictionaries it is possible to replace the random Gaus-

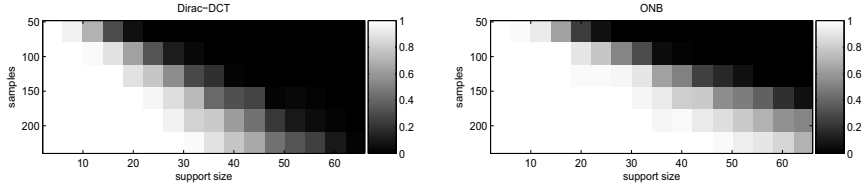

Figure 5.3: Recovery Rates for OMP as a Function of the Support and Sample Sizes

sian/Bernoulli matrix by a random Fourier matrix, see also [44]. This would have the advantage that the Fast Fourier Transform can be used in the algorithms in order to speed up the reconstruction.

6 Classification via Incoherent Subspaces

In this chapter we present a signal model for classification based on a collection of low dimensional subspaces embedded into the high dimensional signal space. We develop an alternate projection algorithm to find such a collection and finally test the classification performance of our scheme in comparison to Fisher's Linear Discriminant Analysis and a recent approach based on sparse approximation.

6.1 Introduction

Let us start with a not so serious example from every day life. The door bell is ringing and we are wondering whether it will be the postman wanting a signature, the plumber coming to fix the toilet or the neighbour complaining about noise, in which case it might be wise not to open the door. What we are facing, while looking through the spy hole and trying to remember what the three candidates look like, is a typical classification problem, ie. given a set of N unit norm training signals $y \in \mathbb{R}^d$ belonging to c classes and a new signal y_{new} find out which class the new signal belongs to. The most common solution approaches follow a two step procedure. First relevant features are selected from the signal. Then the class of the signal is determined by comparing to which features of already labelled signals (nearest neighbour, e.g. [7]) or subspaces spanned by features corresponding to signals in the same class (nearest subspace, cp [32]) the obtained features are closest. In our situation this would mean first focussing on the person's eyes, nose and mouth while ignoring the hairstyle and then comparing them to the eyes, nose and mouth of all possible candidates in previous encounters, in the hope of coming to the right conclusion about opening the door.

In order to formalise both steps we assume the following notation. All signals in class i are collected as columns of the matrix Y_i and these matrices Y_i are in turn combined into a big $d \times N$ data matrix $Y = (Y_1 \dots Y_c) = (y_1^1 \dots y_1^{n_1} \dots y_c^1 \dots y_c^{n_c})$. As this is the simplest and for the chapter most relevant case we will assume that the features are extracted via a linear transform A. This is for instance the case for Fisher's LDA, where A is chosen as the orthogonal projection that maximises the ratio of between-class scatter to that of within-class scatter, [20]. In analogy to the definitions above we define $f_i^k := Ay_i^k$, $F_i := AY_i$ and $F = AY$.* If we denote the Moore-Penrose pseudo-inverse of a

*Note, that in case the features are obtained in a different way the coming results remain valid and interesting when interpreting the features themselves as signals and setting A equal to the identity.

matrix M by $M^\dagger$ and its transpose by $M^\star$ we can summarise the classification procedure as:

Extraction:	$f_{new} = Ay_{new}$	
Labelling:	$\arg\max_i \|F_i^\star f_{new}\|_\infty$	(nearest neighbour)
	$\arg\max_i \|F_i F_i^\dagger f_{new}\|_2$	(nearest subspace)

To see more clearly what happens in the labelling step we expand the expression whose maximum we are seeking for nearest neighbours,

$$\begin{aligned}\|F_i^\star f_{new}\|_\infty &= \|(AY_i)^\star A y_{new}\|_\infty \\ &= \|(\underbrace{A^\star AY_i}_{=:S_i^\infty})^\star y_{new}\|_\infty,\end{aligned}$$

and for nearest subspace. Note that $F_i F_i^\dagger$ as a projection matrix is Hermitian and therefore,

$$\begin{aligned}\|F_i F_i^\dagger f_{new}\|_2 &= \|(F_i F_i^\dagger)^\star f_{new}\|_2 \\ &= \|(AY_i(AY_i)^\dagger)^\star A y_{new}\|_2 \\ &= \|(\underbrace{A^\star AY_i(AY_i)^\dagger}_{=:S_i^2})^\star y_{new}\|_2.\end{aligned}$$

From the two expansion we see that for both classification schemes we can combine the extraction and labelling step using the matrices $S_i^{\infty/2}$:

Extraction&		
Labelling:	$\arg\max_i \|(S_i^\infty)^\star y_{new}\|_\infty$	(nearest neighbour)
	$\arg\max_i \|(S_i^2)^\star y_{new}\|_2$	(nearest subspace)

This formulation should make most mathematical hearts skip a beat and give them the itch to generalise. And indeed first there is no reason why we should restrict ourselves to using sensing matrices of the form $S_i^\infty = A^\star AY_i$ and $S_i^2 = A^\star AY_i(AY_i)^\dagger$ when we could us any $s_i \times d$ matrix S_i, where s_i itself becomes a parameter of choice. Second instead of measuring the two or infinity norm of the sensed vector we could measure any other norm. Classification then gets the general form,

$$\arg\max_i \|S_i^\star y_{new}\|. \tag{6.1}$$

However before trailing off in mathematical bliss let us check how this generalisation could be helpful for the door opening problem. Constructing our sensing matrices through the transform A corresponds to mentally going through the eyes, mouths and noses of the three candidates in previous situations and comparing them to the eyes, mouth and nose of the person in front of the door. Under normal circumstances this approach will work fine but in our not so serious example the problem is that the plumber and the postman are identical twins. So the comparison of features we extracted will give us the same response for the plumber and the postman, even though we can probably distinguish the neighbour. Fortunately - for us - the postman once had an unlucky encounter with the neighbour's dog, which left him with a small scar on the right cheek. This scar sets him apart from his twin, the plumber. The freedom in choice of the sensing matrix, now allows us to remember different features for different people. So for the neighbour we remember eyes,

mouth and nose, for the plumber eyes, nose and scar-free cheek and for the postman eyes, nose and scarred cheek. Going through this list of individual features we see that the person's eyes, mouth and nose do not resemble those of the neighbour in any situation, that his eyes and mouth resemble those of the plumber and postman and that the right cheek seems to be scar-free. Thus we should hurry up and open the door.
In the next section we derive desirable properties of a collection of sensing matrices $S = (S_1 \dots S_c)$ and which norm to choose for the classification procedure through the development and study of a class model based on incoherent subspaces. The third section is dedicated to the development of an algorithm to calculate such a collection and the fourth to test its performance for face recognition. In the last section we summarise our findings, point out connections to related approaches and outline possibilities for future work.

6.2 Class Model

Assume that our favourite mathematical tool, the oracle, has already told us the best norm to use for classification in our data-set. Then a naive way of formulating the problem of finding a good collection of sensing matrices with the help of our training data Y would be: find a collection S that using the prescribed method will correctly classify all our training data and hope that it will work also for all signals to come, ie

$$\text{find } S \text{ s.t } \forall i, j \neq i, k : \frac{\|S_j^\star y_i^k\|}{\|S_i^\star y_i^k\|} < 1. \tag{6.2}$$

While this approach gives us some ideas about how S should look like it still is too general to derive an algorithm. For instance for stable classification the ratio of norms should not be smaller than just one but smaller than a constant $\mu < 1$ and to pick out information of the same order of magnitude and thus prevent mistaking noise for features the sensing matrices for every class S_i should be somehow balanced. To see what these extra constraints for the collection S should be and how the norm should be chosen at the same time, we will develop a class model inspired by the door opening problem.
There we remembered for every person, class, a set of independent features that described the person well. If we model these independent features as orthonormal vectors $f_i^l \in \mathbb{R}^d$ collected in the matrix $F_i = (f_i^1, \dots f_i^s)$, we can write any image of a person i, i.e. signal y_i^k in the class i, as combination of these class specific features and some rest r_i^k, orthogonal to the feature span,

$$y_i^k = F_i x_i^k + r_i^k, \qquad r_i^k \perp sp(F_i). \tag{6.3}$$

For simplicity we assume that the number of independent features per class $s_i = s$ is constant, even though one can imagine situations, where different classes could require different numbers of features for their description. Having defined these features the interesting next step is how to translate that they describe a person/class well. An obvious idea would be to ask for the class specific part of the signal to have higher energy than the rest but, thinking back to the example of face recognition, it is unlikely that the intuitively important features, mouth eyes and scar or nose, contain more energy than the hair and the rest of the face. On the other hand, keeping in mind that we want to do classification by checking which sensing matrix S_i gives the largest response measured in some norm, what we actually want is not that the energy of the class specific part of the signal is larger than the remaining signal part but larger than the energy captured by the features of any other class. Since the set of features of each class F_j forms an orthonormal system, the captured energy

can be easily calculated as $\|F_j^\star y_i^k\|_2$ and what we need is that

$$\forall i,k,j\neq i:\ \frac{\|F_j^\star y_i^k\|_2}{\|F_i^\star y_i^k\|_2}<1. \tag{6.4}$$

Inserting the expression for y_i^k and using the triangular equation we can bound the ratio above as

$$\begin{aligned}\frac{\|F_j^\star y_i^k\|_2}{\|F_i^\star y_i^k\|_2} &\leq \frac{\|F_j^\star F_i x_i^k\|_2+\|F_j^\star r_i^k\|_2}{\|x_i^k\|_2}\\ &\leq \|F_j^\star F_i\|_{2,2}+\frac{\|F_j^\star r_i^k\|_2}{\|x_i^k\|_2},\end{aligned} \tag{6.5}$$

where $\|\cdot\|_{2,2}$ denotes the $(2,2)$ operator norm. For general $1\leq q,p\leq\infty$ the (q,p) operator norm is defined as $\|M\|_{q,p}:=\max_{\|x\|_q=1}\|Mx\|_p$. Assume that both terms of the last bound are small. This means that no combination of features in one class can be well represented by any other set of features and that for every signal the non class specific part of a signal does not have a lot of its energy in the span of features of another class. Then using the feature sets as sensing matrices $S_i=F_i$ and measuring the response in the Euclidean norm will lead to stable classification. Sometimes, as in the introductory example of identical twins, it might however happen that two different classes share one or more features. In this case we have to amend the class model by adding a model on the coefficients of the features for all signals in all classes. One possibility is to assume that all features contribute equally to the class specific part of the signals. Given such a flat distribution of the coefficients $x_i^k(l)$ of all features f_i^l, i.e. their absolute values are constant c, we can exploit the resulting difference of various norms of the coefficient sequence when bounding the ratio we need to be small,

$$\begin{aligned}\frac{\|F_j^\star y_i^k\|_2}{\|F_i^\star y_i^k\|_2} &\leq \frac{\|F_j^\star F_i x_i^k\|_2+\|F_j^\star r_i^k\|_2}{\|x_i^k\|_2}\\ &\leq \|F_j^\star F_i\|_{q,2}\frac{\|x_i^k\|_q}{\|x_i^k\|_2}+\frac{\|F_j^\star r_i^k\|_2}{\|x_i^k\|_2},\end{aligned} \tag{6.6}$$

The norm of flat sequences is smallest for $q=\infty$, leading to $\|x_i^k\|_\infty/\|x_i^k\|_2=s^{-1/2}$ and making this a promising choice for a good further bound. The $(\infty,2)$ norm of $F_j^\star F_i$ can be roughly estimated as

$$\begin{aligned}\|F_j^\star F_i\|_{\infty,2} &= \max_{\|x\|_\infty=1}\|F_j^\star F_i x\|_2\\ &= \max_{\|x\|_\infty=1}\Big(\sum_k\Big(\sum_l\langle f_j^k,f_i^l\rangle x_l\Big)^2\Big)^{1/2}\\ &\leq \Big(\sum_k\Big(\sum_l|\langle f_j^k,f_i^l\rangle|\Big)^2\Big)^{1/2},\end{aligned}$$

and we finally get that

$$\frac{\|F_j^\star y_i^k\|_2}{\|F_i^\star y_i^k\|_2}\leq\left(\frac{\sum_k\big(\sum_l|\langle f_j^k,f_i^l\rangle|\big)^2}{s}\right)^{1/2}+\frac{\|F_j^\star r_i^k\|_2}{\|x_i^k\|_2}. \tag{6.7}$$

The first term of this new bound can be smaller than one even if some of the entries of $F_j^\star F_i$ are as large as 1 provided the rest is small, meaning that in case of balanced coefficients classification will be successful even when two classes share the same feature.

To get the last estimate we exploited the advantageous ratio between the infinity and the Euclidean norm. Pursuing this line of thought, good ratios, and remembering that we can use any (p-) norm for classification, immediately leads to the idea of replacing the Euclidean with the 1-norm, which compares even more favorably to the ∞-norm, and to investigate in general the link between coefficient distributions and (q, p) bounds to characterise the interplay of the feature matrices. Going through the calculations analogue to the ones above we get

$$\frac{\|F_j^\star y_i^k\|_p}{\|F_i^\star y_i^k\|_p} \leq \|F_j^\star F_i\|_{q,p} \frac{\|x_i^k\|_q}{\|x_i^k\|_p} + \frac{\|F_j^\star r_i^k\|_p}{\|x_i^k\|_p}. \tag{6.8}$$

The minimal ratio for balanced coefficients we get for $p = 1$ and $q = \infty$, i.e. $\|x_i^k\|_\infty / \|x_i^k\|_1 = 1/s$. For the $(\infty, 1)$ norm of $F_j^\star F_i$ we have the following crude bound,

$$\begin{aligned} \|F_j^\star F_i\|_{\infty,1} &= \max_{\|x\|_\infty = 1} \|F_j^\star F_i x\|_1 \\ &= \max_{\|x\|_\infty = 1} \sum_k |\sum_l \langle f_j^k, f_i^l \rangle x_l| \\ &\leq \sum_{k,l} |\langle f_j^k, f_i^l \rangle|, \end{aligned} \tag{6.9}$$

which leads us to the following estimate for the ratio of two class responses measured in the 1-norm

$$\frac{\|F_j^\star y_i^k\|_1}{\|F_i^\star y_i^k\|_1} \leq \frac{\sum_{k,l} |\langle f_j^k, f_i^l \rangle|}{s} + \frac{\|F_j^\star r_i^k\|_1}{\|x_i^k\|_1}. \tag{6.10}$$

As before we see that the first term can be smaller than one even if two classes share several features or have quite similar features. The second term can actually be bounded by the analogue term in the Euclidean norm since for a perfectly flat sequence $|x_i| = c$ we have $\|x\|_1 = \sqrt{s}\|x\|_2$ and, in general, $\|x\|_1 \leq \sqrt{s}\|x\|_2$, combining to

$$\frac{\|F_j^\star r_i^k\|_1}{\|x_i^k\|_1} \leq \frac{\|F_j^\star r_i^k\|_2}{\|x_i^k\|_2}. \tag{6.11}$$

So it will be at worst as large as the energy of the non class specific part of a signal in the span of features of another class.

Let's assume now that the coefficients of the class specific features follow the completely opposite distribution. They are not well balanced but extremely sparse, i.e. only one of them is non-zero. In this case the norm of the coefficient sequence is the same for all p, so we cannot profit of a beneficial ratio. However, we can choose p, q to minimise the norm of the interplaying feature matrices. This minimum is attained for $p = \infty$, $q = 1$ and we have $\|F_j^\star F_i\|_{1,\infty} = \max_{k,l} |\langle f_j^k, f_i^l \rangle|$, leading to

$$\frac{\|F_j^\star y_i^k\|_\infty}{\|F_i^\star y_i^k\|_\infty} \leq \max_{k,l} |\langle f_j^k, f_i^l \rangle| + \frac{\|F_j^\star r_i^k\|_\infty}{\|x_i^k\|_\infty}. \tag{6.12}$$

What we can see is, that in case of a sparse coefficient distribution we need the correlation between all feature vectors to be small and the response of the non class specific part to be small. On the other hand it is not a problem if all features of one class can be represented by those of any other. Of course there is ample opportunity to develop more class models, assuming different distributions on the coefficients and using more exotic norms or using different assumptions on the features, i.e. non-orthogonal, but instead of losing ourselves in too much detail we will go on and find a practical

way to calculate sensing or feature matrices for classification based on the three main models.

6.3 Finding Feature/Sensing Matrices

From the analysis in the last section we can derive two types of conditions that the collection of feature or sensing matrices F needs to satisfy. The first type describes how features from different classes should interact, i.e. the interplay measured in the appropriate matrix norm should be small, and the second type how the features should interact with the training data, i.e. the ratio of the response without to within class should be small. The problem with both kinds of conditions is they are not linear and difficult to handle. For instance calculating the $(2,2)$-norm is equivalent to finding the largest singular value and calculating the $(\infty,1)$-norm is even np-hard. We will therefore start with a very simple approach, and in the last section point out how to extend it to include more complicated constraints. So instead of requiring explicitly that the interplay between features from different classes is small, hereby avoiding to investigate what small means quantitatively, we will hope that this will come as free side effect from regulating the interaction with the training data, and simply ask that F is a collection of orthonormal systems of rank s. The condition that the ratio between the response of the training data within to without class is small will be replaced by requiring the response within class to be equal to a constant β_p and without class smaller than a constant μ_p. Define the two sets $\mathcal{F}_s$ and $\mathcal{F}_\mu$ as

$$\begin{aligned}\mathcal{F}_s &:= \{F = (F_1,\dots,F_c) :\ F_i^\star F_i = I_s\}\\ \mathcal{F}_\mu &:= \{F :\ \|F_i^\star y_i^k\|_p = \beta_p,\ \|F_j^\star y_i^k\|_p \le \mu_p,\ \forall k, i, j \neq i\},\end{aligned} \tag{6.13}$$

then our problems could be summarised as finding a matrix in the intersection of the two sets, i.e. $F \in \mathcal{F}_s \cap \mathcal{F}_\mu$. However, since this intersection might be empty, we should rather look for a pair of matrices, each belonging to one set, with minimal distance to each other measured in some matrix norm, eg. the Frobenius norm, denoted by $\|\cdot\|_{\mathbf{2}}$*,

$$\min \|F_s - F_\mu\|_{\mathbf{2}} \text{ s.t. } F_s \in \mathcal{F}_s\, F_\mu \in \mathcal{F}_\mu. \tag{6.14}$$

One line of attack is to use an alternate projection method, i.e. we fix a maximal number of iterations, an initialisation for F_s^0 and then in each iterative step do:

- find a matrix $F_\mu^k \in \arg\min_{F\in\mathcal{F}_\mu} \|F_s^{k-1} - F\|_{\mathbf{2}}$
- check if $\|F_s^{k-1} - F_\mu^k\|_{\mathbf{2}}$ is smaller than the distance of any previous pair and if yes store F_s^{k-1}
- find a matrix $F_s^k \in \arg\min_{F\in\mathcal{F}_s} \|F_\mu^k - F\|_{\mathbf{2}}$
- check if $\|F_s^k - F_\mu^k\|_{\mathbf{2}}$ is smaller than the distance of any previous pair and if yes store F_s^k

If both sets are convex, the outlined algorithm is known as Projection onto Convex Sets (POCS) and guaranteed to converge. Non convexity of possibly both sets, as is the case here, results in much more complex behaviour. Instead of converging, the algorithm just creates a sequence (F_μ^k, F_s^k) with at least one accumulation point. We will not discuss all the possible difficulties here but refer to [57], where all details, proofs and background information can be found and wherein the authors conclude that alternate projection is a valid strategy for solving the posed problem.
To keep the flow of the chapter, we will not discuss the two minimisation problems that need to

*We use this notation instead of the more common variant $\|\cdot\|_F$ to avoid confusion.

be alternatively solved here. The interested reader can find them, including the exact parameter settings in the simulations of the next section, in the appendix of [51]. Instead we will discuss how to set the parameters β_p, μ_p and possible choices for the initialisation F_s^0.
The motivation for our choice of β_p is the best case situation. An orthonormal system of s feature vectors can maximally take out all the energy of a signal,

$$\|F^\star y\|_2 \leq \|y\|_2. \tag{6.15}$$

As the signals are assumed to have unit norm, this energy is at most one and we set $\beta_2 = 1$. The maximal 1-norm of the vector $F_i^\star y$ of length s with energy 1 is $\sqrt{s}$. This is attained when all features of one class take out the same energy, i.e. the absolute values of the entries in $F^\star y$ are all equal to $1/\sqrt{s}$. This leads to $\beta_1 = \sqrt{s}$. The infinity norm $F_i^\star y$ corresponds to the maximal inner product between one of the feature vectors and the signal. As both the feature vector and the signals are normalised, this can be at most one and so we set $\beta_\infty = 1$.
From the discussion in the last section we see that the parameter μ reflects how much the spaces containing the class specific part overlap. If we have $d \geq c \cdot s$, it is theoretically possible to have c subspaces of dimension s which are mutually orthogonal to each other, and μ could be zero. As soon as the above inequality is reversed, because for instance the actual dimension of the span of all features, i.e. $rank(F)$, is smaller than d, not all subspaces corresponding to the different classes can be orthogonal but will have to overlap. How the size of the overlap should be measured, is determined by the choice of p-norm for classification. For instance for $p = 2$ the overlap is measured by $\|F_j^\star F_i\|_{2,2}$ and from theory about Grassmannian manifolds, see [57], we know that the maximal overlap between two of c subspaces of dimension s embedded in the space $\mathbb{R}^d$ can be lower bounded by

$$\max_{i \neq j} \|F_j^\star F_i\|_{2,2}^2 \geq \frac{s \cdot c - d}{d(c-1)}. \tag{6.16}$$

The problem with setting μ as above is that we are not controlling the interaction between the sets of features directly but only indirectly over the training data. There the worst case might not be assumed and so μ as above would be too large. Also for the cases $p = 1, \infty$ we do not have a similar bound. Therefore instead of trying to analyse theoretically how to set μ, where we have to deal with too many unknowns, we use the above bound as an indication of order of magnitude and, when testing our scheme on real data, vary the parameter μ. Lastly for the initialisation for each class we choose the orthogonal system that maximises the energy taken from this class opposed to the energy taken from the other classes, i.e.

$$F_{s,i}^0 = \arg \min_{F_i^\star F_i = I_s} \|F_i^\star Y_i\|_{\mathbf{2}}^2 - \sum_{j \neq i} \|F_i^\star Y_j\|_{\mathbf{2}}^2. \tag{6.17}$$

This problem can be easily solved, by considering the rewritten version of the function to minimise,

$$\min_{F_i^\star F_i = I_s} \operatorname{trace}\left(F_i^\star (Y_i Y_i^\star - \sum_{j \neq i} Y_j Y_j^\star) F_i\right). \tag{6.18}$$

If $UDU^\star$ is an eigenvalue decomposition of the symmetric (Hermitian) matrix $Y_i Y_i^\star - \sum_{j \neq i} Y_j Y_j^\star$, then the minimum is attained for $F_{s,i}^0$ consisting of the s eigenvectors corresponding to the s largest eigenvalues.

$s\backslash\frac{\mu}{\sqrt{s}}$	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
2	60	56	56	57	60	58	60	61	66	64	69
3	52	46	48	46	51	51	53	58	62	61	61
4	62	52	54	55	55	56	56	54	55	57	61
5	64	59	56	56	55	58	61	63	66	68	68
6	61	54	57	54	56	59	62	58	61	71	71
7	57	55	57	55	59	57	58	62	61	68	69

Table 6.1: Number of misclassified images for $p = 1$ and varying values s and μ.

$s\backslash\mu$	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
1	57	58	59	58	60	59	59	58	58	58	62
2	51	49	51	51	51	55	57	57	59	58	56
3	47	42	45	50	53	53	54	61	62	61	64
4	46	42	41	41	47	48	51	62	63	61	63
5	48	43	40	44	50	51	52	55	55	59	61
6	49	45	42	45	49	48	51	54	54	57	58
7	45	43	43	43	45	45	48	53	51	54	52

Table 6.2: Number of misclassified images for $p = 2$ and varying values s and μ.

6.4 Testing

To test the proposed scheme we used a subset of images from the AR-database, [36]. For each of the 126 people there are 26 frontal images of size 165×120 taken in two separate sessions. The images include changes in illumination, facial expression and disguises. For the experiment we selected 50 male and 50 female subjects and for each of them took the 14 images with just variations in illumination and facial expression, neutral, light from the right and left, front light, angry, happy, sleepy. The all together 700 images from the first session were used as training data and the 699 images* from the second session for testing. Every image was converted to grayscale and then stored as a 19800 dimensional column vector. The images from the first session were stored in the 19800×700 matrix Y^1 and those from the second in the 19800×699 matrix Y^2. In order to speed up the calculations, we first applied a unitary transform, which does not change the geometry of the problem, but reduces the size of the matrices, i.e. we did a reduced QR-factorisation decomposing Y^1 into the 19800×700 matrix Q with orthogonal columns and the 700×700 upper triangular matrix R and set $\tilde{Y}^1 = Q^\star Y^1 = R$ and $\tilde{Y}^2 = Q^\star Y^2$.

We tested the proposed scheme for all three choices of p and varying values of μ_p scaling from 0 to 10% of β_p and number of features per class varying from 1 to 7. The choice of the maximal outside-class contribution $\mu_{\max} = 0.1\beta_p$ was inspired by the bound in (6.16). If we take as effective signal dimension $d = 700$ and assume that the space should not only accommodate the 100 different people in our training set but all people, i.e. we let c go to infinity, the bound approaches $\sqrt{s/d}$ which is 0.1 if $s = 7$ and 0.0378 if $s = 1$. The maximal number of features per class is 7, since we only have 7 test images and so it does not make sense to look for spaces of higher dimension containing all test images. Note also that for $s = 1$ the three schemes are the same, so the results are only displayed once. For each set of parameters we calculated the corresponding feature matrix using the algorithm described in the last section on the images from the first session. We then classified the images from the second session using the appropriate p-norm. The results are shown in Tables 6.1, 6.2 and 6.3.

*700 minus corrupted image w-027-14.bmp

$s\backslash\mu$	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
2	55	62	59	54	56	52	54	61	63	64	62
3	55	63	58	56	60	58	59	63	65	69	69
4	55	64	60	57	59	58	58	61	67	70	67
5	55	60	59	55	58	57	57	60	66	71	69
6	55	61	59	54	57	56	56	65	67	72	69
7	55	61	59	55	56	54	55	66	66	71	70

Table 6.3: Number of misclassified images for $p = \infty$ and varying values s and μ.

As we can see we get the best performance for $p = 2$, followed by $p = 1$ and $p = \infty$. This comes as no surprise when considering the structure of our data. Intuitively the important features of a face are eyes, nose and mouth. Since the people in the pictures have different facial expression, usually not all of these features will be active explaining why $p = 1$ is not the most appropriate model. On the other hand we can expect to have more than one feature active at the same time even if not to the same extent. Using $p = \infty$ we lose the information given by these secondary active features while with $p = 2$ we still incorporate it into the final decision.
We can also see that 0.1% of μ as maximally allowed outside class 'energy' seemed to have been a good choice as we can always see a small decrease and large increase of the error going from 0 to 0.1, with the best range for $p = 1$ and $p = 2$ between 0.01 and 0.03 and for $p = \infty$ between 0.02 and 0.06. For $p = 1$ we get better performance for the lower dimensions, which seems reasonable because there the equal energy distribution over the features is easier achieved. For $p = 2$ on the other hand the better performance is achieved with higher dimensions, which are able to capture more important side details. Finally for $p = \infty$ the results seem equal for all dimensions. A possible explanation is given by the initialisation, which ensures that for all dimensions the first, most promising direction is included.
Still in all three cases in the most promising ranges the proposed scheme outperforms a standard method like Fisher's LDA, [20]. The best result by LDA is obtained when using the highest possible number of discriminant axes $c - 1 = 99$. In this case nearest neighbour classification, corresponding to $p = \infty$ but with non orthogonal features, fails to identify 59 images, and nearest subspace classification, corresponding to $p = 2$ fails to identify 71 images. When concentrating on the results for $p = 2$, which is the most sensible choice given the structure of the data, $p = 2$, we also see that the scheme performs well in comparison to a recent, successful method based on ℓ_1 minimisation, [58]. The best result reported there is a success rate of 94.99%, meaning 35 misclassified images, which is 5 images better than our best case of 40 errors. The advantage of our method is that it is a lot simpler. Not taking the calculation of the feature matrices into account, as this part of the pre-processing, all that has to be done to classify a new data vector is to multiply it with the feature matrix, $cs(2d-1)$ operations, calculate the norms for each class, $c(2s-1)$ operations in the computationally worst case $p = 2$ and find the maximum, $c - 1$ operations. Taken all together this results in less than $cs(2d + 1)$ operations, which is basically the cost of the matrix vector product. The ℓ_1 minimisation method however requires on top of extracting d_f features, $d_f(2d-1)$ operations if it is done linearly, the solution of the convex optimisation problem

$$\min \|z\|_1 \text{ s.t. } \|f_{new} - Fz\|_2 \leq \varepsilon, \tag{6.19}$$

where F in this case is the $d_f \times N$ matrix containing the features of all the training data, which contributes significantly to the overall cost, especially if the number of training signals is large.

6.5 Discussion

We have presented a class model based on incoherent subspaces and linked to that a classification scheme. From a more practical viewpoint we have developed an algorithm to calculate these subspaces, i.e. the feature matrices, and shown that the scheme gives promising results on the AR database. The idea that each class should have its own representative system, learned from the training data can already be found in [52]. There frames or dictionaries for texture classification are learned, such that each provides a sparse representation for its texture class. The new texture then gets the label of the texture frame providing the sparsest representation. In [35], the same basic idea is used but the learning is guided by the principle that the dictionaries should also be discriminant, while in [47] both learning principles are combined, i.e. the dictionaries should be discriminant and approximative. This third scheme can be considered as a more general and more complicated version of our approach. Alternatively our approach can be considered to be a hybrid of Nearest Subspace respectively Nearest Neighbour and the discriminative and approximative frame scheme, in so far as it is linear but has individual features for every class.
The idea to use a collection of subspaces for data analysis can also be found in [34], where the subspaces are used to model homogenous subsets of high-dimensional data which together can capture the heterogenous structures.
For the future there remain some interesting directions to explore. Firstly the possibilities of the subspace classification approach do not seem exhausted using the proposed algorithm. Ironically this fact revealed itself through a mistake in the minimisation procedure, resulting in matrix pairs with distances larger than the optimal ones, and sensing matrices giving better classification results, i.e. in the best case an error of only 35 misclassified images. The main difference of these fake optimal matrices to the sensing matrices corresponding to the actual minima, seemed to be that, while capturing approximately the same 'energy' within class, they were more accurate in respecting the without class energy bound, i.e. less overshooting of the maximally allowed value μ. This overshooting for the real minimal is a result of imposing not only $\|F_i y_j^k\|_2 \leq \mu$ but also $\|F_i y_i^k\|_2 = \beta$, which forces the optimal feature matrix to balance the error incurred by not attaining β within class and the error incurred by being larger than μ without class. A promising idea to avoid the overshooting would be to change the problem formulation and ask to maximise the 'energy' within class subject to keeping the 'energy' without class small, i.e. in the case $p = 2$ solve,

$$\max \sum_i \|F_i^\star Y_i\|_{\mathbf{2}}^2 \text{ s.t. } F_i^\star F_i = I_s \text{ and } \|F_i x_j^k\|_2 \leq \mu,\ \forall k, j \neq i. \tag{6.20}$$

Lastly our approach allows to impose additional constraints on F, like incoherence of the subspaces between each other, e.g. $\|F_i^\star F_j\|_{2,2} \leq \nu$ for $p = 2$, or low rank of the whole feature matrix to reduce the cost of calculating $F^\star y_{new}$. Another possibility to reduce computational cost if d and N are very large, especially in the training step, would be to first take random samples of the training data, which reduce their dimension but very likely preserve the geometrical structure, as described in [1] and used in [58]. Alternatively to reduce the dimension of F one can apply our scheme on classical features, like Eigen or Laplace features, instead of directly on the raw training data.

7 Dictionary Identification

At the beginning of Chapter 2 we introduced the two main questions, when dealing with dictionaries and sparsity. The first, how to find a sparse representation for a signal given the dictionary and the second, how to find a dictionary that gives sparse representations for a class of signals. Here we finally turn to this second question.

7.1 Introduction

Sparse signals are useful. They are easy to store and to compute with and as we have seen in Chapter 5 they are also easy to capture. On the other hand, as has as well become apparent in the first few chapters, it is far from easy to find sparse representations/approximations. Solving the original problem $P(0)$, compare Table 3.1, of finding the approximation with the most zero coefficients turned out to be np-hard, thus necessitating the development of alternative strategies. Checking in any of the already cited publications, e.g [16, 21, 55, 56], when popular methods like thresholding, matching pursuits, basis pursuit will succeed (with high probability) you will more likely than not find a statement starting with 'given a dictionary $\mathbf{\Phi}$ and a signal having an S-sparse approximation/representation ...', which points exactly to the remaining problem. If you have a class of signals and you would like to find sparse approximations someone has to give you the right dictionary. For many signal classes good dictionaries like time-frequency or time scale dictionaries are known and from theoretical study of your signal class you might be able to identify one that will fit well. However, if you run into a new class of signals, chances that the best fit will already be known are quite slim and it can be a time consuming overkill to develop a deep theory like that of wavelets every time. An attractive alternative approach is dictionary learning, where you try to infer the dictionary that will give you good sparse representations for your whole signal class from a small portion of training signals.

Considering the extensive literature available for the sparse decomposition problem, surprisingly little work has been dedicated to theoretical dictionary learning so far. There exist several dictionary learning algorithms [3, 19, 29, 30], but only recently people have started to consider also the theoretical aspects of the problem. Dictionary learning finds its roots in the field of Independent Component Analysis (ICA) [10], where many identifiability results are available, which however rely on asymptotic statistical properties under independence assumptions. Georgiev, Theis and Cichocki [23] as well as Aharon and Elad [2] describe more geometric identifiability conditions on the (sparse)

coefficients of training data in an ideal (overcomplete) dictionary. Both approaches to the identifiability problem rely on rather strong sparsity assumptions, and require a huge amount of training samples. In addition to a theoretical study of dictionary identifiability, both cited papers provide algorithms to perform the desired identification. Unfortunately the naive implementation of these provably good dictionary recovery algorithms seems combinatorial, which limits their applicability to low dimensional data analysis problems and renders them fragile to outliers, i.e. training signals without a sparse enough representation. In this chapter we will study the question when a dictionary can be learned via ℓ_1-minimisation [43, 60], and thus by a non-combinatorial algorithm.

7.2 Dictionary Learning via ℓ_1-Minimisation

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

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

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

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

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

$$\min_{\mathbf{\Phi} \in \mathcal{D}} \mathcal{C}_0(\mathbf{\Phi}, Y). \tag{7.1}$$

The problem is that already finding the representation with minimal non-zero coefficients for one signal in a given dictionary is np-hard, which makes trying to solve (7.1) indeed a daunting task. Fortunately the problem above is not only daunting but also rather uninteresting, since it is not stable with respect to noise or suited to handle signals that are only compressible. Thus the idea of learning a dictionary via ℓ_1-minimisation is motivated on the one hand by the goal to have a criterion that is taking into account that the signals might be noisy or only compressible and on the other by the success of the Basis Pursuit principle for finding sparse representation. There the ℓ_0-pseudo norm was replaced with the ℓ_1-norm, compare Table 3.1, which also promotes sparsity but is convex and continuous. The same strategy can be applied to the dictionary learning problem and the ℓ_0 cost function can be replaced with the ℓ_1-cost function

$$\mathcal{C}_1(\mathbf{\Phi}, Y) := \min_{X \mid \mathbf{\Phi} X = Y} \|X\|_1, \tag{7.2}$$

where $\|X\|_1 := \sum_n \|x_n\|_1$. Several authors [41, 42, 60] have proposed to consider the corresponding minimisation problem

$$\min_{\mathbf{\Phi} \in \mathcal{D}} \mathcal{C}_1(\mathbf{\Phi}, Y). \tag{7.3}$$

Unlike for the sparse representation problem, where this change meant a convex relaxation, the dictionary learning problem (7.3) is still *not convex* and cannot be immediately addressed with generic convex programming algorithms. However, it seems better behaved than the original problem (7.1) because of the continuity of the criterion with respect to increasing amounts of noise, which makes it more amenable to numerical implementation.

Looking at the problem above, we see that in order to solve it we still need to define $\mathcal{D}$, the set of admissible dictionaries. Several families of dictionaries can be considered such as discrete libraries of orthonormal bases (wavelet packets or cosine packets, for which fast dictionary selection is possible using tree-based searches [12]). Here we focus on the 'non parametric' learning problem where the full $d \times K$ matrix $\mathbf{\Phi}$ has to be learned. Since the value of the criterion (7.3) can always be decreased by jointly replacing $\mathbf{\Phi}$ and X with $\alpha\mathbf{\Phi}$ and X/α, $0 < \alpha < 1$, a scaling constraint is necessary and a common approach is to only search for the optimum of (7.3) within a bounded domain $\mathcal{D}$. A set of possible scaling conditions is defined through inequality constraints of the form $\sum_k \|\varphi_k\|_2^\tau \leq 1$ with $0 < \tau < \infty$, with the standard replacement $\max_k \|\varphi_k\|_2 \leq 1$ when $\tau = \infty$ *. Since the optimum of (7.3) with any of the considered inequality constraints in indeed achieved when there is equality, we define the following constraint manifolds for $0 < \tau < \infty$

$$\mathcal{D}^\tau := \{\mathbf{\Phi}, \sum_k \|\varphi_k\|_2^\tau = 1\}, \tag{7.4}$$

and for $\tau = \infty$:

$$\mathcal{D}^\infty := \{\mathbf{\Phi}, \forall k, \|\varphi_k\|_2 = 1\}. \tag{7.5}$$

The constraint manifolds $\tau = 2, \infty$ are for instance used in [30, 59]. For simplicity reasons we will concentrate here on the case $\tau = \infty$, i.e. $\mathcal{D} := \mathcal{D}^\infty$, and refer to the forthcoming paper [26] for the general case.

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

7.2.1 The Identifiability Problem

One important task would be to develop efficient algorithms for solving the posed minimisation problem (7.3). This numerical part of dictionary learning is also the most commonly studied one. Indeed several algorithms have been proposed which adopt a similar approach to learning a dictionary [19, 30, 43] from training data, and their empirical behaviour has been explored. Here we are interested in the more theoretical problem of *dictionary identifiability*. Assuming that the data Y were generated from an 'ideal' dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$ and 'ideal' coefficients X_0 as $Y = \mathbf{\Phi}_0 X_0$, we want to determine conditions on X_0 and to a lesser extent on $\mathbf{\Phi}_0$ such that the minimisation of (7.3) recovers $\mathbf{\Phi}_0$.

Our objective is therefore similar in spirit to previous work on dictionary recovery [2, 23] which studied the uniqueness of overcomplete dictionaries for sparse component analysis. The main difference is that we specify in advance which optimisation criterion we want to use to recover the dictionary (ℓ_1-minimisation) and attempt to express conditions on the matrix X_0 to guarantee that this method will successfully recover a given class of dictionaries.

A first difficulty we immediately face when talking about recovery are the ambiguities that have been known at least since the development of Independent Component Analysis. Because of the normalisation constraint on the dictionary, the usual scaling ambiguity is avoided, but there remains a permutation and a sign ambiguity. For any permutation matrix $\mathbf{P}$ and any diagonal matrix $\mathbf{D}$

*Other constraints, which replace the norm $\|\varphi_k\|_2$ with, e.g., the norm $\|\varphi_k\|_1$, would also be interesting to study for the dictionary learning problem when it is desirable to obtain not only sparse coefficients but also sparse atoms.

with unit diagonal entries we have $\mathbf{\Phi}X = (\mathbf{\Phi PD})(\mathbf{DP}X)$. Hence Problem (7.3) has not just one but a whole equivalence class of minimisers, each of them corresponding to a matching column resp. row permutation and sign change of $\mathbf{\Phi}$ resp. X. Therefore, we have to relax our requirement and can only ask to find conditions such that minimising (7.3) recovers $\mathbf{\Phi}_0$ up to permutation and sign change. The notation $\mathbf{\Phi} \sim \mathbf{\Phi}_0$ will indicate this indeterminacy, meaning that $\mathbf{\Phi} = \mathbf{\Phi}_0\mathbf{PD}$ for some permutation matrix $\mathbf{P}$ and diagonal matrix $\mathbf{D}$ with unit diagonal entries.
Ideally, we would like to characterise coefficient matrices X_0 such that, for any $\mathbf{\Phi}_0 \in \mathcal{D}$ or at least for a reasonable subset of $\mathcal{D}$ such as, for instance, 'incoherent' dictionaries, the *global minima* of

$$\min_{\mathbf{\Phi}\in\mathcal{D}} \mathcal{C}_1(\mathbf{\Phi}, \mathbf{\Phi}_0 X_0) \tag{7.6}$$

can only be found at $\mathbf{\Phi} \sim \mathbf{\Phi}_0$. An even more ambitious goal would be to characterise coefficient matrices such that the *local minima* of (7.6) can only be found at $\mathbf{\Phi} \sim \mathbf{\Phi}_0$, which would guarantee that numerical optimisation algorithms cannot be trapped in spurious local minima, and would behave somewhat independently of their initialisation. This objective raises two complementary questions:

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

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

Here we will concentrate on the first question. Unfortunately, in the study of the ℓ_1-minimisation based dictionary recovery problem, several difficulties arise at once, some due to the possible over-completeness and non-orthogonality of the dictionary, others due to the difficulty of globally characterising the optima of a globally nonconvex problem which admits exponentially many solutions because of the permutation and sign indeterminacies. Therefore instead of characterising directly the local minima of the problem (7.6) we consider the related problem

$$\min_{\mathbf{\Phi}\in\mathcal{D}, X|\mathbf{\Phi}X=Y} \|X\|_1. \tag{7.7}$$

After introducing some notations we provide conditions when a pair $(\mathbf{\Phi}_0, X_0)$ is a local minimum of the ℓ_1-norm $\|X\|_1$ over the constraint manifold

$$\mathcal{M}(Y) := \{(\mathbf{\Phi}, X), \mathbf{\Phi} \in \mathcal{D}, \mathbf{\Phi}X = Y\}. \tag{7.8}$$

In Section 7.5 we specialise to the case of the dictionary being a basis to get to a more concrete sufficient local recovery condition, which we illustrate with an easy example in Section 7.6. This sufficient recovery condition is used in Section 7.7 to derive how many training signals with coefficients generated by a random process are typically needed to guarantee that a basis constitutes a local minimum of the ℓ_1-criterion. The last section is dedicated to the discussion of the results obtained and to point out future research directions.

7.3 Notations

To state the main lemmata and express the local identifiability conditions, we will adopt the following notation conventions.

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

For any $K \times N$ matrix A and index set $\Omega \subset \{1,\dots,K\} \times \{1,\dots N\}$, the notation A_Ω will refer ubiquitously either to the vector $(A_{kn})_{(k,n)\in\Omega}$ or or the $K \times N$ matrix which matches A on Ω and is zero elsewhere.

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

$$\langle AB, C\rangle_F = \text{trace}(B^\star A^\star C) = \text{trace}(A^\star C B^\star) = \langle A, CB^\star\rangle_F \tag{7.9}$$

and similar relations such as

$$\langle \text{diag}(A), B\rangle_F = \langle A, \text{diag}(B)\rangle_F. \tag{7.10}$$

Zero-Diagonal and Diagonal Decompositions
We will use the following simple lemma.

Lemma 7.3.1. *Consider two matrices* $\mathbf{A}$,$\mathbf{B}$ *and let* $\mathbf{A} = \mathbf{Z}_1 + \mathbf{\Delta}_1$, $\mathbf{B} = \mathbf{Z}_2 + \mathbf{\Delta}_2$ *be their unique decomposition into a sum of a zero-diagonal and a diagonal matrix. Then*

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

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

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

□

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

$$\mathbf{M}_0 := \mathbf{\Phi}_0^\star\mathbf{\Phi}_0 - I. \tag{7.11}$$

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

7.4 Local Identifiability Conditions

Just as in the representation problem in Table 3.1, where the ℓ_1-cost is not a smooth function of x as soon as x has at least one zero entry, the cost in Equation (7.7) is not a smooth function of $(\mathbf{\Phi}, X)$ whenever X has at least one zero entry. Therefore, one cannot fully characterise the local minima

*We will generally distinguish column vectors from row vectors using subscript *vs* superscript indices.

of the cost function (7.7) as a subset of the zeroes of a 'gradient' of the ℓ_1-cost function with respect to $(\mathbf{\Phi}, X)$, as this gradient does not exist everywhere. Here, on the opposite, we want to understand the effect of the non-smooth behaviour of the cost function and to exploit it to characterise its local minima. For that we will develop a replacement for the 'gradient' which accounts for the fact that the ℓ_1-cost function indeed admits one-sided directional derivatives everywhere.

For the study of local minima $(\mathbf{\Phi}_0, X_0)$ of (7.7) we first need a characterisation of the tangent space $T_{(\mathbf{\Phi}_0,X_0)}\mathcal{M}(Y)$ to the constraint manifold $\mathcal{M}(Y)$ at the point $(\mathbf{\Phi}_0, X_0)$.

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

The tangent space $T_{(\mathbf{\Phi}_0,X_0)}\mathcal{M}(Y)$ to the constraint manifold $\mathcal{M}(Y)$ at the point $(\mathbf{\Phi}_0, X_0)$ is the collection of the derivatives $(\mathbf{\Phi}', X') := (\mathbf{\Phi}'(0), X'(0))$ of all smooth functions $\epsilon \mapsto (\mathbf{\Phi}(\epsilon), X(\epsilon))$ which satisfy $\forall\epsilon, (\mathbf{\Phi}(\epsilon), X(\epsilon)) \in \mathcal{D}$ and $(\mathbf{\Phi}(0), X(0)) = (\mathbf{\Phi}_0, X_0)$.
To characterise the tangent space $T_{\mathbf{\Phi}_0}\mathcal{D}$ and $T_{(\mathbf{\Phi}_0,X_0)}\mathcal{M}(Y)$ in the following two lemmata, we use the decomposition $\mathbf{\Phi}_0^\star\mathbf{\Phi}_0 = I + \mathbf{M}_0$ introduced in Equation (7.11) and the notion of *admissible* matrices. A square $K \times K$ matrix C is said to be admissible if $\mathbf{\Phi}' := \mathbf{\Phi}_0 \cdot C \in T_{\mathbf{\Phi}_0}\mathcal{D}$.

Lemma 7.4.1. *Let $\mathbf{\Phi}_0 \in \mathcal{D}$ be a complete dictionary with nonzero columns.*

a. Any matrix $\mathbf{\Phi}' \in T_{\mathbf{\Phi}_0}\mathcal{D}$ can be written as $\mathbf{\Phi}' = \mathbf{\Phi}_0 \cdot C$ for some admissible C.

b. The matrix C is admissible if, and only if there exists a zero-diagonal matrix $\mathbf{Z}$ such that

$$C = \mathbf{Z} - \operatorname{diag}(\mathbf{M}_0\mathbf{Z}) \tag{7.12}$$

Proof: The first claim is a trivial consequence of the completeness of $\mathbf{\Phi}_0$, which shows that any matrix can be written as $\mathbf{\Phi}_0 \cdot C$, and the definition of an admissible matrix.
For the second part note that the constraint $\|\varphi_k\|_2 = 1$, $\forall k$ can be rewritten as $\operatorname{diag}(\mathbf{\Phi}^\star\mathbf{\Phi}) = \mathbf{I}$. Taking the derivative, it follows that $\mathbf{\Phi}' \in T_{\mathbf{\Phi}_0}\mathcal{D}$ if, and only if, $\operatorname{diag}(\mathbf{\Phi}_0^\star\mathbf{\Phi}') = 0$. Writing $\mathbf{\Phi}' = \mathbf{\Phi}_0 \cdot C$ and decomposing $C = \mathbf{Z} + \mathbf{\Delta}$ into a zero-diagonal and a diagonal matrix, we obtain from Lemma 7.3.1

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

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

Lemma 7.4.2. *The pair $(\mathbf{\Phi}', X')$ is in the tangent space $T_{(\mathbf{\Phi}_0,X_0)}\mathcal{M}(Y)$ if, and only if, there exists an arbitrary admissible matrix C and an arbitrary element $\mathbf{V}$ of $\mathcal{N}(\mathbf{\Phi}_0)$ such that*

$$\mathbf{\Phi}' = \mathbf{\Phi}_0 \cdot C \tag{7.13}$$
$$X' = -CX_0 + \mathbf{V}. \tag{7.14}$$

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

Using this explicit expression for elements of tangent space we can now turn to the main result of this section, the characterisation of local minima.

7.4.2 Characterisation of Local Minima

Lemma 7.4.3. *Consider a complete dictionary $\mathbf{\Phi}_0 \in \mathcal{D}$ and a coefficient matrix X_0 such that $\mathbf{\Phi}_0 X_0 = Y$. Define the $K \times K$ matrix*

$$\mathbf{U} := \operatorname{sign}(X_0) X_0^\star - \mathbf{M}_0^\star \operatorname{diag}(\|x^k\|_1). \tag{7.15}$$

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

$$|\langle \mathbf{Z}, \mathbf{U} \rangle_F + \langle \mathbf{V}, \operatorname{sign}(X_0) \rangle_F| < \|(\mathbf{Z}X_0 + \mathbf{V})_{\overline{\Lambda}}\|_1. \tag{7.16}$$

then $(\mathbf{\Phi}_0, X_0)$ is a strict local minimum of (7.7).

b. If the reversed strict inequality holds in (7.16) for some zero-diagonal $\mathbf{Z}$ and some $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$ such that $\mathbf{Z}X_0 + \mathbf{V} \neq 0$, then $(\mathbf{\Phi}_0, X_0)$ is not *a local minimum of (7.7).*

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

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

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

$$\nabla^+_{\mathbf{\Phi}', X'} \|X\|_1 := \lim_{\epsilon \to 0, \epsilon > 0} \frac{\|X(\epsilon)\|_1 - \|X_0\|_1}{\epsilon} = +\|(X')_{\overline{\Lambda}}\|_1 + \langle X', \operatorname{sign}(X_0) \rangle_F \tag{7.17}$$

$$\nabla^-_{\mathbf{\Phi}', X'} \|X\|_1 := \lim_{\epsilon \to 0, \epsilon < 0} \frac{\|X(\epsilon)\|_1 - \|X_0\|_1}{\epsilon} = -\|(X')_{\overline{\Lambda}}\|_1 + \langle X', \operatorname{sign}(X_0) \rangle_F, \tag{7.18}$$

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

$$|\langle X', \operatorname{sign}(X_0) \rangle_F| < \|(X')_{\overline{\Lambda}}\|_1. \tag{7.19}$$

Vice-versa, the ℓ_1-criterion does not admit a local minimum at $(\mathbf{\Phi}_0, X_0)$ if there exists some $(\mathbf{\Phi}', X')$ in the tangent space $T_{(\mathbf{\Phi}_0, X_0)} \mathcal{M}(Y)$ yielding the reversed strict inequality.
Using Lemma 7.4.2 we get that the ℓ_1-criterion admits a local minimum at $(\mathbf{\Phi}_0, X_0)$ if for all admissible C and all $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$ such that $\mathbf{V} \neq CX_0$ we have

$$|\langle CX_0 + \mathbf{V}, \operatorname{sign}(X_0) \rangle_F| < \|(CX_0 + \mathbf{V})_{\overline{\Lambda}}\|_1. \tag{7.20}$$

The rest of the proof consists in rewriting (7.20) using Lemma 7.4.1 and the properties (7.9) and (7.10). First, using (7.9), Inequality (7.20) is equivalent to

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

Second, by Lemma 7.4.1, the admissible matrices are exactly the matrices $C = \mathbf{Z} - \operatorname{diag}(\mathbf{M}_0 \mathbf{Z})$, with $\mathbf{Z}$ an arbitrary zero-diagonal matrix. Since $(\mathbf{\Delta} \cdot X_0)_{\overline{\Lambda}} = 0$ for any diagonal matrix $\mathbf{\Delta}$, we get

$(CX_0)_{\overline{\Lambda}} = (\mathbf{Z}X_0)_{\overline{\Lambda}}$ for any admissible matrix. The inequality is therefore equivalent to

$$|\langle \mathbf{Z} - \operatorname{diag}(\mathbf{M}_0\mathbf{Z}), \operatorname{sign}(X_0)X_0^\star\rangle_F + \langle \mathbf{V}, \operatorname{sign}(X_0)\rangle_F| < \|(\mathbf{Z}X_0 + \mathbf{V})_{\overline{\Lambda}}\|_1 \tag{7.21}$$

with arbitrary zero-diagonal $\mathbf{Z}$ and $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$.
Third, since $\operatorname{diag}(\operatorname{sign}(X_0)X_0^\star) = \operatorname{diag}(\|x^k\|_1)$, we observe using (7.9) and (7.10) that

$$\langle \operatorname{diag}(\mathbf{M}_0\mathbf{Z}), \operatorname{sign}(X_0)X_0^\star\rangle_F, = \langle \mathbf{M}_0\mathbf{Z}, \operatorname{diag}(\operatorname{sign}(X_0)X_0^\star)\rangle_F = \langle \mathbf{Z}, \mathbf{M}_0^\star \operatorname{diag}(\|x^k\|_1)\rangle_F. \tag{7.22}$$

Hence Inequality (7.21) is equivalent to

$$\left|\langle \mathbf{Z}, \operatorname{sign}(X_0)X_0^\star - \mathbf{M}_0^\star \operatorname{diag}(\|x^k\|_1)\rangle_F + \langle \mathbf{V}, \operatorname{sign}(X_0)\rangle_F\right| < \|(\mathbf{Z}X_0 + \mathbf{V})_{\overline{\Lambda}}\|_1.$$

□

7.5 Local Identifiability Conditions for Basis Learning

The characterisation of local minima derived in the last section is very general but also still quite abstract as it relies on the auxiliary matrices $\mathbf{Z}$ and V. Here we specialise our results to the case of a basis, i.e. when the number of atoms equals the signal dimension $K = d$ and the atoms of $\mathbf{\Phi}_0$ are linearly independent. This leads to get a more concrete if only sufficient local identifiability condition. To formulate the condition, we introduce the following block decomposition of the matrix X_0 (see Figure 7.1):

- x^k is the k-th row of X_0;
- Λ_k is the set indexing the nonzero entries of x^k and $\overline{\Lambda}_k$ the set indexing its zero entries;
- s^k is the row vector $\operatorname{sign}(x^k)_{\Lambda^k}$;
- X_k (resp. $\bar{X}_k$) is the matrix obtained by removing the k-th row of X_0 and keeping only the columns indexed by Λ_k (resp. $\overline{\Lambda}_k$) .

We also define m_k the k-th column of the matrix $\mathbf{M}_0$ and $\bar{m}_k := (\langle \varphi_\ell, \varphi_k\rangle)_{1\le\ell\le K, \ell\neq k}$, the k-th column of the matrix $\mathbf{M}_0$ without the zero entry corresponding to the diagonal.

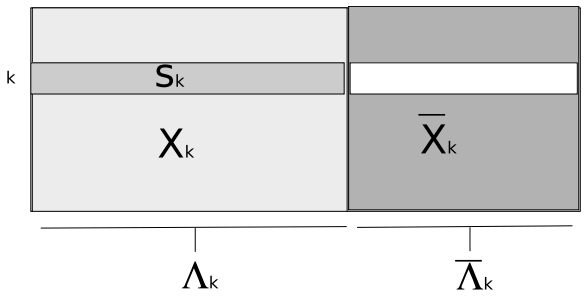

Figure 7.1: Block decomposition of the matrix X_0 with respect to a given row x^k. Without loss of generality, the columns of X_0 have been permuted so that the first $|\Lambda_k|$ columns hold the nonzero entries of x^k while the last $|\overline{\Lambda}_k|$ hold its zero entries.

Theorem 7.5.1. *Consider a $K \times N$ matrix X_0. If for every k there exists a vector d_k with $\max_k \|d_k\|_\infty < 1$ such that*

$$\bar{X}_k d_k = X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k. \tag{7.23}$$

then $(\mathbf{\Phi}_0, X_0)$ constitutes a strict local minimum of the ℓ_1-criterion.

The proof of the Theorem is based on the next lemma, which decouples the recovery condition in (7.16) into conditions expressed independently for each k.

Lemma 7.5.2. *Assume that $\mathbf{\Phi}_0$ is a basis. The recovery condition in* (7.16) *is satisfied for all nonzero zero-diagonal matrices $\mathbf{Z}$ if and only if for all k and for all $z \in \mathbb{R}^{K-1}\backslash\{0\}$ we have*

$$|\langle X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k, z\rangle| < \|\bar{X}_k^\star z\|_1. \tag{7.24}$$

Proof: When $\mathbf{\Phi}_0$ is a basis the null space is $\mathcal{N}(\mathbf{\Phi}_0) = \{0\}$ and the recovery condition (7.16) is satisfied for all nonzero zero-diagonal matrices $\mathbf{Z}$ and $\mathbf{V} \in \mathcal{N}(\mathbf{\Phi}_0)$ such that $\mathbf{Z}X_0 + \mathbf{V} \neq 0$ if, and only if, for all nonzero zero-diagonal matrices $\mathbf{Z}$ we have

$$|\langle \mathbf{Z}, \mathbf{U}\rangle_F| < \|(\mathbf{Z}X_0)_{\overline{\Lambda}}\|_1. \tag{7.25}$$

Denote z^k the k-th row of the zero diagonal matrix $\mathbf{Z}$, a row vector in $\mathbb{R}^K$ with a zero entry at the k-th coordinate, and $\bar{z}^k$ the row vector in $\mathbb{R}^{K-1}$ obtained by removing this zero entry. Observe that the k-th row of $\mathbf{Z}X_0$ is $z^kX_0 = \bar{z}^kX_0^k$ where X_0^k is X_0 with the k-th row removed. As a consequence the right hand side is decomposed into the sum

$$\|(\mathbf{Z}X_0)_{\overline{\Lambda}}\|_1 = \sum_k \|(z^kX_0)_{\overline{\Lambda}_k}\|_1 = \sum_k \|(\bar{z}^kX_0^k)_{\overline{\Lambda}_k}\|_1 = \sum_k \|\bar{z}^k(X_0^k)_{\overline{\Lambda}_k}\|_1 = \sum_k \|\bar{z}^k\bar{X}_k\|_1. \tag{7.26}$$

Now we decompose the left-hand side into a similar sum. First, we observe that

$$\langle \mathbf{Z}, \operatorname{sign}(X_0)X_0^\star\rangle_F = \langle \mathbf{Z}X_0, \operatorname{sign}(X_0)\rangle_F = \sum_k \langle z^kX_0, \operatorname{sign}(x^k)\rangle = \sum_k \langle \bar{z}^kX_0^k, \operatorname{sign}(x_k)\rangle$$
$$\langle \mathbf{Z}, \mathbf{M}_0^\star \operatorname{diag}(\|x^k\|_1)\rangle_F = \sum_k \langle z^k, m_k^\star \operatorname{diag}(\|x^k\|_1)\rangle = \sum_k \langle z^k, \bar{m}_k^\star \operatorname{diag}(\|x^j\|_1)_{j\neq k}\rangle.$$

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

$$\langle \bar{z}^kX_0^k, \operatorname{sign}(x^k)\rangle = \langle \bar{z}^k[X_k; \bar{X}_k], [s^k; 0]\rangle = \langle \bar{z}^kX_k, s^k\rangle = \langle \bar{z}^k, s^kX_k^\star\rangle.$$

and (7.25) holds for all nonzero zero-diagonal matrix $\mathbf{Z}$ if, and only if,

$$\left|\sum_k \langle \bar{z}^k, s^kX_k^\star - \bar{m}_k^\star \operatorname{diag}(\|x^j\|_1)_{j\neq k}\rangle\right| < \sum_k \|\bar{z}^k\bar{X}_k\|_1,$$

for all $(\bar{z}^k)_{k=1}^K$ with at least one row vector $\bar{z}^k \neq 0$. After transposing all the expressions it is easy to check that a necessary and sufficient condition is

$$|\langle X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k, z\rangle| < \|\bar{X}_k^\star z\|_1,\ \forall k,\ \forall z \neq 0.$$

□

Proof: [Theorem 7.5.1] For d_k with $\max_k \|d_k\|_\infty < 1$ as in (7.23) we get

$$|\langle X_k(s^k)^\star - \mathrm{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k, z\rangle| = |\langle \bar{X}_k d_k, z\rangle| = |\langle d_k, \bar{X}_k^\star z\rangle| \leq \|d_k\|_\infty \|\bar{X}_k^\star z\|_1 < \|\bar{X}_k^\star z\|_1,$$

which by Lemma 7.5.2 guarantees recovery. □

The lemma above is also the starting point to showing via duality analysis that Condition 7.23 is not only sufficient but also necessary. We refer to [26] for more details.

7.6 Example - Ideally Sparse Training Data

Assume that the coefficient matrix X_0 has the following structure:

a. each column x_n is 'ideally' sparse, in the sense that it has exactly one nonzero component. This means that each training sample $y_n = \mathbf{\Phi}_0 \cdot x_n$ is colinear to some dictionary vector;

b. each row x^k has at least one nonzero component, meaning that the direction of each dictionary vector is represented at least once in the training samples.

Using Theorem 7.5.1 let us check for which bases $\mathbf{\Phi}_0$ such properties of X_0 imply that the pair $(\mathbf{\Phi}_0, X_0)$ is a local minimum. We can rearrange the matrix X_0 so that first we have all the columns who have the non-zero entry in the first row, then the ones with the non-zero in the second row etc.

$$\tilde{X}_0 = \begin{pmatrix} \tilde{x}^1 & 0 & \dots & 0 \\ 0 & \tilde{x}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{x}^K \end{pmatrix}.$$

The first observation from the rearrangement above is that for each k the split into X_k and $\bar{X}_k$ will result in a zero matrix X_k because the only nonzero entries are on the k-th row. Thus we have $X_k(s^k)^\star = 0$ and just need to show that we can find d_k with $\|d_k\|_\infty < 1$ such that $\mathrm{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k = \bar{X}_k d_k$. This means that for every component $\bar{m}_k(i) = \langle \varphi_k, \varphi_i\rangle$ we need to satisfy

$$\forall i: \ \langle \varphi_k, \varphi_i\rangle = \frac{\langle \bar{x}_k^i, d_k\rangle}{\|x^i\|_1},$$

where $\bar{x}_k^i$ denotes the i-th row of $\bar{X}_k$. Because of the ideally sparse structure of X_0 the index sets Ω_k^i where the rows $\bar{x}_k^i$ are non zero do not overlap, i.e. for $i \neq j$ we have $\Omega_k^i \cap \Omega_k^j = \emptyset$, and the conditions we need to satisfy are independent. So if we choose d_k such that $d_k|_{\Omega_k^i} = c_k^i \mathrm{sign}(\bar{x}_k^i)|_{\Omega_k^i}$ we see that we should have

$$|\langle \varphi_k, \varphi_i\rangle| = |c_k^i| < 1, \tag{7.27}$$

which is always satisfied and we get that any basis will in combination with ideally sparse data constitute a local minimum.

7.7 Probabilistic Analysis

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

7.7.1 The Model

We assume that the entries x_{kn} of the $K \times N$ coefficient matrix X are i.i.d. with $x_{kn} = \varepsilon_{kn} g_{kn}$, where the ε_{kn} are indicator variables taking the value one with probability p and zero with probability $1-p$, i.e. $\varepsilon \sim p\delta_1 + (1-p)\delta_0$. The variables g_{nk} follow a standard Gaussian distribution, i.e. centered with unit variance.
The important role of the indicator variables is to guarantee a strictly positive probability that the entry x_{kn} is exactly zero. The assumption that the g_{nk} are centered Gaussians with unit variance is mainly for simplicity reasons as it allows us to do all proofs using only elementary probability theory. However, we believe that the same results hold for many other distributions as long as they show a certain amount of concentration, as for instance Bernoulli ± 1 with equal probability or any other subgaussian distribution.
Let us start with a geometric interpretation of the necessary recovery conditions.

7.7.2 Geometric Inspiration

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

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

We can see that the probability of satisfying both conditions will largely depend on the number of non zero coefficients in each row. The more zeros the shorter the vectors s^k and x^k, thus the more likely that $\|u_k\|_2 = \|X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k\|_2$ is small, and the higher the dimension of the unit cube, thus more chances its image covers a big ball. So we get a higher probability to recover a basis, the sparser the signals are and the more incoherent the basis is, i.e. the smaller $\|\bar{m}_k\|_2 = \|m_k\|_2$. The following theorem gives concrete estimates, derived by working out the details of the geometric sketch above.

7.7.3 Main Theorem

Theorem 7.7.1. *Denote the event 'the original basis is not a local minimum of the ℓ_1-criterion' shortly by '☹'. If for a basis $\mathbf{\Phi}$ we have $\max_k \|m_k\|_2 < \frac{1-2p}{20}$ and the number of randomly generated training signals exceeds $N > \frac{600(K-1)}{(1-2p)^2}$ where $p < 1/2$, the probability of '☹' decays as*

$$\mathbb{P}(☹) \leq 2K \left[\exp\left((K-1)\log(61\sqrt{\tfrac{K-1}{p}}) - \frac{(1-2p)pN}{13}\right) + \exp\left(-\frac{(1-2p)^2 pN}{800}\right) + (K-1)\exp\left(-\frac{pN}{4}\right) + \exp\left(-2p^2 N\right)\right] \tag{7.28}$$

The crucial probabilities in the bound above are the first exponential because of the term $\mathcal{O}(K \log K)$ and the second one because of the horrible constant 1/800. The third is dominated by the first and for $p > 1/1603$ the last exponential is dominated by the second one. Thus in this case we can get the cruder but more readable bound.

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

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

To state the theorem in a concrete form, we had to make some rough decisions on the way, crudely bounding some intermediate probabilities. The next subsection gives a skeleton of the proof, indicating where these choices had to be made, so in case all parameters, coherence and size of the basis, probability of a coefficient to be non zero and number of training signals, are precisely known, it is easy to retrace the steps and get the optimal bounds. In the course of that we will also prove the following simple but totally abstract theorem.

Theorem 7.7.2. *If for a basis $\mathbf{\Phi}$ we have $\max_k \|m_k\|_2 < (1-p)$, then there exist constants $b > 0$ and $a, c < \infty$, depending only on p, such that for $N > c \cdot d$ we have*

$$\mathbb{P}(\text{☹}) \leq \exp(a \cdot d \log d - b \cdot N). \tag{7.29}$$

7.7.4 Skeleton of the Proof - Probability Split

To estimate the overall probability that the original basis is not a local minimum of the ℓ_1-criterion, we have a look at all aspects of the sufficient condition in (7.23) that could possibly go wrong and bound their probabilities individually. First we can take the union bound over every row index k,

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

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

$$\begin{aligned}\mathbb{P}(\text{☹}_k) &= \sum_{M=0}^{N} \mathbb{P}\big(\text{☹}_k \,|\, |\bar{\Lambda}_k| = M\big) \cdot \mathbb{P}\big(|\bar{\Lambda}_k| = M\big) \\ &\leq \sum_{M=M_l}^{M_u} \mathbb{P}\big(\text{☹}_k \,|\, |\bar{\Lambda}_k| = M\big) \cdot \mathbb{P}\big(|\bar{\Lambda}_k| = M\big) + \mathbb{P}\big(|\bar{\Lambda}_k| < M_l \cup |\bar{\Lambda}_k| > M_u\big) \\ &\leq \max_{M_l \leq M \leq M_u} \mathbb{P}\big(\text{☹}_k \,|\, |\bar{\Lambda}_k| = M\big) + \mathbb{P}\big(|\bar{\Lambda}_k| < M_l \cup |\bar{\Lambda}_k| > M_u\big)\end{aligned}$$

*Given only K training signals the dictionary giving the sparsest representation is the set of training signals itself.

To bound the probability of the first term in the expression above, we use the geometric inspiration from Subsection 7.7.2.

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

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

$$\begin{aligned}\mathbb{P}(\circledcirc) \leq \sum_{k=1}^{K} \Big\{ \max_{M_l \leq M \leq M_u} \big[\mathbb{P}\big(\bar{X}_k(Q^M) \not\supseteq B_2^{K-1}(\alpha_M)\big) + \mathbb{P}\big(\|u_k\|_2 > \alpha_M \,|\, |\bar{\Lambda}_k| = M\big)\big] \\ + \mathbb{P}\big(|\bar{\Lambda}_k| < M_l \cup |\bar{\Lambda}_k| > M_u\big)\Big\}. \quad (7.30)\end{aligned}$$

From (7.30) it becomes clear how important it is to carefully choose the parameters M_l, M_u and α_M to keep the sum of all probabilities small. However, to make this choice we first need to estimate the magnitude of the probabilities involved.

7.7.5 Estimating the Individual Probabilities

All estimates are based on concentration of measure results to bound the probability that a random variable deviates a lot from its expected value.
We start with the easiest estimate, the probability of the number of zero coefficients in each row being below M_l or above M_u, using Hoeffding's inequality.

Theorem 7.7.3. *Let $Y_1 \dots Y_N$ be independent random variables. Assume that the Y_n are almost surely bounded, meaning for $1 \leq i \leq N$ we have $\mathbb{P}(Y_n \in [a_n, b_n]) = 1$. Then, for the sum of these variables $S = Y_1 + \dots + Y_N$ we have the inequality*

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

which is valid for positive values of t, where E(S) is the expected value of S.

In each row the number of zero coefficients $|\Lambda_k|$ is N minus the number of non-zero coefficients $|\bar{\Lambda}_k|$, which is the sum of the indicator variables $\sum_n \varepsilon_{kn}$. The ε_{nk} are taking only the values zero and one, so $a_i = 0$, $b_i = 1$ and $\mathbb{E}(\sum_n \varepsilon_{kn}) = pN$ leading to

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

Choosing $t = (1-p)\varepsilon_\Lambda$ and inserting $|\bar{\Lambda}_k| = N - |\Lambda_k|$ we get

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

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

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

So if we set $M_l = N(1-p)(1-\varepsilon_\Lambda)$ and $M_u = N(1-p)(1+\varepsilon_\Lambda)$ we get that

$$\mathbb{P}\big(|\bar{\Lambda}_k| < M_l \cup |\bar{\Lambda}_k| > M_u\big) \leq 2\exp(-2N(1-p)^2\varepsilon_\Lambda^2).$$

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

Lemma 7.7.4. *Let A be a matrix of size $d \times M$. The image of the unit cube Q^M by A contains a Euklidean ball of size α if and only if for all x with $\|x\|_2 = 1$ there exists a $v \in Q^M$, i.e. $\|v\|_\infty \leq 1$ such that $|\langle Av, x\rangle| \geq \alpha$.*

Proof: It will be easier to prove the converse statement:

$$A(Q^M) \not\supseteq B_2^d(\alpha) \qquad \Leftrightarrow \qquad \exists x, \|x\|_2 = 1, \text{ s.t. } \forall v \in Q^M, |\langle Av, x\rangle| < \alpha$$

While the $\Leftarrow$ direction is obvious the $\Rightarrow$ direction is slightly more tricky.
The image of Q^M by A is a convex polygon, that is symmetric around the origin. Let $\beta < \alpha$ be the radius of the largest ball that can be inscribed into $A(Q^M)$. Choose $\pm x$ a pair of vectors where the ball $B_2^d(\beta)$ touches the surface of the polygon. There the tangent planes to the ball $h^+ : \langle y, \frac{x}{\|x\|_2}\rangle = \beta$, $h^- : \langle y, \frac{x}{\|x\|_2}\rangle = -\beta$ are parallel to the facets of the polygon and as $A(Q^M)$ is convex, it is enclosed between them, i.e. $A(Q^M) \subseteq \{y : |\langle y, \frac{x}{\|x\|_2}\rangle| \leq \beta\}$. Thus for the unit norm vector $x_\beta = \frac{x}{\|x\|_2}$ and for all $v \in Q^M$ we have $|\langle Av, x_\beta\rangle| \leq \beta < \alpha$. □

Lemma 7.7.5. *If there exists an $\varepsilon_\mathcal{N}$-net $\mathcal{N}$ for the unit sphere in $\mathbb{R}^d$ such that for all $x_i \in \mathcal{N}$ we have a $v_i \in Q^M$ such that $|\langle Av_i, x_i\rangle| \geq \alpha$ and $\|A\|_{2,\infty} \leq \beta$, then $A(Q^M) \supseteq B_2^d(\alpha - \beta\varepsilon_\mathcal{N})$.*

Proof: By Lemma 7.7.4 we need to show that for all x with unit norm we can find $v \in Q^M$ such that $|\langle Av, x\rangle| \geq \alpha - \beta\varepsilon_\mathcal{N}$. Since $\mathcal{N}$ is an $\varepsilon_\mathcal{N}$-net we can find $x_0 \in \mathcal{N}$ with $\|x - x_0\|_2 < \varepsilon_\mathcal{N}$. For v_0 we then have

$$|\langle Av_0, x\rangle| \geq |\langle Av_0, x_0\rangle| - |\langle Av_0, x - x_0\rangle| \geq \alpha - \|Av_0\|_2\|x - x_0\|_2 \geq \alpha - \beta\varepsilon_\mathcal{N}.$$

□

As a corollary to the lemma above we get the following probabilistic estimate.

Corollary 7.7.6. *Choose an $\varepsilon_\mathcal{N}$-net $\mathcal{N}$ for the unit sphere in $\mathbb{R}^d$ with $|\mathcal{N}| \leq (\frac{6}{\varepsilon_\mathcal{N}})^d$. For a 'random' $d \times M$ matrix $A = (A_1 \dots A_M)$ we can bound the probability that $A(Q^M)$ covers a ball of radius $\alpha - \beta\varepsilon_\mathcal{N}$ as*

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

Proof: A direct consequence of Lemma 7.7.5 and the following two observations

$$\sup_{\|v\|_\infty \leq 1} |\langle Av, x_i\rangle| = \sup_{\|v\|_\infty \leq 1} |\langle v, A^\star x_i\rangle| = \|A^\star x_i\|_1,$$
$$\|A\|_{2,\infty} \leq \sum_i \|A_i\|_2.$$

□

To finally get a quantitative estimate, we need the following two concentration of measure inequalities, whose proofs can be found in the appendix of [26].

Theorem 7.7.7. *Let $A = (A_1 \dots A_M)$ be a matrix of size $d \times M$, whose entries follow the distribution described in Subsection 7.7.1, $A_{ij} = \varepsilon_{ij} g_{ij}$, $i = 1 \dots d$, $j = 1 \dots M$, and $x \in \mathbb{R}^d$ be a vector with unit*

norm. Then

$$a) \qquad \mathbb{P}\big(\|A^\star x\|_1 \le Mp(\sqrt{\tfrac{2}{\pi}} - \varepsilon_\alpha)\big) \le 2\exp\left(-\frac{\varepsilon_\alpha^2 Mp}{2+\sqrt{2}\varepsilon_\alpha}\right),$$

$$b) \qquad \mathbb{P}\big(\sum_{j=1}^{M} \|A_j\|_2 \ge M\sqrt{pd}(1+\varepsilon_\beta)\big) \le 2\exp\left(-\frac{\varepsilon_\beta^2 M\sqrt{p}}{2\sqrt{p}+\sqrt{2}\varepsilon_\beta}\right).$$

From the first equation we see that α has to smaller than $\sqrt{\frac{2}{\pi}}Mp$. Indeed, since we also have the converse bound, i.e.

$$\mathbb{P}\big(\|Ax\|_1 \ge Mp(\sqrt{\tfrac{2}{\pi}} + \varepsilon_\alpha)\big) \le 2\exp\left(-\frac{\varepsilon_\alpha^2 Mp}{2+\sqrt{2}\varepsilon_\alpha}\right), \tag{7.31}$$

the probability of finding a unit vector x such that $\sup_{\|v\|_\infty \le 1} |\langle Av, x\rangle| = \|A^\star x\|_1 < Mp(\sqrt{\frac{2}{\pi}} + \varepsilon_\alpha)$ rapidly approaches 1, meaning that the radius of the maximal ball cannot exceed $Mp(\sqrt{\frac{2}{\pi}} + \varepsilon_\alpha)$. In an attempt to simultaneously balance the resulting probabilities and keep them readable we choose $\varepsilon_\alpha = \sqrt{2/\pi} - 1/3$, leading to $\alpha = Mp/3$, $\varepsilon_\beta = 1/3$, leading to $\beta = 4M\sqrt{pd}/3$, and $\varepsilon_{\mathcal{N}} = 10^{-1}\sqrt{p/d}$. Using Corollary 7.7.6 we arrive at

$$\mathbb{P}\big(A(Q^M) \not\supseteq B_2^d(\frac{Mp}{5})\big) \le 2(60\sqrt{\tfrac{d}{p}})^d \exp\left(-\frac{Mp}{13}\right) + 2\exp\left(-\frac{M\sqrt{p}}{18\sqrt{p}+3\sqrt{2}}\right)$$

Note that for $p \le \frac{1}{2}$ we have $\exp\left(-\frac{M\sqrt{p}}{18\sqrt{p}+3\sqrt{2}}\right) \le \exp\left(-\frac{Mp}{12}\right)$, which leads to the simpler bound,

$$\mathbb{P}\big(A(Q^M) \not\supseteq B_2^d(\frac{Mp}{5})\big) \le 2\exp\left(d\log(61\sqrt{\frac{d}{p}}) - \frac{Mp}{13}\right). \tag{7.32}$$

Last we will estimate the probability that the vector $u_k = X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k$ is not contained in the Euklidean ball of radius $\alpha = Mp/5$.
One way to make sure $\|u_k\|_2$ is small is to check that both its components are small, i.e. if $\|X_k(s^k)^\star\|_2$ is smaller than $q\alpha$ for some $q \in [0,1]$ and $\|\operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k\|_2$ is smaller than $(1-q)\alpha$, we have $\|X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k\|_2 \le \|X_k(s^k)^\star\|_2 + \|\operatorname{diag}(\|x^j\|_1)_{j\neq k}m_k\|_2 < \alpha$, leading to the bound

$$\begin{aligned}\mathbb{P}\big(\|X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k\|_2 > \alpha \,|\, |\bar{\Lambda}_k| = M\big)\\ \le \mathbb{P}\big(\|X_k(s^k)^\star\|_2 > q\alpha \,|\, |\bar{\Lambda}_k| = M\big) + \mathbb{P}\big(\|\operatorname{diag}(\|x^j\|_1)_{j\neq k}m_k\|_2 > (1-q)\alpha\big).\end{aligned}$$

Using the fact that $\|\operatorname{diag}(\|x^j\|_1)_{j\neq k}m_k\|_2 \le \max_{j\neq k}\|x^j\|_1\|m_k\|_2$ and a union bound over j the second term in the equation above can in turn be bounded as

$$\mathbb{P}\big(\|\operatorname{diag}(\|x^j\|_1)_{j\neq k}m_k\|_2 > (1-q)\alpha\big) \le \sum_{j\neq k}\mathbb{P}\big(\|x^j\|_1\|m_k\|_2 > (1-q)\alpha\big),$$

so that we finally get

$$\begin{aligned}\mathbb{P}\big(\|X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k\|_2 > \alpha \,|\, |\bar{\Lambda}_k| = M\big)\\ \le \mathbb{P}\big(\|X_k(s^k)^\star\|_2 > q\alpha \,|\, |\bar{\Lambda}_k| = M\big) + \sum_{j\neq k}\mathbb{P}\big(\|x^j\|_1\|m_k\|_2 > (1-q)\alpha\big).\end{aligned} \tag{7.33}$$

To keep the sum of the two probabilities small, it its again necessary to carefully choose the size of parameter q, which will depend on the magnitude of $\|m_k\|_2$ measuring the coherence of the basis. It is easy to see that when the basis is orthogonal we have $\|m_k\|_2 = 0$ and can set $q = 1$. For further bounds we need another two concentration of measure results, whose proofs can again be found in the appendix of [26].

Theorem 7.7.8. *a. Let B be a matrix of size $d \times L$, whose entries follow the distribution described in Subsection 7.7.1, $B_{ij} = \varepsilon_{ij} g_{ij}$, $i = 1 \dots d$, $j = 1 \dots L$, and s be a vector of length L with entries $s_j = \pm 1$, $j = 1 \dots L$. Then for $\varepsilon_s > 0$*

$$\mathbb{P}\big(\|Bs\|_2^2 \geq dLp(1+\varepsilon_s)\big) \leq 2\exp\left(-\frac{dp\varepsilon_s^2}{6+2\varepsilon_s}\right). \tag{7.34}$$

b. Let x be a vector of length N, whose entries follow the distribution described in Subsection 7.7.1, $x_i = \varepsilon_i g_i$, $i = 1 \dots N$. Then for $\varepsilon_m > 0$

$$\mathbb{P}\big(\|x\|_1 \geq pN(\sqrt{\tfrac{2}{\pi}} + \varepsilon_m)\big) \leq 2\exp\left(-\frac{pN\varepsilon_m^2}{2+\varepsilon_m/\sqrt{2}}\right). \tag{7.35}$$

We apply the theorem to the matrix X_k, the vector s^k and the vectors x^j to further bound the probability in (7.33). Write shortly $d = K - 1$. If $(q\alpha)^2 > dLp$ and $(1-q)\alpha > \sqrt{\frac{2}{\pi}} pN\|m_k\|_2$, we set $\varepsilon_s = \frac{(q\alpha)^2}{dLp} - 1$ and $\varepsilon_m = \frac{(1-q)\alpha}{pN\|m_k\|_2} - \sqrt{\frac{2}{\pi}}$ to get

$$\mathbb{P}\big(\|X_k(s^k)^\star + \|x_k\|_1 \bar{m}_k\|_2 > \alpha\big) \leq 2\exp\left(-\frac{(q\alpha)^2}{2L}\cdot c_s\right) + 2d\exp\left(-\frac{(1-q)\alpha\sqrt{2}}{\|m\|_k}\cdot c_m\right),$$

$$\text{with} \qquad c_s = \frac{(1-\frac{dLp}{(q\alpha)^2})^2}{1+2\frac{dLp}{(q\alpha)^2}} \qquad \text{and} \qquad c_m = \frac{(1-\sqrt{\frac{2}{\pi}}\frac{pN\|m_k\|_2}{(1-q)\alpha})^2}{1+\frac{pN\|m_k\|_2}{(1-q)\alpha}(2\sqrt{2}-\sqrt{\frac{2}{\pi}})}.$$

Let us investigate the conditions that there exist $\varepsilon_s, \varepsilon_m > 0$ in more detail. For $\alpha = \frac{Mp}{5}$ we first need

$$1 < \frac{q^2p^2M^2}{25dLp}.$$

At worst $M = M_l = (1-\varepsilon_\Lambda)(1-p)N$ and $L = N - M_l = (\varepsilon_\Lambda + p - \varepsilon_\Lambda p)N)$ so we need that,

$$1 < \frac{N}{d}\cdot\frac{q^2p(1-\varepsilon_\Lambda)^2(1-p)^2}{25(\varepsilon_\Lambda + p - \varepsilon_\Lambda p)},$$

which will always be satisfied as soon as the number of signals N is large enough. The second condition

$$1 < \sqrt{\frac{\pi}{2}}\frac{(1-q)pM}{5pN\|m_k\|_2} \tag{7.36}$$

is more interesting as in the worst case for $M = M_l$ it is equivalent to

$$\|m_k\|_2 < \sqrt{\frac{\pi}{2}}\frac{(1-q)(1-\varepsilon_\Lambda)(1-p)}{5} < \sqrt{\frac{\pi}{2}}\frac{(1-p)}{5},$$

which means that as soon as $\|m_k\|_2 \geq \sqrt{\frac{\pi}{2}}\frac{(1-p)}{5}$ we cannot find $\varepsilon_\Lambda, q > 0$ to get an $\varepsilon_m > 0$. Looking back at the estimate of the radius of the maximal ball we see that α necessarily has to be smaller than $\sqrt{\frac{2}{\pi}}Mp$, leading to

$$\|m_k\|_2 < 1 - p.$$

This means that as soon as $\|m_k\|_2 \geq (1-p)$ the size of the vector u_k grows faster than the size of the maximal ball, and recovery can no longer be guaranteed.
However, let's assume that $\|m_k\|_2 < \frac{M}{20N}$ and choose $q = 1/\sqrt{3}$. If $M^2 > 300dL/p$ a long calculation shows that we have

$$\mathbb{P}\big(\|X_k(s^k)^\star - \operatorname{diag}(\|x^j\|_1)_{j\neq k}\bar{m}_k\|_2 > \frac{Mp}{5}\big) \leq 2\exp\left(-\frac{M^2p^2}{400L}\right) + 2d\exp\left(-\frac{Np}{4}\right)$$

If we combine this estimate with the estimate in (7.32), we can bound the probability that u_k is not in the image of the unit cube by $\bar{X}_k$ as, $(d = K-1)$,

$$\mathbb{P}\big(\text{☹}_k \,|\, |\bar{\Lambda}_k| = M\big) \leq 2\exp\left(d\log(61\sqrt{\tfrac{d}{p}}) - \frac{Mp}{13}\right) + 2\exp\left(-\frac{M^2p^2}{400L}\right) + 2d\exp\left(-\frac{Np}{4}\right) \tag{7.37}$$

Keeping in mind that $L = N - M$, we see that the expression above is the smaller the larger M is. Thus if we want to bound it over $M_l \leq M \leq M_u$ we need to insert the minimal value $M = M_l = (1-\varepsilon_\Lambda)(1-p)N$. For aesthetic reasons we choose $\varepsilon_\Lambda = p/(1-p)$, leading to $M_l = (1-2p)N$ and $N - M_l = 2pN$. Putting this together with the estimate that $M \geq M_l$ we get that if $\max_k \|m_k\|_2 < \frac{1-2p}{20}$ and $N > \frac{600(K-1)}{(1-2p)^2}$ the probability of not recovering the dictionary as local minimum of the ℓ_1-criterion can be bounded as

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

7.8 Discussion

We have developed some algebraic conditions on a dictionary coefficient pair to constitute a local minimum of the ℓ_1 dictionary learning criterion. In case the dictionary is an incoherent basis we have shown that for coefficient matrices generated from a random sparse model the resulting basis coefficient pair suffices these conditions with high probability as long as the number of training signals grows like $d\log d$. These are exciting new results but since dictionary learning is a relatively young field they lead to more open questions. For the special case when the dictionary is assumed to be a basis it would be desirable to show the converse direction, i.e. if the coherence of the basis is too high and the training signals are generated by the same random sparse model, the basis coefficient pair will not be a local minimum. Ideally this breakdown coherence $\max_k \|m_k\|_2$ would be the same or close to $(1-p)$. Another helpful result would be to prove that under the random model there exists only one local minimum which then has to be the global one, and could be found with simple descent algorithms. Numerical experiments in two dimensions support this hypothesis. Figure 7.2 is a plot of the ℓ_1-cost $\|\mathbf{\Phi}^{-1}Y\|_1$ for all possible two-dimensional bases, where both atoms are parametrised by their angle θ_i to the x-axis, $\theta_i \in [0, \pi]$. The $N = 500$ training signals $Y = \mathbf{\Phi}_0 X_0$ were generated using the random sparse model with $p = 0.5$. As can be seen the only two local

minima are at the original dictionary $\mathbf{\Phi}_0$ and at the dictionary corresponding to $\mathbf{\Phi}_0$ with permuted columns (the sign ambiguity is avoided by restricting the angles to the interval $[0, \pi]$).

Finally much harder research will have to be invested to extend the current results to the over-

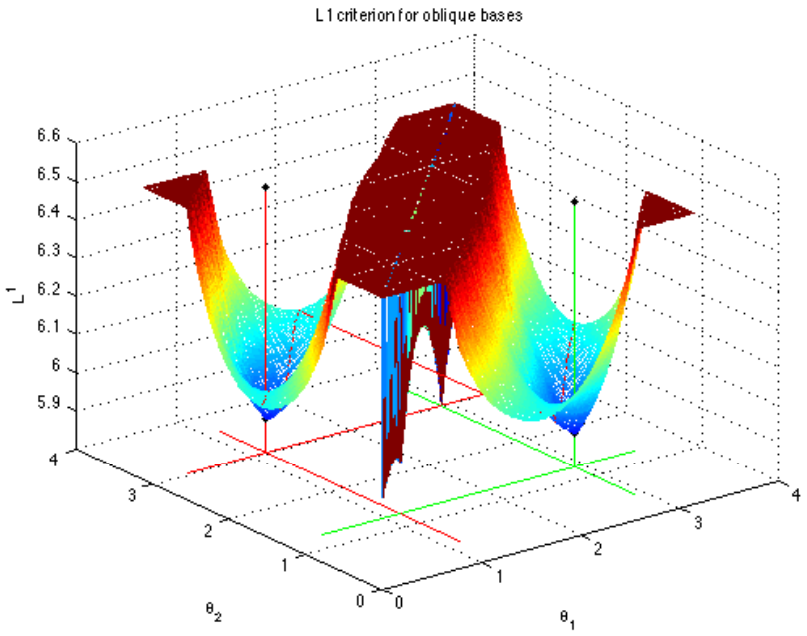

Figure 7.2: ℓ_1-cost as a function of all two-dimensional bases

complete and the noisy case. In the overcomplete case the null space has to be taken into account, which prevents a straightforward generalisation from the intrinsic conditions to the explicit ones. In the noisy case already the formulation of the problem has to be changed as we cannot expect the best dictionary for the noise contaminated training data to be exactly the same as the original dictionary but only close to it.

8 Outlook

In the first part of the thesis we have seen that even though finding sparse representations is hard, the situtation is not hopeless. In particular Chapter 2 showed that sensing dictionaries can improve algorithms like Thresholding and (O)MP, and Chapters 3/4 that on average the behaviour of both algorithms is quite good. Finding sparse representations is by now a huge field of research, with new algorithms and variants of existing ones, both general or specialised to certain dictionaries being developed every day. In short the field is being thoroughly explored. The same can be said for the topic presented in Chapter 5. Compressed Sensing is new, hot and sexy. The already existing literature is enormous, as can be seen on the Compressive Sensing Resources website at

`http://www.dsp.ece.rice.edu/cs/}.`

The situation is different for the subjects broached in the last two chapters. While classification itself is quite a big and well explored field as well, the dictionary or subspace view seems quite novel. However, while all the ideas presented here can certainly be further developed, as pointed out at the end of Chapter 6, the main message to be learned is that every element or group of elements in a dictionary can have a meaning. The same idea had already been touched at the beginning of Chapter 4 when discussing the applications of multichannel signal approximations, where every atom corresponded to a thought. Thus, keeping this connection between atoms and meanings in mind can help bring new views to many data mining problems.
Dictionary learning finally is a young and very important field. Indeed any theory about finding sparse representations or compressed sensing is only useful if you can actually find a dictionary providing these sparse representations. In that sense dictionary learning is also further research into sparse representations or compressed sensing. At the moment there exist only a handful of algorithms, some of which are too inefficient to work for real applications, and a little bit of theory. However, for real life applications what is needed are fast algorithms that can handle big data sizes and a theoretical framework to guarantee that they work, which makes exploring the directions pointed out at the end of Chapter 7 all the more important.

Bibliography

[1] D. Achlioptas (2001). Database-friendly random projections. In *Proc. 20th Annual ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems*, pp. 274–281.

[2] M. Aharon, M. Elad, A. Bruckstein (2006). On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Journal of Linear Algebra and Applications* **416**:48–67.

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

[4] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin (2007). A simple proof of the restricted isometry property for random matrices. *Constructive Approximation* .

[5] D. Baron et al. (2005). *Distributed Compressed Sensing.* Tech. rep., Rice University.

[6] T. Blumensath, M. Davies (submitted). Iterative Hard Thresholding for compressed sensing. *Applied Computational Harmonic Analysis* .

[7] R. Brunelli, T. Poggio (1993). Face recognition: Features vs. templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**(10):1042–1053.

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

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

[10] J.-F. Cardoso (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE. Special issue on blind identification and estimation* **9**(10):2009–2025.

[11] V. Cevher, M. Duarte, R. Baraniuk (2008). Distributed target localization via spatial sparsity. In *EUSIPCO, Lausanne, Switzerland.*

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

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

[14] D. Donoho (2006). Compressed Sensing. *IEEE Transactions on Information Theory* **52**(4):1289–1306.

[15] D. Donoho (2006). For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* **59**(6):797–829.

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

[17] D. Donoho, J. Tanner (2006). Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *Preprint arXiv:math.MG/0607364* .

[18] I. Drori (2007). Fast ℓ_1 minimization by iterative thresholding for multidimensional NMR spectroscopy. *EURASIP Journal on Advances in Signal Processing* .

[19] D. J. Field, B. A. Olshausen (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**:607–609.

[20] R. Fisher (1936). The use of multiple measures in taxonomic problems. *Ann. Eugenics* **7**:179–188.

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

[22] J. J. Fuchs (1998). Detection and estimation of superimposed signals. In *Proc. IEEE ICASSP98*, vol. 3, pp. 1649–1652.

[23] P. Georgiev, F. J. Theis, A. Cichocki (2005). Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks* **16**(4):992–996.

[24] A. Gilbert, J. Tropp (2007). Signal recovery from random measurements via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory* **53**(12):4655–4666.

[25] R. Gribonval, H. Rauhut, K. Schnass, P. Vandergheynst (2008). Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *Journal of Fourier Analysis and Applications* **14**(5):655–687.

[26] R. Gribonval, K. Schnass (2009). Dictionary identifiability. *in preparation* .

[27] R. Horn, C. Johnson (1985). *Matrix Analysis*. Cambridge University Press.

[28] P. Indyk (2008). Explicit constructions for compressed sensing of sparse signals. In *19th Symposium on Discrete Algorithms*.

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

[30] K. Kreutz-Delgado et al. (2003). Dictionary learning algorithms for sparse representation. *Neural Computations* **15**(2):349–396.

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

[32] K. Lee, J. Ho, D. Kriegman (2005). Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(5):684–698.

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

[34] Y. Ma, A. Yang, H. Derksen, R. Fossum (2008). Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM Review* **50**(3):413–458.

[35] J. Mairal et al. (2008). *Discriminative Learned Dictionaries for Local Image Analysis*. IMA Preprint Series 2212, University of Minnesota.

[36] A. Martinez, B. R. (1998). *The AR face database*. Technical Report 24, CVC.

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

[38] F. Naini, R. Gribonval, L. Jacques, P. Vandergheynst (2009). Compressive sampling of pulse trains: Spread the spectrum! In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

[39] D. Needell, J. Tropp (2008 (accepted)). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied Computational Harmonic Analysis* .

[40] D. Needell, R. Vershynin (DOI: 10.1007/s10208-008-9031-3). Uniform Uncertainty Principle and signal recovery via Regularized Orthogonal Matching Pursuit. *Foundations of Computational Mathematics* .

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

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

[43] M. Plumbley (2007). Dictionary learning for ℓ_1-exact sparse coding. In M. Davies, C. James, S. Abdallah (eds.), *International Conference on Independent Component Analysis and Signal Separation*, vol. 4666, pp. 406–413, Springer.

[44] H. Rauhut (2007). Random sampling of sparse trigonometric polynomials. *Applied Computational Harmonic Analysis* **22**(1):16–42.

[45] H. Rauhut (2008). Stability results for random sampling of sparse trigonometric polynomials. *IEEE Transactions on Information Theory* **54**(12):5661–5670.

[46] H. Rauhut, K. Schnass, P. Vandergheynst (2008). Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory* **54**(5):2210–2219.

[47] F. Rodriguez, G. Sapiro (2008). *Sparse Representations for Image Classification: Learning Discriminative and Reconstructive Non-Parametric Dictionaries*. IMA Preprint Series 2213, University of Minnesota.

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

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

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

[51] K. Schnass, P. Vandergheynst (submitted). Classification via incoherent subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

[52] K. Skretting, J. Husoy (2006). Texture classification using sparse frame-based representations. *EURASIP Journal on Applied Signal Processing* **2006**:11.

[53] T. Strohmer, R. Heath (2003). Grassmannian frames with applications to coding and communication. *Applied Computational Harmonic Analysis* **14**(3):257–275.

[54] D. Studer, U. Hoffmann, T. Koenig (2006). From EEG dependency multichannel matching pursuit to sparse topographic EEG decomposition. *Journal of Neuroscience Methods* **153**(2):261–275.

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

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

[57] J. Tropp, I. Dhillon, R. Heath Jr, T. Strohmer (2005). Designing structured tight frames via an alternating projection method. *IEEE Transactions on Information Theory* **51**(1):188–209.

[58] J. Wright et al. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(2).

[59] M. Yaghoobi, T. Blumensath, M. Davies (submitted). Dictionary learning for sparse approximation with the majorization method. *IEEE Transactions on Signal Processing* .

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

Curriculum Vitae

Personal Details

Name: **Karin Schnass**
Date of Birth: 3rd of May, 1980
Place of Birth: Klosterneuburg, Austria
Citizenship: Austrian
Email: karin.schnass@epfl.ch
Web page: http://lts2www.epfl.ch/∼schnass/

Education

2005 - 2009: Ph.D. student at the Signal Processing Laboratories, Swiss Federal Institute of Technology (EPFL).
2004 April: Masters thesis on Gabor Multipliers at NuHAG with Prof. H.G. Feichtinger.
2001/02 Winter: Erasmus semester at the University of Leeds, UK.
1998 - 2004: Study of Mathematics at the University of Vienna, Austria, Faculty of Mathematics.
1998 - 2000: Study of Biology at the University of Vienna, Austria, Faculty of Life Sciences.
1990 - 1998: Scientific High School, Klosterneuburg.

Work Experience

2005 - 2009: Research Assistant at the Signal Processing Laboratories, Swiss Federal Institute of Technology (EPFL).
May 2004 - January 2005: Leonardo Internship at Philips Research, Eindhoven, NL.
October 2003 - January 2004: Mathematics teaching assistant at the University of Vienna and the University of Natural Resources and Applied Life Sciences, Vienna.
before: Postgirl, waitress, zoo guide, swimming instructor.

Languages

German: native
English: fluent
French: decent
Italian: decent
Dutch: basic

INTERESTS

Sports: cycling, swimming, tennis, skiing, hiking
Leasure: travelling, reading, dancing
Art: guitar, painting, sewing

Personal Publications

Journal Papers

R. Gribonval, K. Schnass (2009). Dictionary identification. *in preparation.*

K. Schnass, P. Vandergheynst (2009). Classification via incoherent subspaces. submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

R. Gribonval, H. Rauhut, K. Schnass, P. Vandergheynst (2008). Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *Journal of Fourier Analysis and Applications* **14**(5):655–687.

H. Rauhut, K. Schnass, P. Vandergheynst (2008). Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory* **54**(5):2210–2219.

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

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

Conference Papers

R. Gribonval, K. Schnass (2009). Basis identification from random sparse samples. *SPARS09.*

R. Gribonval, K. Schnass (2008). Dictionary identifiability from few training samples. *EUSIPCO08.*

R. Gribonval, K. Schnass (2008). Some recovery conditions for basis learning by l1-minimization. *ISCCSP08.*

K. Schnass, P. Vandergheynst (2008). Dictionary learning based dimensionality reduction for classification. *ISCCSP08.*

R. Gribonval, B. Mailhe, H. Rauhut, K. Schnass, P. Vandergheynst (2007). Average case analysis of multichannel thresholding. *ICASSP08.*